Marko Bachl
Analyse rezeptionsbegleitend gemessener Kandidatenbewertungen in TV-Duellen

Analyse rezeptionsbegleitend gemessener Kandidatenbewertungen in TV-Duellen

Erweiterung etablierter Verfahren und Vorschlag einer Mehrebenenmodellierung

Marko Bachl

Dissertation, Universität Hohenheim, 2014

Titel: Individuelle Kandidatenbewertungen während des gesamten Duells

Druck und Verlag: epubli GmbH, Berlin
www.epubli.de

ISBN 978-3-7375-0138-5

Printed in Germany

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Danksagung

Mein Dank geht zu allererst an Prof. Frank Brettschneider, der diese Dissertation als Erstgutachter betreut und mir – nicht nur für diese Arbeit – außerordentliche wissenschaftliche Freiheiten und Unterstützung gewährt hat. Ebenso danke ich Prof. Bertram Scheufele für sein sehr hilfreiches Gutachten und Prof. Michael Schenk für die Übernahme des Prüfungsvorsitzes.

Ein empirisches Forschungsprojekt ist selten die Leistung einer einzigen Person. Daher danke ich allen aktuellen und ehemaligen Kolleginnen und Kollegen, die an der Durchführung der TV-Duell-Studie vor der Landtagswahl 2011 in Baden-Württemberg beteiligt waren. Mein Dank gilt auch allen anderen Kolleginnen und Kollegen, mit denen ich seit Oktober 2008 am Institut für Kommunikationswissenschaft der Universität Hohenheim zusammenarbeiten durfte. Ganz besonders möchte ich mich bei meiner langjährigen Kollegin Catharina Vögele bedanken, die nicht nur an allen TV-Duell-Studien unseres Fachgebiets beteiligt war, sondern auch diesen Text mit wichtigen Korrekturen verbessert hat.

Die Idee, Mehrebenenmodelle zur Analyse der rezeptionsbegleitend gemessenen Kandidatenbewertungen einzusetzen, ist aus Gesprächen und einem Workshop mit Michael Scharkow und Jens Vogelgesang hervorgegangen. Ihnen bin ich in vielerlei Hinsicht und weit über dieses Projekt hinaus zu großem Dank verpflichtet. Michael Scharkow danke ich zudem dafür, dass er als einzige Person ein Kapitel dieser Arbeit in einer Version gelesen hat, die sich zum Wohle aller Leserinnen und Leser nicht mehr in diesem Buch befindet.

Schließlich danke ich meiner Familie und ganz besonders meiner Freundin Julia Daubertshäuser. Sie haben es überhaupt erst ermöglicht, dass ich diese Arbeit schreiben konnte.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

Ausgangslage und Zielsetzung

Wenn eine kommunikationswissenschaftliche Frage darauf abzielt, welche subjektiven Wahrnehmungen, Urteile oder Emotionen *während* der Rezeption kontinuierlicher Stimuli auftreten, stoßen die klassischen Modi der Befragung an ihre Grenzen. Als eine Möglichkeit, solche subjektiven Prozesse rezeptionsbegleitend zu erheben, hat sich die Real-Time Response (RTR) Messung als „[k]ontinuierliche Befragung in Echtzeit“ (Maurer, 2013b, S. 219) im Methodenrepertoire der empirischen Kommunikationsforschung etabliert. In einer Rezeptionsstudie, die eine RTR-Messung umfasst, nutzen die Teilnehmerinnen und Teilnehmer ein technisches Gerät, um kontinuierlich eine oder mehrere Fragen zu beantworten (vgl. Kapitel 3.1). Biocca, David und West (1994, S. 16) fassen den besonderen Nutzen der rezeptionsbegleitenden Messungen treffend zusammen:

> Such systems, by allowing subjects to continuously report their changing mental states, evaluations, and opinions, are well-suited to explore theoretical issues regarding the cognitive processing of continuous messages.

Der kontinuierliche, prozessuale Charakter der Medienrezeption und -wirkung ist in vielen Feldern der Rezeptions- und Wirkungsforschung relevant: Von der Unterhaltungsforschung über die politische Kommunikationsforschung bis hin zur Werbewirkungsforschung werden Studien durchgeführt, die Publikumsreaktionen während der Medienrezeption erfassen (vgl. Kapitel 3.2). Ein spezifisches Forschungsfeld, in dem RTR-Messungen häufig eingesetzt werden und sich als ertragreich erwiesen haben, ist die Wahrnehmung und Wirkung von politischen TV-Debatten (vgl. Kapitel 2 und 3.4). In diesen Studien werden die Teilnehmer vor und nach dem Duell befragt. Während des Duells geben sie mit RTR-Geräten Auskunft über den Eindruck, den sie im Moment von den Kandidaten haben. Häufig wird zusätzlich eine quantitative Inhaltsanalyse des Debatteninhalts durchgeführt (vgl. exemplarisch das in Kapitel 4 beschriebene Studiendesign). Durch die Verknüpfung der Daten aus Befragungen, RTR-Messung und Inhaltsanalyse versprechen diese Studien

wertvolle Einblicke in die individuelle Verarbeitung der TV-Debatten in Abhängigkeit von den Voreinstellungen der Rezipienten und den Debatteninhalten. Mit kaum einem anderen Design können wir diese Prozesse anhand extern valider Stimuli so detailliert erfassen, beschreiben und erklären.

Damit die Potenziale der RTR-Messungen genutzt werden können, muss zum einen ihre Datenqualität sichergestellt werden. Hierzu findet bereits seit geraumer Zeit eine methodologische Auseinandersetzung mit der Erhebungsmethode statt (vgl. Kapitel 3.3). Im Vordergrund stehen bei diesen Arbeiten die Auswirkungen der rezeptionsbegleitenden Messung auf die externe Validität der Studien sowie die Reliabilität und die Validität der erhobenen Daten. Die vorliegenden Befunde zeigen, dass RTR-Messungen die externe Validität zumindest im Vergleich zu anderen in einem Laborsetting durchgeführten Rezeptionsstudien nicht übermäßig einschränken. Auch erweisen sich die erhobenen Daten als hinreichend reliabel und valide, um ihre Verwendung in quantitativen Analysen rechtfertigen zu können.

Zum anderen müssen die Verfahren, die zur Auswertung der RTR-Messungen verwendet werden, dem Ziel der Analysen angemessen sein. Während sich vielfältige methodologische Literatur mit der RTR-Messung als Erhebungsinstrument beschäftigt, liegen kaum explizite Auseinandersetzungen mit den Analysemöglichkeiten der erhobenen Daten vor. Zu den wenigen bemerkenswerten Ausnahmen gehören der weiterhin zentrale Überblicksartikel von Biocca et al. (1994) und eine aktuelle Dissertation von Nagel (2012) (vgl. Kapitel 5). Dies ist aus mindestens zwei Gründen bedauerlich: *Erstens* sind Studien mit RTR-Messungen sehr ressourcenintensiv, gerade wenn sie wie die TV-Duell-Studien während der Ausstrahlung eines echten Medienereignisses durchgeführt werden. Die notwendige Hardware ist teuer, der Personalaufwand seitens der Forscher ist groß, und auch die Studienteilnehmer nehmen einen erheblichen Aufwand auf sich. Angesichts der Ressourcen, die in die Datenerhebung investiert werden, ist es schwer nachvollziehbar, dass für die anschließende Datenanalyse häufig kaum Aufwand betrieben wird.

Zweitens sind Datensätze aus solchen Studien im Vergleich zu einfachen Befragungsdaten sehr komplex (vgl. Kapitel 5.1). Dies gilt vor allem dann, wenn Informationen aus der Befragung vor der Stimulusrezeption und aus einer quantitativen Analyse des Stimulusinhalts mit den RTR-Messungen verknüpft werden sollen. Die komplexe Datenstruktur macht es schwierig, mithin sogar unmöglich, die klassischen Lehrbuchverfahren sinnvoll einzusetzen. Die meisten (veröffentlichten) Arbeiten begegnen diesem Hindernis, indem sie vor der eigentlichen Analyse eine Aggregation der Daten vornehmen. Die Zusammenfassung erfolgt dabei entlang (mindestens) einer dieser beiden Dimensionen (J. Maier, 2013, S. 178-180; Maurer & Reinemann, 2009, S. 6-9):

1. *Aggregation über Personen*: Es werden eine oder mehrere RTR-Zeitreihen gebildet, indem die Messungen von allen oder von Teilgruppen der Rezipienten zu einem Messzeitpunkt zusammengefasst werden. In einer typischen TV-Duell-Studie werden so die bekannten RTR-Kurven gebildet, die über die mittlere Bewertung der Kandidaten durch das gesamte Publikum oder durch die politischen Lager im Verlauf der Debatte Auskunft geben (vgl. Kapitel 5.2).
2. *Aggregation über Messzeitpunkte*: Es werden für jeden Rezipienten eine oder mehrere Variablen gebildet, indem alle oder ein Teil der RTR-Messungen des Rezipienten zusammengefasst werden. In einer TV-Duell-Studie können diese Variablen als Indikator dafür genutzt werden, wie ein Rezipient die Kandidaten während der gesamten Debatte (z.B. Bachl, 2013b; J. Maier, 2007) oder während einzelner Debattenausschnitte (z.B. während der bildungspolitischen Duellabschnitte, Bachl & Vögele, 2013) bewertet.

Die Aggregationen helfen dabei, die Komplexität der Daten zu reduzieren und sie so der Datenanalyse mit den geläufigen Verfahren zugänglich zu machen. Wie noch ausführlich zu zeigen sein wird, entstehen durch die Aggregationen jedoch einige Probleme (vgl. Kapitel 5). Im besten Fall werden lediglich wertvolle Informationen, die in den Rohdaten noch enthalten sind, vernachlässigt. Im schlechtesten Fall kann die Aggregation aber auch zu theoretisch unangemessenen Interpretationen und fehlerhaften Inferenzschlüssen führen. Ohne den folgenden Ausführungen zu weit vorgreifen zu wollen, kann die wesentliche Schwäche der aggregationsbasierten Vorgehensweise bereits mit einem einfachen Abgleich mit dem oben angeführten Zitat von Biocca et al. (1994, S. 16) dargelegt werden. Mit dem Einsatz einer RTR-Messung ist die Hoffnung verbunden, die kognitive Verarbeitung einer kontinuierlichen Botschaft zu beschreiben und zu erklären. Wir sind also an *individuellen* und *kontinuierlichen* Prozessen interessiert. Die Berechnung von Zeitreihen nach dem ersten Aggregationsansatz erhält zwar die längsschnittliche Natur der Daten, jedoch werden die Urteile der individuellen Rezipienten zu Personenaggregaten zusammengefasst. Es ist jedoch nicht nur in der Medienwirkungsforschung bekannt, dass mit Analysen solcher Aggregate keine Hypothesen zu individuellen Prozessen getestet werden können, ohne das Risiko eines ökologischen Fehlschlusses einzugehen (Yanovitzky & Greene, 2009). Die Individualebene der Theorien zur Wahrnehmung, Verarbeitung und Wirkung der Medienstimuli passt hier nicht zur Aggregatebene der Datenanalyse.

Die Zusammenfassung von RTR-Messungen einzelner Rezipienten nach dem zweiten Ansatz erhält auf Ebene der Personen zwar das Individualniveau der

ursprünglichen Messung. Sie führt aber zu einem Verlust der Variation über die Zeit, verhindert eine Verknüpfung von einzelnen Passagen des Stimulus mit den Reaktionen der Rezipienten und macht so den zentralen Nutzen der rezeptions*begleitenden* Messung zunichte. Gerade bei der Analyse der Reaktionen auf längere, multiepisodische Medienstimuli ist dies bedauerlich, da statt der wiederholten Reaktion auf einen bestimmten Stimulusinhalt nur die durchschnittliche Reaktion auf den Inhalt im gesamten Stimulus untersucht wird. In der Forschung zur Erklärung der unmittelbaren Kandidatenbewertungen in TV-Debatten ist die Verknüpfung mit den Inhalten einzelner Kandidatenaussagen eine zentrale Forschungslücke (McKinney & Carlin, 2004; Reinemann & Maurer, 2008), zu deren Verringerung wir mit derart aggregierten Daten wenig beitragen können.

Die vorliegende Arbeit ist dadurch motiviert, der Analyse der Daten einer RTR-Messung eine Beachtung zu schenken, die dem bei ihrer Erhebung betriebenen Aufwand gerecht wird. Dabei verfolgen wir zwei forschungsleitende Zielsetzungen:

1. Wir diskutieren die Grenzen der auf aggregierten RTR-Messungen basierenden Verfahren und erklären, welche Konsequenzen sie für die Interpretation der Befunde haben. Zudem entwickeln wir das am weitesten verbreitete Verfahren zur explorativen Analyse von RTR-Messungen, die Peak-Spike-Analyse, weiter und geben Empfehlungen, wie sich ihre Befunde theoretisch angemessen darstellen und einordnen lassen (Kapitel 5).

2. Wir schlagen drei Klassen von Mehrebenenmodellen vor, mit denen wesentliche Limitationen der bisher eingesetzten Verfahren zur deduktiven Analyse der RTR-Messungen behoben werden (Kapitel 6).

Neben diesen konkreten Zielsetzungen ist es ein wichtiges allgemeines Anliegen, die Charakteristika der bisher kaum beachteten individuellen RTR-Messungen von den aus zahlreichen Publikationen bekannten aggregierten Verlaufskurven abzugrenzen. Im Verständnis dieses Unterschieds liegt unseres Erachtens der Schlüssel, um die theoretisch-konzeptionellen Modelle, die zur Erklärung der rezeptionsbegleitend gemessenen Konstrukte herangezogen werden, mit dem angemessenen Datenniveau bei ihrer empirischen Analyse zusammenzubringen.

Inhaltlicher Schwerpunkt dieser Arbeit ist die Erklärung der unmittelbaren Bewertung von Stefan Mappus (CDU) und Nils Schmid (SPD) im TV-Duell vor der baden-württembergischen Landtagswahl 2011. Alle Diskussionen datenanalytischer Verfahren werden anhand der Daten aus einer umfangreichen

Rezeptionsstudie zu dieser TV-Debatte (Bachl, Brettschneider & Ottler, 2013a) praktisch veranschaulicht. Unter anderem analysieren wir die Bewertung der beiden Kandidaten in Abhängigkeit von den Voreinstellungen der Rezipienten und den Themen der Kandidatenaussagen. Dabei nehmen wir auch eine Überprüfung von Annahmen vor, die auf den Ansatz des Issue Ownership zurückgehen (Petrocik, 1996). Die zweite weitergehende Analyse befasst sich mit den Effekten der Relationen (Selbstpräsentationen, Angriffe, Verteidigungen, vgl. W. L. Benoit, 2007) in den Kandidatenaussagen im Zusammenspiel mit den Voreinstellungen der Rezipienten. Neben der Weiterentwicklung der analytischen Verfahren leistet diese Arbeit so auch einen Beitrag zu zwei wesentlichen Strängen der kommunikationswissenschaftlichen TV-Debatten-Forschung.

Der direkte Nutzen der vorliegenden Arbeit liegt also in der Weiterentwicklung der Analysewerkzeuge für die Auswertung von mit RTR-Messungen erfassten unmittelbaren Kandidatenbewertungen in TV-Duell-Studien. Mit diesen Verfahren soll es möglich sein, die individuelle und kontinuierliche Verarbeitung und Wirkung politischer Kommunikation präziser zu beschreiben und besser zu verstehen. Die größte Relevanz besitzt diese Arbeit damit sicherlich für diejenigen, die selbst in diesem Forschungsfeld und mit diesem Messinstrument arbeiten. Jedoch lassen sich die vorgestellten Verfahren auch in vielen anderen Kontexten der Kommunikationswissenschaft, in denen prozessbezogene Daten während der Medienrezeption mit RTR-Messungen erfasst werden, anwenden. Auch wenn dies zurzeit nur einen relativ kleinen Anwenderkreis umfasst, ist zu hoffen, dass er sich in den kommenden Jahren erweitert. Die Integration von RTR-Messverfahren in Laborsoftware (z.B. Jarvis, 2012) und Online-Befragungen (z.B. Iyengar, 2011; Kercher, Bachl, Vögele & Vohle, 2012) reduziert die Abhängigkeit von vergleichsweise teurer apparativer Ausstattung und ermöglicht eine weitere Verbreitung der Methode. Zudem stehen durch die Aufnahme einer umfangreichen TV-Duell-Studie in die German Longitudinal Election Study erstmals entsprechende Datensätze für Sekundäranalysen bereit (Rattinger, Roßteutscher, Schmitt-Beck, Weßels & Wolf, 2013). Schließlich ist zu hoffen, dass auch diejenigen von der vorliegenden Arbeit profitieren können, die selbst keine Studien mit RTR-Messungen durchführen, sie aber in der Fachliteratur rezipieren. Die Diskussion der verschiedenen Analyseverfahren soll auch dabei helfen, Befunde aus solchen Arbeiten besser einordnen zu können.

Begriffsklärung

In dieser Arbeit verstehen wir unter RTR-Messungen alle Messverfahren, mit denen Versuchsteilnehmer während der Rezeption eines kontinuierlichen, in

der Regel audiovisuellen, Stimulus bewusst Angaben zu einem vor der Rezeption definierten Konstrukt machen (vgl. zur technischen Durchführung der Messung ausführlich Kapitel 3.1). Für diese Messverfahren kursieren neben „RTR-Messungen“ in der Literatur weitere Bezeichnungen. Vor allem in der englischsprachigen Literatur verbreitet sind die Begriffe „continuous response measurement (CRM)“ (z.B. Biocca et al., 1994) und „moment-to-moment responses (MTM)“ (z.B. Baumgartner, Sujan & Padgett, 1997). Wir verwenden im Folgenden die Bezeichnung RTR-Messung, da sie sich in der deutschsprachigen Literatur weitestgehend durchgesetzt hat.

Abgegrenzt werden muss das hier behandelte RTR-Verfahren von apparativen Beobachtungsverfahren, insbesondere von kontinuierlichen psychophysiologischen Messmethoden (vgl. für einen Überblick Fahr & Hofer, 2013). Auch physiologische Messungen dienen dazu, einen Einblick in Wahrnehmungen und Gefühlszustände während der Medienrezeption zu erhalten. Allerdings erfassen sie physiologische Indikatoren, die von den Studienteilnehmern in der Regel nicht bewusst beeinflusst werden können. Die von physiologischen Messverfahren erzeugte Datenstruktur hat mit den hier behandelten Daten von RTR-Messungen einige Charakteristika gemein. Ob und inwieweit die im Folgenden präsentierten Vorschläge für ihre Analyse ebenfalls hilfreich sind, können und wollen wir in dieser Arbeit jedoch nicht weiter vertiefen. Wenn wir im Folgenden auch allgemein von rezeptionsbegleitenden Messungen sprechen, sind – sofern nicht ausdrücklich darauf hingewiesen wird – von den Rezipienten bewusst wiedergegebene Angaben im Sinne der RTR-Messungen gemeint.

Aufbau der Arbeit

Im folgenden Kapitel 2 begründen wir, warum TV-Debatten lohnenswerte Untersuchungsgegenstände für die politische Kommunikationsforschung sind. Im Mittelpunkt des Kapitels stehen mögliche Forschungsfragen, die mit einer TV-Duell-Studie beantwortet werden können, wenn die unmittelbaren Bewertungen der Kandidaten rezeptionsbegleitend erfasst werden. Kapitel 3 vermittelt anschließend einen Überblick über den Einsatz von RTR-Messungen in der Kommunikationswissenschaft und verwandten Disziplinen. Hier stellen wir zuerst dar, welche Techniken zur RTR-Messung genutzt werden, beschreiben typische Charakteristika von RTR-Studien auch jenseits der Forschung zu TV-Debatten und fassen den methodologischen Forschungsstand zusammen. Im abschließenden Teilkapitel 3.4 widmen wir uns ausführlicher empirischen Studien, in denen unmittelbare Kandidatenbewertungen während TV-Debatten erklärt werden. Dabei unterscheiden wir zwischen induktiv angelegten Arbeiten,

in denen ausgehend von den empirisch beobachteten Kandidatenbewertungen mögliche Erklärungsansätze identifiziert werden, und deduktiven Analysen, die Annahmen über die Effekte bestimmter Eigenschaften des Debatteninhalts auf die Kandidatenbewertungen prüfen. Neben den Befunden werden auch das methodische Vorgehen und die zur Erklärung herangezogenen theoretischen Ansätze thematisiert.

Der empirische Teil beginnt mit Kapitel 4, in dem die Rezeptionsstudie zum TV-Duell vor der baden-württembergischen Landtagswahl 2011 vorgestellt wird. Besondere Beachtung finden Reliabilität und Validität der RTR-Messungen, da ihre Analyse in dieser Arbeit im Mittelpunkt steht. Kapitel 5 setzt sich mit den Verfahren auseinander, die sich bisher für die Analyse der unmittelbaren Kandidatenbewertungen während TV-Debatten etabliert haben. Zuerst gehen wir auf das Verfahren der Peak-Spike-Analyse ein, das als Standardverfahren in induktiv angelegten Arbeiten gelten kann. Wir diskutieren die Probleme dieses Analyseansatzes und schlagen einige Erweiterungen vor. Auf dieser Basis geben wir auch Empfehlungen für die zukünftige Verwendung des Verfahrens. Im Anschluss wenden wir uns den etablierten Verfahren zur deduktiven Analyse zu. Wir argumentieren, dass Ergebnisse, die auf diesen Verfahren basieren, nur sehr bedingt aussagekräftig sind, da sie entweder auf einer Aggregation der individuellen RTR-Messungen über Personen oder auf einer Zusammenfassung vieler Messzeitpunkte basieren. Vor allem die erstgenannten Verfahren sind daher nicht zu einer Überprüfung von Annahmen geeignet, die auf Ansätzen der individuellen Wahrnehmung und Informationsverarbeitung beruhen. Zudem sind viele Inferenzschlüsse auf Grundlage der Ergebnisse aller etablierten Verfahren fehlerbehaftet.

Um die Limitationen der bisher zur deduktiven Analyse eingesetzten Verfahren zu beheben, schlagen wir in Kapitel 6 eine Mehrebenenmodellierung der unmittelbaren Kandidatenbewertungen in TV-Debatten vor. Das Kapitel ist in vier Teile gegliedert. Zuerst führen wir das einfache Wachstumskurvenmodell ein (Kapitel 6.1). Die Modellklasse ermöglicht es, die dynamische Veränderung der unmittelbaren Bewertungen eines Kandidaten während einer Antwort durch Merkmale der individuellen Rezipienten zu erklären. In Kapitel 6.2 präsentieren wir das kreuzklassifizierte Mehrebenenmodell. Hier wird berücksichtigt, dass jede RTR-Bewertung von genau einem Rezipienten stammt und dass jede RTR-Bewertung zu genau einer Kandidatenaussage abgegeben wird. Dadurch wird es ermöglicht, die Bewertungen gleichzeitig in Abhängigkeit von Merkmalen der Rezipienten und der Aussagen sowie ihren Interaktionen zu analysieren. Das in Kapitel 6.3 vorgestellte kreuzklassifizierte Wachstumskurvenmodell kombiniert die beiden genannten Modellklassen. So ist es möglich, die dynamische Veränderung der Bewertungen innerhalb vieler

Kandidatenaussagen systematisch durch Merkmale der Aussagen und Merkmale der Rezipienten zu erklären. In allen drei Teilkapiteln geben wir zu Beginn eine kurze Einführung in die formale Funktionslogik der Modellklassen, bevor wir sie an praktischen Beispielen zur Bewertung der Kandidaten im TV-Duell vor der baden-württembergischen Landtagswahl 2011 demonstrieren. Das letzte Teilkapitel 6.4 diskutiert die Stärken und Schwächen der vorgeschlagenen Modelle und regt weitere Modellerweiterungen an.

Im abschließenden Fazit ordnen wir die Befunde in den größeren Rahmen der TV-Duell-Studien ein und geben einen Ausblick zur Verwendung der vorgestellten Verfahren jenseits von Studien zu diesem Untersuchungsgegenstand (Kapitel 7).

2 Bedeutung von TV-Debatten und ihrer empirischen Untersuchung

Im folgenden Kapitel begründen wir, warum die empirische Untersuchung von TV-Debatten im Allgemeinen und speziell mit Studiendesigns, die eine rezeptionsbegleitende Messung der Kandidatenbewertungen enthalten, lohnenswert ist. Dazu stellen wir zuerst die gesellschaftliche Bedeutung von TV-Debatten als zentrale Ereignisse moderner Medienwahlkämpfe dar. Anschließend diskutieren wir, welche Nutzen die Kommunikationsforschung aus empirischen Studien zu TV-Debatten über die Untersuchung eines wichtigen Wahlkampfereignisses hinaus ziehen kann. Aus diesen Überlegungen leiten wir schließlich ab, wie Studien, in denen die Bewertung der Kandidaten auch während der Debattenrezeption erfasst wird, zur Beantwortung kommunikationswissenschaftlicher Fragestellungen beitragen. Ziel dieses Teilkapitels ist es explizit *nicht*, einen umfassenden Überblick über die Forschung zu TV-Debatten zu geben. Eine solche Literaturschau wäre in Anbetracht der wissenschaftlichen Aufmerksamkeit, die diesem Untersuchungsgegenstand entgegengebracht wird, alleine einer Monographie angemessen. An dieser Stelle sei daher auf einige aktuellere Überblickswerke verwiesen (z.B. W. L. Benoit, 2013; W. L. Benoit, Hansen & Verser, 2003; Maurer & Reinemann, 2007a; McKinney, 2007; McKinney & Carlin, 2004; Racine Group, 2002; Reinemann & Maurer, 2008).

Gesellschaftliche Relevanz

TV-Debatten der Spitzenkandidaten für das Amt des Regierungschefs sind zentrale Ereignisse in modernen Medienwahlkämpfen. Ihre besondere Relevanz ergibt sich aus der Größe und der Zusammensetzung des Publikums, der speziellen Präsentationsform der politischen Inhalte und aus den ihnen zugeschriebenen Effekten. Das laut McKinney und Carlin (2004, S. 203) meist genannte Argument für die Bedeutung der TV-Debatten ist ihre große Reichweite. In den USA, dem Ursprungsland des Formats, sind Fernsehdebatten führender Politiker seit Langem sehr populär.[1] In vielen anderen Ländern wurden sie in den letzten Jahrzehnten eingeführt und haben sich beim Publikum

1 Vgl. dazu die Reichweiten der Debatten in US-Präsidentschaftswahlkämpfen auf der Website der „Commission on Presidential Debates“: `www.debates.org`.

durchgesetzt (Reinemann & Maurer, 2008, S. 5058-5059). Auch in Deutschland waren die TV-Duelle der Kanzlerkandidaten seit ihrer Einführung vor der Bundestagswahl 2002 stets das singuläre Medienereignis mit der größten Reichweite im Bundestagswahlkampf: Zwischen 14 und 21 Millionen Zuschauer verfolgten diese Duelle (Geese, Zubayr & Gerhard, 2005; Gscheidle & Gerhard, 2013; Zubayr, Geese & Gerhard, 2009; Zubayr & Gerhard, 2002). Die Debatten der Spitzenkandidaten vor Landtagswahlen erreichen ebenfalls regelmäßig einen signifikanten, wenn auch geringeren Anteil der Wähler (Vögele, Brettschneider & Bachl, 2013, S. 32-33).

Neben der schieren Größe sind die Zusammensetzung des Publikums und die Einstellung der Zuschauer gegenüber dem Format bemerkenswert. Auch Wähler, die sich weniger stark für Politik interessieren und nur selten Kontakt zu politischen Informationen haben, schalten bei diesem Medienereignis ein (z.B. Dehm, 2002; J. Maier & Faas, 2011; McKinney & Carlin, 2004). Die Zuschauer schreiben den Duellen eine große Bedeutung zu, unterstellen ihnen große Wirkung und haben die Motivation, durch die Sendung etwas über die Kandidaten zu lernen (Reinemann & Maurer, 2008; Zubayr & Gerhard, 2002). Vor dem Hintergrund der zunehmenden Differenzierung und Fragmentierung des medialen (Informations-) Angebots in der modernen Mediengesellschaft (z.B. Schulz, 2011, S. 20-41) sind Größe und Charakteristika des Publikums von TV-Debatten besonders relevant. Nur noch wenigen politischen Informationsformaten gelingt es, breit gestreute Wählerschichten in diesem Umfang zu erreichen.

Das Publikum alleine begründet jedoch noch nicht die gesellschaftliche Relevanz dieser Medienereignisse – die Debatten müssen dazu auch Wirkungen auf die Zuschauer haben. Zahlreiche Studien weisen nach, dass die Debatten eine Vielzahl von direkt oder indirekt mit der Wahlentscheidung verbundenen Einstellungen, Kognitionen und Verhaltensabsichten beeinflussen können (vgl. z.B. für Systematisierungen McKinney & Carlin, 2004; Reinemann & Maurer, 2008; Zhu, Milavsky & Biswas, 1994; für eine Meta-Analyse von 33 Wirkungsstudien zu US-amerikanischen Debatten W. L. Benoit, Hansen & Verser, 2003; für vergleichende Sekundäranalysen von Befragungsdaten zu mehreren Debatten u.a. Blais & Perrella, 2008; J. Maier & Faas, 2011; Schrott & Lanoue, 2013). Aus der Kombination von großer Reichweite und nachgewiesenen Wirkungen auf das Publikum lässt sich ableiten, dass die TV-Debatten letztendlich eine mitentscheidende Rolle im Wahlkampf spielen können. Als Fazit lässt sich hier der Befund einer Analyse repräsentativer Befragungsdaten zu den Kanzlerduellen 2002, 2005 und 2009 von J. Maier und Faas (2011, S. 86) zitieren:

> TV debates have the power to alter the electoral support of candidates and political parties. This, of course, can have an impact on the course of the campaign and – especially if the race is close (which was the case in 2002 and 2005) – might have the power to change the outcome of an election.

Relevanz für die politische Kommunikationsforschung

Dass die politische Kommunikationsforschung den TV-Debatten eine große Relevanz zumisst, zeigt sich nicht zuletzt an der großen Zahl der empirischen Studien, die sich mit diesem Gegenstand beschäftigen. In ihrem Review der (englischsprachigen) Forschungsliteratur aus dem Jahr 2004 zählen McKinney und Carlin (2004, S. 204) bereits mehr als 800 Arbeiten, die sich mit unterschiedlichen Aspekten dieses Untersuchungsgegenstands beschäftigen. In den letzten zehn Jahren dürften alleine durch Studien zu späteren Debatten noch einige mehr hinzugekommen sein. Auch in der Forschung zu Wahlkämpfen in Deutschland haben die TV-Duelle einen festen Platz gefunden. Zu allen Kanzlerduellen seit den zwei Debatten zwischen Gerhard Schröder und Edmund Stoiber im Bundestagswahlkampf 2002 liegen zahlreiche Publikationen vor.[2] Ein weiterer Indikator dafür, dass TV-Duell-Studien als Bestandteil der Wahlkampfforschung institutionalisiert sind, ist die Aufnahme eines entsprechenden Moduls in die „German Longitudinal Election Study" (GLES) (Rattinger et al., 2013). Im Rahmen der nationalen Wahlstudien zu den Bundestagswahlen 2009 und 2013 wurden jeweils auch groß angelegte Rezeptionsstudien zu den TV-Duellen zwischen Angela Merkel und Frank-Walter Steinmeier bzw. Peer Steinbrück durchgeführt.

Die bedeutende Stellung, die TV-Duell-Studien in der akademischen Wahlkampf- und Kommunikationsforschung einnehmen, begründet sich maßgeblich durch die große gesellschaftliche Relevanz, die dem Medienereignis zugeschrieben wird. Wenn ein Ereignis potenziell einen großen Einfluss auf den Verlauf eines Wahlkampfs nehmen kann, dann muss sich auch die Forschung damit aus-

2 Zum Zeitpunkt des Verfassens dieser Arbeit bezieht sich dies auf die Bundestagswahlkämpfe 2002, 2005 und 2009. Neben den Studien mit rezeptionsbegleitender Erfassung der Kandidatenbewertungen von Faas und Maier (2004a), Maurer und Reinemann (2003), Maurer, Reinemann, Maier und Maier (2007) und Faas, Maier, Maier und Brettschneider (2009), aus denen viele der im Forschungsstand in Kapitel 3.4 vorgestellten Publikationen hervorgegangen sind, seien hier beispielhaft genannt: Donsbach, Jandura und Hastall (2002); Hofrichter (2004); Klein (2005); Klein und Rosar (2007); J. Maier und Faas (2011); Müller (2003); Scheufele, Schünemann und Brosius (2005); Tapper und Quandt (2006, 2010). Die zuvor in Bundestagswahlkämpfen üblichen „Elefantenrunden" mit Spitzenvertretern aller im Bundestag vertretenen Parteien erhielten dagegen verhältnismäßig wenig Aufmerksamkeit (z.B. Schrott, 1990a, 1990b).

einandersetzen. TV-Debatten sind darüber hinaus aber auch äußerst geeignete Untersuchungsgegenstände für die empirische Wahl- und Kommunikationsforschung. In seinem Überblickswerk zur politischen Kommunikation hat Schulz (2011, S. 217) Wahlkämpfe als „eine Art Forschungslabor“ beschrieben, das sich besonders gut dazu eignet, Grundlagenforschung in diesem Feld zu betreiben. In der natürlich geschaffenen „Laborsituation“ nehmen TV-Debatten als Bestandteile der massenmedial vermittelten Kommunikation eine besondere Rolle ein. Denn hier treten die wichtigsten Akteure des Wahlkampfs vor ein großes und vielfältiges Publikum, um die wichtigsten Themen des Wahlkampfs zeitlich komprimiert und in einer relativ standardisierten Präsentationsform zu kommunizieren. Faas und Maier (2004a, S. 56) haben zur Beschreibung dieses Formats die Analogie der „Wahlkämpfe im Miniaturformat“ geprägt, McKinney und Carlin (2004, S. 204) bezeichnen die Inhalte einer TV-Debatte als „a capsule summary of campaign issues“. Nicht übergangen werden sollen an dieser Stelle kritische Stimmen zu TV-Duellen der Spitzenkandidaten der größeren (Volks-) Parteien in politischen Systemen, in denen nicht Personen, sondern Parteien zur Wahl stehen (z.B. Donsbach, 2002; Donsbach et al., 2002). Diesen Duellen fehlen mit den Stimmen der übrigen Parteien natürlich wesentliche Bestandteile des Wahlkampfs. Beispielsweise war der Spitzenkandidat der Grünen und spätere Ministerpräsident Winfried Kretschmann zum in dieser Arbeit untersuchten TV-Duell überhaupt nicht eingeladen (Vögele, 2013). Die Analogie der Debatten als Miniaturwahlkämpfe reicht hier nur so weit, wie sie sich auf die Auseinandersetzung der Kandidaten bezieht, die nach Ansicht der Ausrichter der Debatten die größten Chancen haben, die kommende Regierung anzuführen. Folgen wir aber den genannten Analogien, so können wir aus TV-Duell-Studien nicht nur etwas über Inhalte, Wahrnehmungen und Wirkungen einer TV-Debatte lernen, sondern sie auch zur Beantwortung weiter gefasster Fragen der Wahl- und Kommunikationsforschung nutzen.

Zum einen können wir die Befunde als Indikatoren für die politische Kommunikation in den Wahlkämpfen, in deren Kontext die Debatte stattfindet, auffassen. Aus den Inhalten der Kandidatenaussagen können wir auf die Botschaftsstrategien der Spitzenkandidaten und ihrer Wahlkampfteams im Wahlkampffinale schließen. Es ist davon auszugehen, dass die Kandidaten die Debatten auch dazu nutzen, die wichtigsten Botschaften noch einmal für das große TV-Publikum zusammenzufassen (vgl. z.B. zu den Strategien der Kandidaten für das TV-Duell vor der Landtagswahl 2011 in Baden-Württemberg Krafft & Zaiss, 2013). Ebenso kann die Themenauswahl der Sendungsredaktion und der Moderatoren als ein Indikator für journalistische Relevanzzuschreibungen in diesem Wahlkampf gelten (W. L. Benoit & Hansen, 2001; J. Maier & Maier, 2013). Die in einer TV-Duell-Studie festgestellten Wirkungen der Debatte

auf das Publikum werden in den größeren Kontext der Meinungsdynamiken in den letzten Wahlkampfwochen gestellt, um die Bedeutung der Debatteneffekte, aber auch von Medienwirkungen im Allgemeinen, abschätzen zu können. Wenn zudem die Reaktionen des Publikums auf die einzelnen Kandidatenaussagen rezeptionsbegleitend gemessen werden, können die Befunde schließlich als Indikatoren dafür dienen, wie die Kernbotschaften der Kandidaten von den Wählern auch außerhalb des TV-Duells im Wahlkampf beurteilt werden.

Zum anderen werden Studien zu *einzelnen* TV-Duellen genutzt, um allgemeine, vom Kontext der untersuchten Debatte und des sie umgebenden Wahlkampfs losgelöste Aussagen zu treffen bzw. allgemeine Hypothesen zu überprüfen. Diese Indikatorfunktion der TV-Duell-Studien ist vor allem für die Medienwirkungsforschung relevant.[3] Das Forschungsinteresse gilt dann nicht mehr der Bewertung *dieser* Kandidaten in *diesem* Duell oder den Wirkungen *dieses* Duells in *diesem* Wahlkampf, sondern den Erklärungen für die Bewertung von Kandidaten in TV-Debatten oder den Wirkungen von TV-Debatten im Allgemeinen. Diese Überlegung kann noch einen Schritt weiter geführt werden. Die TV-Duelle können in Medienwirkungsstudien auch als Beispiele allgemeiner Stimuli betrachtet werden, anhand derer die Wahrnehmungen und Wirkungen massenmedial vermittelter (politischer) Kommunikation untersucht werden. Der Umstand, dass diese Stimuli TV-Debatten sind, ist hier nur noch eine Randbedingung, unter der die Gültigkeit allgemeiner Hypothesen der Rezeptions- und Wirkungsforschung überprüft wird.

Ausgehend von den Wirkungen der TV-Duelle als stellvertretende Stimuli wird in Wirkungsstudien also auch auf die Wirkungen von anderen politischen Botschaften im selben Wahlkampf, von TV-Duellen im Allgemeinen oder ganz generell von massenmedial vermittelter (politischer) Kommunikation geschlossen. Die Validität solcher Inferenzschlüsse lässt sich am besten im Vergleich mit den typischen Designs der Medienwirkungsforschung einordnen: der Kombination von Befragungen und Inhaltsanalysen und dem Experimentaldesign

3 Studien zu den Debatteninhalten und darauf aufbauenden Inferenzschlüssen auf die Kandidaten und Journalisten als Kommunikatoren lassen sich verhältnismäßig einfach um Inhaltsanalysen weiterer, auch in der Vergangenheit liegender TV-Debatten erweitern, um über mehrere TV-Debatten hinweg verallgemeinerbare Schlüsse zu ziehen. Siehe hierzu z.B. stellvertretend für die umfangreiche Arbeit von Benoit und Kollegen zur (international) vergleichenden Inhaltsanalyse von TV-Debatten (W. L. Benoit, 2007, 2013), die vergleichende Analyse der Inhalte aller Debatten im US-amerikanischen Präsidentschaftswahlkampf 2000 von Carlin, Morris und Smith (2001), die automatisierte Inhaltsanalyse von Kampagnenkommunikation in US-amerikanischen Präsidentschaftswahlkämpfen inklusive vieler TV-Debatten von 1948 bis 1996 von Hart und Jarvis (1997) oder die Inhaltsanalyse sämtlicher TV-Debatten der Spitzenkandidaten vor Bundes- und Landtagswahlen in Deutschland von Jansen und Maier (2013). Für Rezeptions- und Wirkungsstudien ist eine solche Erweiterung sehr aufwändig, für vergangene Duelle unmöglich.

(Maurer, 2013a). Abzuwägen sind die interne und die externe Validität des Inferenzschlusses auf die Wirkungen anderer Stimuli (Bortz & Döring, 2006, S.53).[4] Die interne Validität meint in diesem Kontext die Kausalattribution der Effekte zu bestimmten Charakteristika des medialen Stimulus. Die externe Validität meint die Übertragbarkeit der Effekte auf andere Stimuli, die ähnliche Charakteristika wie der untersuchte Stimulus aufweisen.

In Studien, die Daten aus Befragungen und Inhaltsanalysen auf Individualniveau kombinieren, wird zuerst die Nutzung der Medienangebote durch die Befragten erfasst. Dann werden inhaltsanalytisch in den von den Befragten genutzten Medien die für die Forschungsfrage interessanten Charakteristika der Medieninhalte bestimmt. Aus der Nutzungshäufigkeit der Medien und dem Vorkommen der relevanten Charakteristika in den Medien kann für jeden Befragten eine Kontaktwahrscheinlichkeit bestimmt werden (z.B. Maurer, 2012; Wolling & Wirth, 2012). Diesem Vorgehen kann grundsätzlich eine hohe externe Validität zugeschrieben werden, da es in der Bandbreite von potenziell zu erfassenden Stimuli nur durch praktische Erwägungen beschränkt ist. Weder die Abfrage der Mediennutzung noch das Untersuchungsmaterial der Inhaltsanalysen lassen sich bis hin zur Vollständigkeit erweitern. Grundsätzlich können so die Effekte vieler medialer Stimuli erfasst werden, eine Übertragbarkeit der Befunde auf weitere ähnliche Stimuli sollte dadurch gegeben sein. Die interne Validität solcher Studien ist dagegen als geringer einzuschätzen, da sich die Effekte nicht eindeutig kausal auf die Rezeption bestimmter Charakteristika der Stimuli zurückführen lassen, sondern nur mit Kontaktwahrscheinlichkeiten angenähert werden.

Bei der Medienwirkungsforschung mit Experimenten ist das Verhältnis von externer zu interner Validität genau umgekehrt. Hier werden den Rezipienten in randomisierten Gruppen verschiedene Versionen eines Stimulus präsentiert, bei dem im Idealfall nur genau das Merkmal manipuliert wird, dessen Wirkung von Interesse ist. Dementsprechend ist eine hohe interne Validität gegeben, da Effekte genau diesem Merkmal kausal zugeschrieben werden können. Fraglich ist allerdings, inwieweit sich von einer geringen Zahl künstlich geschaffener bzw. manipulierter Stimuli auf eine größere Population von echten

4 In der folgenden Argumentation geht es ausschließlich um die Validität des Inferenzschlusses von den Wirkungen eines TV-Duells als Stimulus auf die Wirkungen anderer Stimuli. Andere Faktoren, welche die Validität einer Studie ganz allgemein beeinflussen – u.a. Ziehung und Zusammensetzung der Stichprobe, Versuchsanordnung und -durchführung, Instrumente der Datenerhebung und Analyseverfahren (Bortz & Döring, 2006, S.53-58) – spielen bei diesen Überlegungen zunächst keine Rolle. Auf die Bedeutung der Stichprobenziehung kommen wir im nächsten Abschnitt zu sprechen. Die Validität von typischen Studiendesigns, in denen RTR-Messungen eingesetzt werden, sowie Reliabilität und Validität von RTR-Messungen werden in Kapitel 3.3 diskutiert.

Wahlkampfbotschaften, TV-Duellen oder anderen massenmedial vermittelten (politischen) Inhalten schließen lässt (Faas & Huber, 2010; Klimmt & Weber, 2013; Trepte & Wirth, 2004). Bis zur wiederholten Replikation des Experiments mit anderen Stimuli, bei denen aber dasselbe Merkmal manipuliert wird, ist die Generalisierbarkeit kritisch zu betrachten.

Eine Wirkungsstudie mit einem TV-Duell (während dessen Ausstrahlung im Wahlkampf) als Stimulus bietet hier einen vielversprechenden Mittelweg zwischen den beiden typischen Designs. Im Vergleich mit der Kombination aus Befragung und Inhaltsanalyse ist die Kausalattribution der Effekte zu den Inhalten des medialen Stimulus besser abgesichert, wenn die Medienrezeption mit der Rezeption des TV-Duells gleichgesetzt wird. Einschränkungen müssen aber vor allem bei Inferenzen auf die Wirkungen von Botschaften außerhalb des TV-Duells gemacht werden. Form und Inhalte einer TV-Debatte sind der Analogie der Miniaturwahlkämpfe zum Trotz natürlich nicht mit allen rezipierten Inhalten gleichzusetzen. Im Vergleich zu den häufig in Experimentalstudien eingesetzten künstlichen Stimuli verspricht eine TV-Debatte jedoch höhere externe Validität. Dieses Argument leitet sich direkt aus der Analogie der TV-Debatten als Miniaturwahlkämpfe ab. Der Stimulus ist ein bedeutender Bestandteil des echten Medienwahlkampfs, dessen Inhalte von den wichtigsten Akteuren gestaltet werden, und der eine große Zahl der Wählerinnen und Wähler erreicht. Das komprimierte zeitliche Format und die relativ starke Standardisierung der Kandidatenauftritte durch die Regeln der Debatte erlauben die Untersuchung der Rezeption im unmittelbaren zeitlichen Umfeld bzw. sogar während der öffentlichen Ausstrahlung. Die Orientierung der Datenerhebung an der Live-Sendung des Stimulus macht jedoch das ohnehin schwierige Unterfangen, bestimmte Merkmale eines Medienstimulus so zu variieren, dass die externe Validität zumindest halbwegs gewahrt bleibt, fast unmöglich.[5] Die interne Validität als Möglichkeit des kausalen Nachweises von Effekten (vor allem bestimmter Merkmale) des Debatteninhalts ist daher geringer als im Experimentaldesign.

Vor dem Hintergrund dieser Überlegungen nehmen TV-Duell-Studien, in denen die Bewertungen der Kandidaten rezeptionsbegleitend während der Debatte erfasst werden, eine besondere Rolle ein. In diesen Studien ist es möglich, die Reaktionen der Rezipienten auf Variationen der Inhalte innerhalb der TV-Debatte zu untersuchen (Reinemann & Maurer, 2007a, S. 23-34). Auch wenn

5 Zwei sinnvolle Ausnahmen sind hier denkbar: Zum einen kann der Modus für die Präsentation des gesamten Stimulus experimentell variiert werden, z.B. audiovisuell vs. nur auditiv (Faas & Maier, 2004b) oder HDTV vs. Nicht-HDTV (Bos, van Doorn & Smanik, 2012). Zum anderen kann eine Kontrollgruppe eingesetzt werden, die statt des TV-Duells einen nicht politischen Stimulus sieht, um Effekte des gesamten TV-Duells nachzuweisen.

dadurch nicht die interne Validität eines randomisierten Experiments erreicht wird, lässt sich durch eine Verknüpfung der Charakteristika des Stimulus und den darauf folgenden Bewertungen durch die Rezipienten genauer bestimmen, von welchen Inhalten die Effekte ausgehen. Auch die externe Validität bezüglich Schlussfolgerungen auf die Wirkungen anderer medialer Stimuli ist hier größer, da eine Übertragung der Befunde auf medial vermittelte (politische) Botschaften möglich ist, deren Charakteristika nur einigen Ausschnitten aus der Debatte gleichen.

Ableitungen für Wirkungsstudien zu TV-Debatten mit rezeptionsbegleitender Messung der Kandidatenbewertung

In Wirkungsstudien zu TV-Debatten werden die Effekte untersucht, die von der Rezeption der Inhalte eines der bedeutendsten Ereignisse des Wahlkampfs ausgehen. Dabei versuchen die Forscherinnen und Forscher häufig, die Befunde nicht nur in Bezug auf die TV-Debatte als eigentlichen Untersuchungsgegenstand zu interpretieren, sondern sie auch darüber hinaus als Indikatoren für die Effekte anderer Stimuli aufzufassen. Für diese Inferenzschlüsse sind Bewertungen der Kandidaten, die während der Rezeption der Debatte erfasst werden, besonders hilfreich, da hier der zeitliche Bezug zu einzelnen Inhalten der Debatte hergestellt werden kann. Im Folgenden präzisieren wir anhand abstrakter Forschungsfragen, welche Inferenzschlüsse auf Basis dieser Messungen möglich sind und welche Voraussetzungen die Studien dafür erfüllen müssen.[6]

Zunächst können die rezeptionsbegleitend gemessenen Kandidatenbewertungen ohne die Absicht eines Inferenzschlusses für die Personen-Stichprobe der Studie und die vorliegende Debatte deskriptiv betrachtet werden. Die Frage lautet dann:

> Frage 1: Welche unmittelbaren Bewertungen geben *diese Zuschauer* in Abhängigkeit von den Inhalten *dieser Debatte* ab?

Eine solche deskriptive Betrachtung stellt fast immer den ersten Schritt einer Analyse dar. Wir wollen zunächst sehen, welche Informationen die Daten unserer Stichprobe enthalten. Auch können wir Vergleiche anstellen, welche Inhalte von den Zuschauern in der Stichprobe besser oder schlechter bewertet werden,

6 Grundsätzlich setzten die folgenden Überlegungen voraus, dass die rezeptionsbegleitende Messung der Kandidatenbewertungen reliabel und valide erfolgt. Den methodologischen Forschungsstand zu dieser Frage stellen wir in Kapitel 3.3 dar. Die Reliabilität und Validität unserer eigenen RTR-Messungen untersuchen wir in Kapitel 4.2.

oder an welchen Stellen sich die Bewertungen von Teilstichproben der Zuschauer unterscheiden. Doch die reine Deskription der Daten ist in aller Regel nicht das eigentliche Forschungsziel. Vielmehr wollen wir Inferenzschlüsse ziehen, um allgemeine Aussagen zu treffen bzw. zu überprüfen. Ausgehend von den rezeptionsbegleitenden Bewertungen der Kandidaten sind zwei Richtungen der Inferenzen denkbar: Zum einen wollen wir fast immer etwas über die Bewertungen durch andere Zuschauer aussagen. Hierbei handelt es sich um den in Wirkungsstudien üblichen Inferenzschluss auf andere Rezipienten außerhalb der Stichprobe. Zum anderen sollen die Befunde der TV-Duell-Studie – wie oben ausführlich diskutiert – zusätzlich auch Verallgemeinerungen hinsichtlich der Wirkungen von anderen Inhalten der Kampagnen- und Medienkommunikation zulassen. In diesem Fall wird ein Inferenzschluss auf andere Stimuli angestrebt.

Greifen wir die bereits diskutierten möglichen Forschungsziele einer TV-Duell-Studie auf, so sind wir zunächst daran interessiert, etwas über die Bedeutung der Debatte im Kontext des laufenden Wahlkampfs auszusagen. Bezüglich der unmittelbaren Bewertungen der Kandidaten ist die Frage zu beantworten:

> Frage 2: Welche unmittelbaren Bewertungen geben *andere Zuschauer* in Abhängigkeit von den Inhalten *dieser Debatte* ab?

Gefragt wird hier nach einem Inferenzschluss auf die Grundgesamtheit aller Personen, die das TV-Duell verfolgt haben, bzw. auf Teilgruppen des gesamten Duellpublikums. Nur wenn ein solcher Inferenzschluss möglich ist, können wir anhand der Studie überhaupt etwas über die gesellschaftliche Relevanz der Bewertung der Kandidaten im TV-Duell aussagen. Natürlich ist es klar, dass eine vollkommene Repräsentativität für das gesamte Debattenpublikum in solchen Rezeptionsstudien nicht erreicht werden kann. Sie ziehen meist keine Zufallsstichprobe, unterliegen durch den hohen Aufwand auf Seiten der Probanden einer starken Selbstselektion und sind nicht zuletzt durch die notwendige technische Ausstattung lokal an die Erhebungsstandorte gebunden (vgl. dazu auch die in Kapitel 3.4 berichteten Stichproben der TV-Duell-Studien mit rezeptionsbegleitenden Messungen). Wenn wir diese Kriterien zugrunde legen, können wir uns auf Basis dieser Studien jeden Inferenzschluss sparen und die Ergebnisse strikt im Hinblick auf die Stichprobe interpretieren. Trotz allem werden die Ergebnisse der TV-Duell-Studien auch als Indikator für die Debattenwahrnehmung in der Grundgesamtheit betrachtet. Um die Indikatorfähigkeit der Studien zu verbessern, werden fast immer quotierte Stichproben hinsichtlich der politischen Voreinstellungen gebildet und Gewichtungen der Stichprobe hinsichtlich der relevanten Grundgesamtheit vorgenommen. Wenn

wir unter diesen Bedingungen die Indikatorfunktion der Stichprobenergebnisse akzeptieren, können wir sie auf die Grundgesamtheit übertragen. Die dafür verwendeten statistischen Verfahren müssen dann der Unsicherheit, die sich bei einem Inferenzschluss von der Stichprobe der Studienteilnehmer auf die Grundgesamtheit aller Zuschauer der Debatte ergibt, Rechnung tragen.

In einem weiteren Schritt interessieren wir uns für die Indikatorfunktion der unmittelbaren Kandidatenbewertungen für die Bewertung ähnlicher politischer Botschaften der Kandidaten und ihrer Parteien im laufenden Wahlkampf:

> Frage 3: Welche Bewertungen geben *die Wählerinnen und Wähler* zu *anderen Botschaften* ab, die den Aussagen der Kandidaten in dieser Debatte in bestimmter Hinsicht ähnlich sind?[7]

Die Beantwortung dieser Frage erfordert sowohl einen Inferenzschluss auf andere Personen als auch einen Inferenzschluss auf andere Stimuli. Für den Schluss auf die Grundgesamtheit der Wählerinnen und Wähler gelten im Prinzip dieselben Überlegungen wie zum Schluss auf die Grundgesamtheit anderer Zuschauer. Allerdings ist zu problematisieren, dass die Verzerrung der Stichprobe einer TV-Duell-Studie zugunsten der politisch Interessierten hier stärker ins Gewicht fallen dürfte, da sich der Inferenzschluss auch auf Wählerinnen und Wähler beziehen soll, die die Debatte nicht rezipiert haben.

Der Inferenzschluss von den Inhalten des TV-Duells auf andere, ähnliche Botschaften im Wahlkampf beruht auf zwei analytischen Schritten. Zuerst verstehen wir die Inhalte des TV-Duells als eine Abfolge vieler Stimuli. Papastefanou (2013, S. 12) hat diesen Umstand treffend beschrieben:

> Watching a TV debate means that respondents are exposed not to a single stimulus, but to a stream of audio and visual stimuli, as they are expressed voluntarily or involuntarily by the participants with verbal, facial, gestures and posture expressions.

Abstrakt ausgedrückt können wir aus diesem „stream of stimuli" analytisch einzelne Stimuli mit den Merkmalen, die für unser Forschungsinteresse relevant sind, extrahieren. Dann untersuchen wir deren Effekte auf die unmittelbaren Bewertungen der Kandidaten. Konkret könnten so beispielsweise alle Aussagen der Kandidaten zur Schulpolitik in der TV-Debatte identifiziert werden. Dann würde mit den Daten der rezeptionsbegleitenden Messung untersucht, wie die Kandidaten infolge schulpolitischer Aussagen bewertet werden.

7 Zur Verbesserung der Lesbarkeit verwenden wir im Folgenden Begriffe wie „Aussagen" oder „Botschaften". Auch wenn diese Begriffe vor allem auf die verbale Ebene der Kandidatenauftritte hinweisen, sind auch immer die nonverbalen Elemente gemeint.

Im zweiten Schritt wollen wir darauf schließen, wie Wahlkampfbotschaften, die diesen Stimuli in Hinblick auf die Merkmale, nach denen sie extrahiert wurden, ähnlich sind, durch die Wählerinnen und Wähler bewertet werden. Konkret wollen wir also aus den Bewertungen der Kandidaten infolge ihrer schulpolitischen Aussagen in der Debatte ableiten, wie die schulpolitischen Wahlkampfbotschaften von der Wählerschaft bewertet werden. Um die Validität dieses Inferenzschlusses zu beurteilen, müssen wir einschätzen, für welche Grundgesamtheit(en) der politischen Botschaften die Stichprobe der aus dem TV-Duell extrahierten Stimuli als repräsentativ gelten kann. Dieser Überlegung liegt die Idee zugrunde, dass die konkreten Aussagen der Kandidaten in der Debatte (teils strategisch, teils unbewusst ausgewählte) Realisationen aus der Grundgesamtheit aller Kampagnenbotschaften, für die sie und ihre Partei in diesem Wahlkampf stehen, sind (vgl. zur Auffassung manifester Textinhalte als Realisationen aus einer Grundgesamtheit an latenten Inhalten K. Benoit, Laver & Mikhaylov, 2009; Scharkow, 2012, S. 22-25). Wie es um die Repräsentativität bestimmter Aussagen in einem Duell für eine Grundgesamtheit der Wahlkampfbotschaften bestellt ist, lässt sich kaum nach einem harten Kriterium prüfen und auch nur im Einzelfall – für ein bestimmtes Duell in einem bestimmten Wahlkampf – qualitativ einschätzen. Für eine grundsätzliche Indikatorfunktion, wie sie auch in der Analogie der Debatten als Miniaturwahlkämpfe enthalten ist, sprechen die Befunde verschiedener Inhaltsanalysen. So kann gezeigt werden, dass in den Duellen in der Regel die wichtigen Themen eines Wahlkampfs vorkommen (z.B. Bachl, Käfferlein & Spieker, 2013b; Maurer, 2007) und dass die Rhetorik anderer Kommunikationsmaßnahmen einer Kampagne mit der Rhetorik der Kandidaten in den Debatten übereinstimmt (Hart & Jarvis, 1997). Beispielsweise wäre zu fragen, inwiefern die Aussagen der Kandidaten zur Schulpolitik im TV-Duell sich mit den Botschaften der Kandidaten und ihrer Parteien im Wahlkampf decken. Beeinträchtigt wird die Validität solcher Inferenzschlüsse dagegen durch die Tatsache, dass vor allem die Präsentationsform der Aussagen im Format des TV-Duells eher speziell erscheint und nicht direkt mit der Rezeption anderer Kampagnenkommunikation oder anderer massenmedialer Stimuli gleichgesetzt werden kann.

Wenn ein solcher Inferenzschluss nach allen Abwägungen gerechtfertigt ist, kann er mit geeigneten Analyseverfahren statistisch umgesetzt werden. Wenn sich die Inferenz nur auf die Bewertung eines Stimulus bezieht (Im Beispiel: Es gibt nur eine schulpolitische Aussage eines Kandidaten), dann hat dies für die Wahl des statistischen Verfahrens keine Konsequenzen. Allerdings müssen wir bedenken, dass nach der Logik, dass der Stimulus eine Stichprobe aus der Grundgesamtheit aller möglichen Stimuli ist, eine Inferenz auf der Basis von n = 1 gezogen wird. Eine Verallgemeinerung über eine exakte Wiederholung

dieses Stimulus hinaus wäre kaum angemessen. Sinnvoller sind Inferenzen, wenn aus dem TV-Duell mehrere Stimuli mit vergleichbaren Merkmalen extrahiert werden. Wenn beispielsweise mehrere Aussagen eines Kandidaten zu seiner Schulpolitik in der Debatte vorkommen, können die Bewertungen während dieser Aussagen als mehrere manifeste Indikatoren für die latente Bewertung der schulpolitischen Position des Kandidaten durch die Zuschauer bzw. Wähler gelten. In diesem Fall dürfte die Inferenz auf die Effekte eines Merkmals des Stimulus in weiteren, ähnlichen Stimuli außerhalb der Stichprobe valider sein als in einem (einfachen) Experiment, da der Inferenzschluss eben nicht nur auf dem Stichprobeneffekt eines einzelnen Stimulus basiert.

Diese wiederholte Messung der Reaktionen auf ein Merkmal der TV-Debatte muss dann aber auch adäquat in einem inferenzstatistischen Verfahren modelliert werden, um zu einem validen Inferenzschluss auf Stimuli mit vergleichbaren Merkmalen außerhalb des TV-Duells zu kommen. Ein solches Verfahren muss – wie auch beim üblichen Inferenzschluss auf Personenebene – die Varianz zwischen den einzelnen Messungen ebenso berücksichtigen wie die Zahl der bewerteten Stimuli in der Stichprobe. Da für die Beantwortung der Frage sowohl Inferenzen auf andere Personen als auch auf andere Inhalte notwendig sind, müssen die Verfahren in der Lage sein, beide Quellen der statistischen Unsicherheit zu berücksichtigen.

Lösen wir uns vom Kontext des Wahlkampfs, in dem eine TV-Debatte stattfindet, können wir zuerst nach der generellen Wirkung bestimmter Merkmale des Debatteninhalts auf die Bewertung von Kandidaten während eines TV-Duells fragen:

> Frage 4: Welche Bewertungen geben *andere Zuschauer* zu *Aussagen in einer anderen TV-Debatte* ab, die den Aussagen der Kandidaten in dieser Debatte in bestimmter Hinsicht ähnlich sind?

Dies ist die abstrakt formulierte Forschungsfrage, die uns in der kommunikationswissenschaftlichen Forschung zur Erklärung der unmittelbaren Bewertung der Kandidaten in TV-Duellen vorrangig interessiert. Denn es geht uns in der Regel nicht darum, zu erklären, warum ein bestimmter Kandidat während eines bestimmten TV-Duells zu einem Zeitpunkt positiv oder negativ bewertet wird. Vielmehr möchten wir untersuchen, ob bestimmte Merkmale des Auftritts eines Kandidaten generell zu einer besseren oder schlechteren unmittelbaren Bewertung führen. Konkret könnte es beispielsweise – wie auch in der vorliegenden Arbeit – um die Frage gehen, ob Angriffe oder Verteidigungen der Kandidaten anders bewertet werden als Selbstpräsentationen.

Schließlich können wir noch einen Schritt weitergehen und uns in den Fragestellungen völlig von der Situation des TV-Duells lösen, um anhand der

rezeptionsbegleitend gemessenen Kandidatenbewertungen allgemeine Fragen der Rezeptions- und Wirkungsforschung zu untersuchen:

> Frage 5: Welche Bewertungen geben *andere Rezipienten* zu *anderen massenmedial vermittelten (politischen) Aussagen* ab, die den Aussagen der Kandidaten in dieser Debatte in bestimmter Hinsicht ähnlich sind?

In diesem Sinne argumentiert beispielsweise Nagel (2012, S. 89-95), dass sich das hinsichtlich der visuellen Gestaltung relativ standardisierte Format des TV-Duells dazu eignet, den relativen Einfluss des nonverbalen Verhaltens von Politikern auf deren Bewertung anhand eines extern validen Stimulus zu untersuchen. Bei dieser Untersuchung ist der Umstand, dass die Kandidaten in einem TV-Duell bewertet werden, nur eine Randbedingung. Interessant ist, inwiefern Merkmale des nonverbalen Verhaltens der hier auftretenden Politiker zur Erklärung ihrer unmittelbaren Bewertung beitragen.

Die Fragen 4 und 5 erfordern wiederum zwei Inferenzschlüsse: von den Rezipienten des TV-Duells in der Stichprobe auf andere Personen sowie von den Inhalten des TV-Duells auf andere Inhalte. Auch hier treffen damit die bereits ausführlich dargestellten Überlegungen grundsätzlich zu. Allein die Referenzrahmen für die Beurteilung der Validität der Inferenzen verändern sich. Für den Schluss auf andere Personen ist ganz allgemein keine Repräsentativität im Hinblick auf eine enger definierte Grundgesamtheit notwendig. Um die Annahme zu prüfen, dass Angriffe und Verteidigungen unterschiedliche Bewertungen verursachen, reicht eine beliebige Stichprobe aus. Wenn der Prozess, der zu den unterschiedlichen Bewertungen führt, allgemeingültig ist, dann muss er sich auch in jeder beliebigen Stichprobe zeigen (Hayes, 2005, S. 41). Ein Signifikanztest besagt hier lediglich, dass ein Unterschied bei anderen Personen aus derselben, nicht näher definierten Grundgesamtheit ebenfalls zu finden wäre. Die Zusammensetzung der Stichprobe ist allerdings für die Untersuchung von Medienwirkungen in der politischen Kommunikation nicht völlig unerheblich, da häufig politische Prädispositionen die Wirkungen moderieren (z.B. Iyengar & Simon, 2000; Zaller, 1992). Solche Interaktionseffekte zwischen politischen Voreinstellungen und Merkmalen der Stimuli können nur untersucht werden, wenn auch genügend Personen mit unterschiedlichen Voreinstellungen in der Stichprobe bzw. Grundgesamtheit enthalten sind.

Der Schluss auf die Aussagen in anderen TV-Duellen, wie er zur Beantwortung der vierten Frage erforderlich ist, ist dann valide möglich, wenn wir Effekte von Merkmalen untersuchen, die sich konzeptionell gut über den Kontext des untersuchten TV-Duells hinaus verallgemeinern lassen. Typische Beispiele

hierfür sind Effekte von Angriffen, Verteidigungen und Selbstpräsentationen, die als Merkmale der Auseinandersetzung von mehreren Kandidaten in einer Debatte so allgemein definiert sind, dass sich die Aussagen jeder Debatte nach diesem Schema klassifizieren lassen (vgl. dazu ausführlicher den Exkurs zu Relationen in Kapitel 3.4, S. 78ff.). Es muss jedoch bedacht werden, dass die Effekte der ausgewählten Merkmale immer unter den Randbedingungen des jeweiligen TV-Duells untersucht werden. So ist es durchaus denkbar, dass z.B. die Effekte von Angriffen von der Persönlichkeit des angreifenden Kandidaten abhängen. Wird eine zuvor getroffene Annahme zum Effekt von Angriffen nicht gestützt, so ist die Annahme für diese Randbedingung nicht haltbar und muss in ihrer Allgemeingültigkeit abgeschwächt werden. Anhand einer Studie zu einer einzigen Debatte, in der nur wenige Kandidaten antreten, ist es aber nicht möglich, Interaktionen mit solchen Randbedingungen systematisch zu untersuchen.

Ob von den inhaltlichen Merkmalen einer TV-Debatte valide Inferenzschlüsse auf Merkmale massenmedial vermittelter (politischer) Aussagen im Allgemeinen möglich sind, kann kaum generell beantwortet werden. So ist beispielsweise die ausführliche Argumentation von Nagel (2012, S. 89-95), warum sich das TV-Duell-Format zur Bestimmung des relativen Einflusses verbaler und nonverbaler Merkmale auf die Bewertung von Politikern bei Fernsehauftritten eignet, durchaus überzeugend. Weniger geeignet scheinen dagegen beispielsweise Inferenzschlüsse von den Bewertungen der schulpolitischen Aussagen eines Kandidaten auf die Wirkung eines redaktionellen Nachrichtenbeitrags über seine Schulpolitik. Daher sollte die Übertragbarkeit im Einzelfall jeweils explizit deutlich gemacht werden, wenn solche stark verallgemeinernden Inferenzschlüsse angestrebt werden.

Schließlich gilt auch für die statistischen Verfahren, mit denen die Inferenzschlüsse zu den Fragen 4 und 5 umgesetzt werden, dass sie entsprechend der Argumentation zu Frage 3 die Unsicherheiten des Schlusses auf andere Personen und des Schlusses auf andere Inhalte adäquat abbilden müssen.

Zusammenfassung

Insgesamt können wir festhalten, dass TV-Duell-Studien schon für sich alleine wichtige Forschungsgegenstände untersuchen, der empirischen Kommunikationsforschung aber auch darüber hinaus eine ganze Reihe von Möglichkeiten zum Beantworten unterschiedlicher Fragestellungen bieten. Besonders groß ist dabei das Potenzial von Studien, in denen die Bewertungen der Kandidaten durch die Rezipienten kontinuierlich während der Debatte erfasst werden. Durch diese Messungen wird eine Verknüpfung bestimmter Debatteninhal-

te mit den Kandidatenbewertungen möglich, was die interne Validität eines Wirkungsbefunds erhöht. Die externe Validität ist zumindest im Vergleich zu typischen Experimenten mit künstlichem Stimulusmaterial höher, da ein echter Bestandteil des Medienwahlkampfs untersucht wird. Reizvoll ist für die Kommunikationswissenschaft auch die Möglichkeit, Inferenzschlüsse auf eine Population von Rezipienten *und* auf eine Population von Inhalten ziehen zu können. Wie valide diese Inferenzschlüsse jeweils inhaltlich sind, muss für jede einzelne Studie und Analyse bewertet werden. Bei dieser Einschätzung ist immer die Gültigkeit beider Inferenzschlüsse zu beachten:

1. Kann von der Personen-Stichprobe der Studienteilnehmer auf die Grundgesamtheit der Personen geschlossen werden, für deren Bewertungen der Kandidaten wir uns interessieren?

2. Kann von der Stichprobe der aus der untersuchten Debatte extrahierten inhaltlichen Stimuli auf die Grundgesamtheit der Inhalte geschlossen werden, für deren Wirkung wir uns interessieren?

Voraussetzung für die Gültigkeit der Befunde ist dabei immer, dass statistische Verfahren zum Einsatz kommen, die die Unsicherheit beim Schluss von den Stichproben auf beide gewünschten Grundgesamtheiten – die der Personen und die der Inhalte – angemessen berücksichtigen. Nur dann kann das große Potenzial solcher Studien wirklich ausgeschöpft werden.

3 RTR-Messungen in der Kommunikationsforschung

In diesem Kapitel wollen wir einen Eindruck von den Einsatzmöglichkeiten rezeptionsbegleitender Messungen in der Kommunikationsforschung vermitteln. Dank reichlicher einführender Literatur können wir uns auf einige für die vorliegende Arbeit zentrale Aspekte beschränken. Überblicksartikel haben zuletzt Fahr (2008), Maurer (2013b), J. Maier (2013) und Ottler (2013) vorgelegt. Darüber hinaus ist der Grundlagentext von Biocca et al. (1994) weiterhin als zentraler Einstieg in das Thema zu empfehlen. Einen Sammelband mit methodologisch orientierten Beiträgen und Anwendungsbeispielen haben J. Maier, Maier, Maurer, Reinemann und Meyer (2009) herausgegeben. Schließlich findet sich ein Bericht über die historischen Anfänge von RTR-Messungen in der Radio-Programmforschung der 1930er Jahre mit dem „Lazarsfeld-Stanton Program Analyzer“ bei Levy (1982). Die Weiterentwicklungen von diesem Zeitpunkt bis in die 1990er Jahre beschreibt Millard (1992).

Im Folgenden zeigen wir zunächst, welche technischen Möglichkeiten zur Durchführung von RTR-Messungen zur Verfügung stehen. Anschließend stellen wir einige typische Charakteristika von Studien mit RTR-Messungen[8] in der Kommunikationswissenschaft und verwandten Feldern vor. Dann fassen wir die methodologische Auseinandersetzung mit RTR-Messungen zusammen. Im dritten Teilkapitel gehen wir ausführlich auf die empirischen Befunde zu einer für die vorliegende Arbeit zentralen Frage ein: Wie können die mit RTR rezeptionsbegleitend gemessenen Kandidatenbewertungen in TV-Duellen erklärt werden?

3.1 Technik von RTR-Messungen

Die Ausrüstung zur Durchführung einer RTR-Studie besteht aus Anwenderperspektive aus zwei Komponenten: (mindestens) einem Eingabegerät sowie einem zentralen System zum kontinuierlichen Sammeln und Speichern der

8 Im Folgenden bezeichnen wir Studien, in denen RTR-Messungen eingesetzt werden, verkürzt als RTR-Studien.

eingegebenen Daten.[9] Aus Sicht eines Sozialwissenschaftlers, der lediglich RTR-Messungen durchführen will, spielt die Funktionsweise des Systems zum Sammeln und Speichern der Daten im Detail nur eine untergeordnete Rolle. Das System muss gewährleisten, dass die Angaben aller Eingabegeräte in vorher bestimmten äquidistanten Zeitintervallen abgerufen und mit einer Gerät- und Zeitmarke gespeichert werden. Dadurch wird sichergestellt, dass jede Messung einem bestimmten Probanden sowie einem bestimmten Zeitpunkt des rezipierten Stimulus zugeordnet werden kann.

Die Auswahl und Gestaltung des Eingabegeräts, mit dem ein Proband während der Rezeption kontinuierlich Auskunft über das Konstrukt gibt, dessen Veränderung im Zeitverlauf für die Forschungsfrage von Interesse ist, ist dagegen nicht nur von technischer Relevanz. Diese Entscheidung entspricht – gemeinsam mit der Vorgabe der Bedeutung der RTR-Skala (vgl. Kapitel 3.2) – konzeptionell der Gestaltung des Fragebogens in einer schriftlichen Befragung und hat damit Konsequenzen für den Informationsgehalt und das Skalenniveau der erhobenen Rohdaten. Hieraus leitet sich auch ab, wie die Daten weiterverarbeitet (transformiert), analysiert und interpretiert werden dürfen. Ottler (2013, S. 115-117) beschreibt die Bedienungsweise der drei verbreitetsten Typen von Eingabegeräten: *Push-Button-Geräte* mit Druckknöpfen oder Tasten für jede Ausprägung der RTR-Skala; *Dials*, bei denen die Skalenpunkte in einem Kreis angeordnet sind und mit einem Drehregler ausgewählt werden; *Slider*, bei denen die Skalenpunkte auf einer Linie angeordnet sind und mit einem Schieberegler ausgewählt werden. Zudem sind *Joysticks*, bei denen die Ausprägungen der Skala durch die Stellung des Hebels relativ zur Mittelstellung ausgewählt werden, zu nennen. Sie kommen unseres Wissens jedoch nur vergleichsweise selten zum Einsatz (z.B. Ramanathan & McGill, 2007; Wünsch, 2006b). Die drei in den gesichteten Studien am häufigsten eingesetzten Gerätetypen sind in Abbildung 3.1 schematisch dargestellt.

J. Maier (2013, S. 173-174) nennt unter anderem die Kategorien Skalenniveau der Messung, Modus der Messung und simultan erfassbare Dimensionen, um die Vor- und Nachteile von typischen Konfigurationen der Eingabegeräte

9 Der in den 1980er Jahren entwickelte „Warmth monitor" (Aaker, Stayman & Hagerty, 1986) nutzte anfangs noch ein einfaches Blatt Papier, auf dem die Probanden während der Rezeption ihre Wertung mit einem Stift im Zeitverlauf einzeichneten. In den meisten hier vorgestellten Studien handelt es sich bei den Eingabegeräten um Hardware, also Tastaturen, Regler oder andere Geräte, die von den Probanden direkt physisch bedient werden. Es sind jedoch auch Softwareumsetzungen möglich, in denen die Eingabegeräte auf einem Bildschirm angezeigt und mit Tastatur oder Computermaus bedient werden (z.B. Iyengar, 2011; Jarvis, 2012; Kercher et al., 2012; Wolf, 2010). Für die folgenden Überlegungen ist diese Unterscheidung jedoch zunächst irrelevant, da die Gestaltung der virtuellen Eingabegeräte der verbreiteten RTR-Hardware nachempfunden ist.

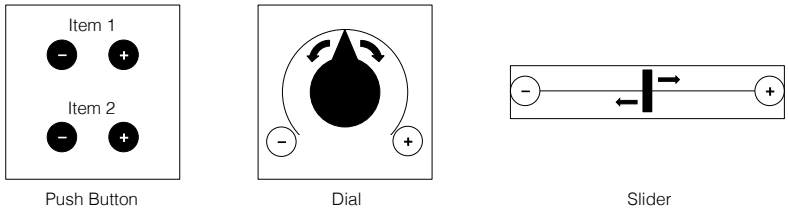


Abbildung 3.1: Schematische Darstellung der drei verbreitetsten Gerätetypen

Push-Button und Dial einzuschätzen. Das Schema kann auch auf die in der Regel gewählten Konfigurationen von Slider und Joysticks angewendet werden.[10] Das Skalenniveau der Messung bezieht sich auf die Rohdaten, die bei der Durchführung der RTR-Messung für jedes Eingabegerät erzeugt werden. Push-Button-Geräte erzeugen lediglich dichotome Rohdaten. Für jedes Zeitintervall (z.B. jede Sekunde) wird für jede Taste gespeichert, ob sie gerade gedrückt ist oder nicht. Dials, Slider und Joysticks werden dagegen mit einer quasi-metrischen Skala versehen. Die Zahl der Skalenpunkte variiert dabei zwischen ordinalen siebenstufigen Skalen, bei denen die Ausprägungen auf dem Eingabegerät verzeichnet sind (z.B. Jackob, Petersen & Roessing, 2008; Maurer et al., 2007), bis hin zu visuell-analogen Skalen, die den Probanden nur End- und Mittelpunkte zeigen und bis zu 1001 Skalenpunkte erfassen (z.B. Fahr, 2006, 2009). Während mit Push-Button-Geräten also nur erfasst wird, ob eine bestimmte Angabe zu einem Zeitpunkt gerade vorliegt, wird mit den übrigen Geräten auch die Intensität einer Angabe zu einem Zeitpunkt erfasst. In dieser Hinsicht enthalten die Rohdaten der Push-Button-Geräte weniger Informationen als die Daten der übrigen Eingabegeräte.

Beim Modus der Messung unterscheidet J. Maier (2013, S. 174) zwischen dem „reset mode“ und dem „latched mode“ (vgl. auch Maurer, 2013b, S. 224). Der Reset Mode meint, dass jede Messung eine Aktivität bzw. Nicht-Aktivität des Probanden widerspiegelt. Dies ist meist bei einer Push-Button-Messung der Fall. In jedem Zeitintervall der Messung wird für jede Taste dichotom der Status „gedrückt“ bzw. „nicht gedrückt“ erfasst. Es kann also genau unterschieden werden, ob der Proband gerade bewusst auf den Stimulus reagiert oder nicht.

10 Wir beschränken uns hier bewusst auf die unseres Wissens verbreitetsten Konfigurationen der (Hardware-) Gerätetypen. Gerade bei RTR-Messungen mit Eingabegeräten, die als Software programmierbar sind, können ohne großen Aufwand auch andere Konfigurationen gewählt werden, wenn dies für das Forschungsinteresse zielführend ist.

Bei Dials oder Slider, die in der Regel im Latched Mode messen, behält das Eingabegerät den gewählten Skalenpunkt so lange bei, bis der Proband eine andere Ausprägung auswählt. Dieser Messwert wird in jedem Messintervall gespeichert. Dadurch lassen die Messwerte häufig mehrere Interpretationen zu. Wenn beispielsweise ein Proband im Laufe des Stimulus seine Angabe von einem neutralen Wert auf einen positiven Wert ändert, kann dies zunächst als ein positiver Effekt des Stimulus zu diesem Zeitpunkt interpretiert werden. Wenn nun im Anschluss über eine längere Zeit dieser positive Wert gemessen wird, so kann dies zweierlei bedeuten: Entweder will der Proband bewusst dauerhaft einen positiven Wert angeben – dann läge ein länger anhaltender positiver Effekt des Stimulus vor. Oder der Wert bleibt unverändert positiv, weil der Proband gerade noch keine neue Antwort geben möchte – dann läge kein Effekt des Stimulus vor, obwohl weiterhin positive Messungen vorliegen. In dieser Hinsicht sind Daten, die im Reset Mode (meist mit Push-Button-Geräten) erfasst werden, informationsreicher, da sie sich eindeutig der Intention des Probanden, zu einem Zeitpunkt eine Antwort zu geben, zuordnen lassen (J. Maier & Faas, 2009).[11]

Schließlich können mit Push-Button-Geräten rezeptionsbegleitend Antworten zu mehreren Items erfasst werden. Dies ist in Abbildung 3.1 angedeutet, indem zu zwei Items positive und negative Wertungen abgegeben werden können. Nach J. Maier (2013, S. 173) ist die maximale Zahl der simultan mit Push-Button-Geräten zu beantwortenden Items „durch den Nutzer und seine Fähigkeit gesetzt gleichzeitig mit mehreren Bewertungsdimensionen umzugehen“. Vor dem Hintergrund der Studien des Autors, die sich meist mit der unmittelbaren Bewertung von Kandidaten in TV-Debatten beschäftigen (z.B. Faas & Maier, 2004a), sind die „Bewertungsdimensionen“ unseres Erachtens besser als Bewertungsobjekte zu verstehen. So gibt es in den Studien zu TV-Duellen zwei Items, mit deren Ausprägungen jeweils ein positiver oder ein negativer Eindruck von Kandidat A oder Kandidat B wiedergegeben wird. In einem eigenen Methodenexperiment zur Bewertung von Politikern in einer Diskussionsrunde mit fünf Teilnehmern haben wir gute Ergebnisse mit fünf Push-Button-Items erzielt (Kercher et al., 2012, S. 8). In diesem Kontext werden alle Bewertungsobjekte nach derselben Dimension, einem allgemeinen (positiven oder negativen) Eindruck, bewertet. Ob tatsächlich eine Bewertung der Kandidaten nach mehreren Dimensionen (z.B. Sympathie und Kompetenz)

11 Wie Joysticks sich hinsichtlich des Modus der Messung einordnen lassen, ist unklar, da dieser Punkt in den Publikationen mit diesem Eingabegerät nicht thematisiert wird (Ramanathan & McGill, 2007; Wünsch, 2006a, 2006b). Nach der Funktionsweise eines handelsüblichen Joysticks für Computerspiele müsste die Messung im Reset Mode stattfinden, da sich diese beim Loslassen in eine neutrale Ausgangsstellung zurückbewegen.

durch dieselben Probanden in Echtzeit möglich wäre, ist offen. Unseres Wissens liegen bisher keine Studien vor, die eine solche Differenzierung von den Probanden gefordert haben. Nichtsdestotrotz ist es mit Push-Button-Geräten generell möglich, den Probanden mehrere Items vorzulegen.

Diese Möglichkeit besteht bei Dials, Slider und Joysticks, deren Skala in ihrer Logik einem einzelnen Rating-Item aus einem Papierfragebogen gleicht, nicht. Beim Einsatz dieser Eingabegeräte kann immer nur ein Item rezeptionsbegleitend gemessen werden.[12] Um diese Einschränkung zu umgehen, haben sich in der Forschungspraxis verschiedene Vorgehen etabliert. Wenn zwei verschiedene inhaltliche Konstrukte rezeptionsbegleitend erfasst werden sollen, besteht eine Möglichkeit darin, die Messungen der Konstrukte auf mehrere Probandengruppen zu verteilen. Beispielsweise lässt Wolf (2010) die Probanden in zwei Gruppen ihren Eindruck von der Kompetenz bzw. der Sympathie der Kandidaten in einem TV-Duell mit Schiebereglern wiedergeben (vgl. Kapitel 3.2 für weitere Beispiele). Alternativ können rezeptionsbegleitende apparative Messungen physiologischer Indikatoren zur Erfassung einer weiteren Dimension – z.B. der Aktivierung (Fahr, 2006; Früh & Fahr, 2006; Früh, 2010) – herangezogen werden. In Studien zur unmittelbaren Bewertung von Politikern in TV-Debatten – und so auch in der vorliegenden Arbeit – stellt sich weniger das Problem unterschiedlicher inhaltlicher Dimensionen. In aller Regel wird hier ein allgemeiner Eindruck von den Kandidaten erhoben. Die RTR-Skala soll aber auf mehrere Kandidaten, und damit auf mehrere Bewertungsobjekte, angewendet werden. In der Literatur finden sich zwei Möglichkeiten, dies mit Dial- oder Slider-Geräten zu erreichen.[13] Die erste Option lässt sich unabhängig von der Zahl der Kandidaten in einer TV-Debatte realisieren. Die Probanden werden angewiesen, die eindimensionale Rating-Skala ihres Eingabegeräts immer auf den Kandidaten anzuwenden, der gerade spricht. Wenn in einem TV-Duell nur zwei Kandidaten teilnehmen, kann auch eine Differentialskala des Eindrucks von beiden Kandidaten verwendet werden (erstmals eingesetzt von Maurer & Reinemann, 2003). Der eine Skalenendpunkt steht dann für einen positiven Eindruck von Kandidat A oder einen negativen Eindruck von Kandidat B, der

12 Grundsätzlich wäre es zwar denkbar, den Probanden mehrere Eingabegeräte dieser Typen zur Verfügung zu stellen, um damit mehrere Items abzubilden. Dies dürfte die Probanden aber ziemlich sicher kognitiv überfordern und wurde nach unserer Kenntnis bisher noch nicht empirisch getestet.

13 Vgl. für eine detaillierte Darstellung, welche Variante in welcher Studie gewählt wurde, den Forschungsstand in Kapitel 3.4.

andere Skalenendpunkt umgekehrt für einen positiven Eindruck von Kandidat B oder einen negativen Eindruck von Kandidat A.[14]

Zusammengefasst können wir feststellen, dass unterschiedliche Typen von Eingabegeräten zur Durchführung von RTR-Messungen existieren, die sich in wesentlichen Kriterien – dem Skalenniveau der Rohdaten, dem Modus der Messung und der Zahl der simultan zu erfassenden Items – unterscheiden. Hinsichtlich dieser Kriterien lassen sich die Push-Button-Geräte von den sich recht ähnlichen Dial- und Slider-Geräten abgrenzen. Eine empirische Antwort auf die Frage, welche Typen sich besser für welche Forschungsinteressen eignen, muss an dieser Stelle ausbleiben. Es existieren kaum vergleichende Studien mit mehreren Gerätetypen, die systematisch deren Vor- und Nachteile überprüfen. J. Maier (2013, S. 176-177) findet keine wesentlichen Unterschiede in der subjektiven Einschätzung von Push-Button- und Dial-Eingabegeräten durch die Probanden. In einem Methodenexperiment, in dem wir Software-Versionen von Push-Button-, Dial- und Slider-Geräten eingesetzt haben, schneiden die Push-Button-Geräte aus Sicht der Probanden etwas besser ab. Insgesamt werden jedoch alle drei Typen als geeignet empfunden (Kercher et al., 2012, S. 12). An dieser Stelle besteht sicherlich noch einiger Bedarf an methodologischer Forschung, die zukünftig vor allem durch das Aufkommen von Software-basierten RTR-Systemen mit flexibel gestaltbaren Eingabegeräten erleichtert werden dürfte (z.B. Kercher et al., 2012).

Bis dahin muss eine Abwägung, welcher Gerätetyp für welche Fragestellung empfohlen werden kann, weitgehend auf Erfahrungsberichten und theoretischen Überlegungen beruhen. J. Maier und Faas (2009) legen ausführlich dar, warum der Entscheidungsprozess der Probanden bei der Abgabe einer Wertung mit Push-Button-Geräten kognitiv weniger fordernd ist (vgl. auch Baggaley, 1987) und daher auch die Bewertung mehrerer Items während der Rezeption ermöglicht. Zudem spricht die aufgrund des Reset Mode eindeutige Interpretation der Daten für diese Eingabegeräte. Im Gegensatz dazu haben Dials und Slider den Vorteil der informationshaltigeren Rohdaten, da auch die Intensität der Wertung erfasst wird. Zur Einordnung von Joysticks lässt sich schließlich nur wenig sagen, da sie nur selten eingesetzt und kaum empirisch evaluiert wurden.

14 Eine weitere Möglichkeit, mehrdimensionale Konstrukte zu messen und dabei auch die Intensität des Urteils ähnlich einer Rating-Skala zu erfassen, bieten die in der Forschung zur Musikrezeption eingesetzten Software-basierten RTR-Eingabegeräte (z.B. Egermann, Nagel, Altenmüller & Kopiez, 2009; Grewe, Nagel, Kopiez & Altenmüller, 2007; Nagel, Kopiez, Grewe & Altenmüller, 2007; Schubert, 1999). Hier wird das emotionale Erleben mittels einer Computermaus auf einer zweidimensionalen Skalenfläche berichtet. Da hierzu allerdings die visuelle Aufmerksamkeit der Rezipienten ständig auf die Skala gerichtet sein muss, kommt diese Variante für kommunikationswissenschaftliche Studien mit audiovisuellen Stimuli nicht infrage.

Meist wird die Wahl des Eingabegeräts ohnehin von forschungspraktischen Zwängen bestimmt. Forscher müssen die RTR-Systeme nutzen, die ihnen zur Verfügung stehen. Dies gilt auch für die Studie, auf der die vorliegende Arbeit basiert. Hier haben wir Dials als Eingabegeräte genutzt. Die genaue Ausgestaltung der Messung wird in Kapitel 4 sowie in Bachl, Brettschneider und Ottler (2013b) ausführlich beschrieben und begründet. Zwangsläufig beziehen sich die Analyseverfahren, die wir in den Kapiteln 5 und 6 entwickeln, zunächst auf die Datenstruktur, die in einer Erhebung mit RTR-Dials erzeugt wird. Wegen der beschriebenen Ähnlichkeiten ist die Übertragbarkeit unserer Befunde auf die ebenfalls häufig eingesetzten Slider ohne große Anpassungen gegeben. Wie sich die Verfahren auf Daten, die mit Push-Button-Geräten erhoben werden, anwenden lassen, diskutieren wir im abschließenden Kapitel 7.

3.2 Typische Charakteristika von RTR-Studien

In diesem Teilkapitel zeigen wir zuerst, in welchen Teilbereichen der Kommunikationswissenschaft und nahestehenden Disziplinen RTR-Studien zu finden sind. Dabei gehen wir auch auf die damit verbundene Frage ein, welche Konstrukte in diesen Studien rezeptionsbegleitend gemessen werden. Anschließend stellen wir typische Forschungsdesigns von RTR-Studien vor. Auf Basis dieser typischen Designs können wir im Fazit (Kapitel 7) Empfehlungen abgeben, welche der in dieser Arbeit speziell für die Analyse von rezeptionsbegleitend gemessenen Kandidatenbewertungen in TV-Duellen diskutierten statistischen Verfahren sich in welchen anderen verbreiteten Forschungsdesigns einsetzen lassen.

Den folgenden Systematisierungen müssen drei Einschränkungen vorausgestellt werden: *Erstens* konzentrieren wir uns im Folgenden auf Studien, in denen die Rezeption von Medienstimuli untersucht wird. Dies sind (mit der Ausnahme von einigen experimentellen Manipulationen) audiovisuelle Stimuli, die in ihrer Gestaltung Ausschnitten aus Informations- und Unterhaltungsformaten gleichen und die irgendeine Art von verbal vermittelter Information enthalten. Nicht berücksichtigt werden damit zum einen Studien, in denen nicht massenmedial vermittelte audiovisuelle Stimuli untersucht werden. So lassen beispielsweise Dalakas (2006a, 2006b) sowie Lemmink und Mattsson (1998, 2002) Videos bewerten, in denen Verhaltensweisen von Servicekräften in Dienstleistungssituationen dargestellt sind. Die Psychologen Levenson und Gottman (1983) lassen Paare retrospektiv Videoaufzeichnungen ihrer eigenen Interaktionen beurteilen. Zum anderen vernachlässigen wir RTR-Studien zur Musikrezeption (z.B. Brittin, 1996; Egermann et al., 2009; Grewe et al., 2007;

Madsen, 1998; Nagel et al., 2007; Schubert, 1999, 2004). Diese Forschung bietet allerdings für einige Bereiche der Unterhaltungsforschung, in denen ebenfalls auf das emotionale Erleben während der Medienrezeption abgezielt wird, einige interessante Anknüpfungspunkte und sollte bei weiterer Forschung in diesem Feld beachtet werden.

Zweitens beschränken wir uns auf Studien, in denen die längsschnittliche Beschreibung und Erklärung der rezeptionsbegleitend gemessenen Konstrukte zumindest eine gewisse Beachtung finden. Dadurch entfallen zum einen Analysen, in denen alle RTR-Messungen während der Rezeption eines Stimulus zu einem einzelnen Wert zusammengefasst werden. Zum anderen verzichten wir auf den Bericht von Studien, die RTR-Messungen nur für längsschnittliche Pre-Tests von audiovisuellen Stimuli einsetzen (z.B. Bradley, 2007) oder die mit RTR-Messungen kontinuierliche *unabhängige* Variablen zur Variation von Merkmalen audiovisueller Stimuli erfassen (z.B. Lang, Sanders-Jackson, Wang & Rubenking, 2013; S. Lee & Lang, 2013; Sparks & Lang, 2010; Wang, Lang & Busemeyer, 2011; Wang, Morey & Srivastava, 2012; Wang, Solloway, Tchernev & Barker, 2012; Woltman Elpers, Wedel & Pieters, 2003).

Drittens erheben wir auch nach den genannten Einschränkungen keinen Anspruch auf eine vollständige Erfassung sämtlicher RTR-Studien. Im Gegenteil sind wir uns bewusst, dass unsere Aufmerksamkeit zu Gunsten der politischen Kommunikationsforschung und speziell zu Gunsten von Studien zu TV-Debatten verzerrt ist. Die Zahl der Beispielstudien zu einem Forschungsfeld, einem zu messenden Konstrukt oder einem typischen Forschungsdesign sollte daher keinesfalls als eine quantifizierende Relevanzzuschreibung zu bestimmten Charakteristika innerhalb der Grundgesamtheit aller RTR-Studien missverstanden werden.

Forschungsfelder und rezeptionsbegleitend gemessene Konstrukte

RTR-Messungen werden in der kommunikationswissenschaftlichen Rezeptions- und Wirkungsforschung, aber auch in benachbarten Disziplinen, eingesetzt, um während der Medienrezeption Eindrücke, Meinungen oder Emotionen zu erfassen (Biocca et al., 1994, S. 16). Im Vergleich zur klassischen Befragung wird die „kontinuierliche Befragung in Echtzeit“ (Maurer, 2013b, S. 219) wegen ihres hohen Aufwands und der mangelnden Verfügbarkeit der notwendigen Technik vergleichsweise selten verwendet. Die generell möglichen Anwendungsbereiche sind jedoch – wie auch bei der klassischen Befragung – weit gestreut. Einen beachtlichen Teil der Veröffentlichungen, in denen im Längsschnitt analysierte RTR-Messungen präsentiert werden, machen Studien zu TV-Debatten aus, die während der Ausstrahlung der Debatten durchgeführt werden (vgl. für die

Relevanz solcher Studien Kapitel 2, für einen Überblick über ihre Befunde Kapitel 3.4). Die unmittelbare Bewertung der Kandidaten in TV-Debatten wird zudem in Experimenten untersucht, in denen (manipulierte) Ausschnitte aus TV-Debatten als Stimuli präsentiert werden. So vergleicht Maurer (2009) die unmittelbaren Bewertungen in Abhängigkeit vom Präsentationsmodus (Audiovisuell vs. Audio vs. Video). Wolf (2010) interessiert sich für die Wirkung von in die Übertragung des Duells eingeblendeten RTR-Bewertungen anderer Zuschauer und Experten. Mit den RTR-Messungen wird in den TV-Duell-Studien meist der *allgemeine Eindruck* erfasst, den die Rezipienten im Moment von den Kandidaten haben. Eine Ausnahme ist die Arbeit von Wolf (2010), in der unterschiedliche Rezipientengruppen entweder ihren Eindruck von der *Sympathie* oder ihren Eindruck von der *Kompetenz* der Kandidaten angeben.

Ebenfalls um den Eindruck von öffentlichen Sprechern, jedoch hinsichtlich spezifischerer Bewertungsmaßstäbe und außerhalb der Situation eines TV-Duells, geht es in zwei weiteren Studien. Jackob et al. (2008) (vgl. auch Roessing, Jackob & Petersen, 2009) erfassen, für wie *überzeugend* die Rezipienten einen Vortrag zum Thema Globalisierung halten, bei dem Rhetorik und Gestik des Sprechers manipuliert wurden. Kercher (2013) misst zu (Ausschnitten von) verschiedenen Politikerreden und -interviews den Eindruck von der *Verständlichkeit* der Politiker.

In der politischen Kommunikations- und Wahlkampfforschung finden sich außerdem einige Arbeiten, die sich mit der unmittelbaren Bewertung von Wahlkampfspots beschäftigen. So lassen J. Maier und Maier (2007) bzw. M. Maier und Maier (2009) die Testzuschauer ihre unmittelbaren Eindrücke von zwei Wahlwerbespots zur Bundestagswahl 2005 bzw. zur Europawahl 2004 mit RTR-Reglern wiedergeben. Kaid (2009) stellt die Ergebnisse dreier Studien vor, in denen die unmittelbaren Rezipienteneindrücke von Wahlwerbespots aus den US-Präsidentschaftswahlkämpfen 1996, 2000 und 2004 erfasst werden. Schließlich untersuchen Iyengar, Jackman, Hahn und Lim (2010) die Reaktionen der Zuschauer auf 28 Spots, die 2006 in Wahlkämpfen um Sitze im US-Senat ausgestrahlt wurden. Die genannten Studien zur Bewertung von Wahlwerbespots setzen RTR-Skalen ein, auf denen der *allgemeine Eindruck* vom Stimulus von *negativ* bis *positiv* angegeben werden kann. Iyengar et al. (2010) fassen die Bedeutung dieser Skala noch etwas weiter und beziehen sie etwas ambivalent sowohl auf die Botschaften der Spots als auch auf die ausgelösten Gefühle der Rezipienten.

Neben den bisher beschriebenen Arbeiten, die sich im weiteren Sinne im Bereich der politischen Kommunikation verorten lassen, befassen sich RTR-Studien auch mit der Rolle von Emotionen und subjektiven Selbsteinschätzungen im Rezeptionsprozess. Bei Fahr (2009) geben unterschiedliche Zuschauer-

gruppen Auskunft darüber, in welchem Maße sie sich gerade durch Ausschnitte aus Talkshows *unterhalten* bzw. *informiert* fühlen. Ähnlich gehen einige Arbeiten vor, in denen die Valenz-Dimension des emotionalen Erlebens auf einer Skala von *unangenehm* bis *angenehm* während der Rezeption von gekürzten Spielfilmen und einer Dokumentation (Wünsch, 2006a), eines TV-Magazins (Fahr, 2006), eines Ausschnitts aus einem Spielfilm (Früh & Fahr, 2006) oder von Nachrichtenbeiträgen (Früh, 2010) erfasst wird. Fahr (2006), Früh und Fahr (2006) sowie Früh (2010) nehmen zusätzlich eine physiologische Messung der Hautleitfähigkeit als Indikator für die emotionale Aktivierung vor. Früh (2010) ergänzt eine weitere Probandengruppe, die per RTR-Messung angibt, für wie *relevant* sie den Inhalt der Nachrichtensendung gerade hält. Alle diese Arbeiten verfolgen unter anderem das Ziel, den Einfluss von Merkmalen der Stimuli auf emotionales Erleben und subjektive Selbsteinschätzung zu bestimmen.

Von Pape und Meyer (2011) analysieren, wie sich die *emotionale Betroffenheit* – operationalisiert mit einer Skala von *überhaupt nicht* bis *sehr betroffen* – beim Sehen einer Dokumentation über Kindesmissbrauch verändert. Die Psychologen Fredrickson und Kahneman (1993) vergleichen den Verlauf des allgemein als *negativer* bis *positiver Affekt* gemessenen Befindens während der Rezeption von zwölf emotionalen Clips, die in ihrer Länge manipuliert wurden. Ziel ist es, herauszufinden, ob die unterschiedlichen dynamischen Verläufe Einfluss auf die retrospektiv berichtete Emotion haben. Am Einfluss der technisch-formalen Gestaltung auf das *(Un-) Wohlbefinden* der Rezipienten sind Lambooij, Ijsselsteijn und Heynderickx (2011) interessiert. Sie untersuchen mit einer RTR-Skala von „bad“ bis „excellent comfort“ (S. 212), wie sich verschiedene Techniken des dreidimensionalen Fernsehens auswirken.

Zwei weitere Studien sind inhaltlich ebenfalls der Rezeptionsforschung zuzuordnen, messen jedoch mit RTR Eindrücke vom Stimulus anstatt subjektive Auskünfte über das Empfinden der Rezipienten. Die Arbeit der Wirtschaftswissenschaftler Ramanathan und McGill (2007) untersucht anhand des rezeptionsbegleitend gemessenen *allgemeinen Gefallens* einer TV-Show („dislike very much“–„like very much“, S. 509), ob der gemeinsame Medienkonsum zu ähnlichen Eindrücken vom Stimulus führt. Dazu werden die RTR-Messungen von Probanden, die sich in verschiedenem Maße gegenseitig beobachten konnten, miteinander verglichen. Shapiro und Chock (2003) lassen unterschiedliche Probanden sieben Ausschnitte aus unterhaltenden TV-Serien nach den *spezifischen Charakteristika* „typicality of people“, „typicality of events“ und „perceived reality“ (S. 181) in Echtzeit bewerten. Zwei weitere Gruppen geben für dieselben Stimuli während der Rezeption an, ob sie *interessant* sind bzw. ob sie gerade *gefallen*. Das Forschungsinteresse richtet sich dann auf die Zusammenhänge zwischen diesen Konstrukten.

Die bisher aufgezählten Studien dienen in erster Linie einem akademischen Erkenntnisinteresse. Solche oder ähnliche Analyse können darüber hinaus auch genutzt werden, um die untersuchten Stimuli zu optimieren. In der angewandten Programmforschung ist dies nach den Berichten einiger Autoren verbreitet (z.B. Biocca et al., 1994; Fahr, 2006; Maurer, 2013b). Da diese Studien nur selten publiziert werden, können wir sie nicht in dieser Literaturschau aufführen. Über zwei Anwendungen, die an der Optimierung der Untersuchungsgegenstände interessiert sind, kann hier abschließend berichtet werden. Zum einen liegen Studien vor, die Informationsmaterial aus Präventionskampagnen evaluieren. Einen sehr direkten Weg dazu wählen Baggaley et al. (1992). Sie lassen Probanden, die in einer Klinik für sexuell übertragbare Krankheiten rekrutiert wurden und damit aus der relevanten Zielgruppe stammen, sechs Public Service Announcements zum Thema HIV / AIDS in Hinblick auf ihre *Nützlichkeit* bewerten: „[T]he 'good' response would be appropriate whenever they [die Probanden, M.B.] saw or heard anything which they felt would persuade viewers to take AIDS seriously and of the need to take precautions against HIV infection" (S. 86). Nach einem ähnlichen Verfahren werden in einer Studie von Tedesco und Ivory (2009) drei Aufklärungsvideos für eine jüngere Zielgruppe zum Thema „Sexuell übertragbare Krankheiten" von Studierenden nach den Kriterien „informative value" und „anxiety-producing value" (S. 184) evaluiert. Einen indirekten Ansatz der Evaluation wählen dagegen Algie und Rossiter (2010) sowie Rossiter und Thornton (2004). Sie untersuchen, inwieweit Aufklärungsspots zu den Gefahren des zu schnellen Fahrens in der Lage sind, die Rezipienten wie beabsichtigt in einen *angespannten Zustand* zu versetzen.

Zum anderen werden auch in der kommerziellen Werbeforschung RTR-Studien zur Spot-Evaluation eingesetzt. Zu diesen Verfahren sind vor allem methodologische Beiträge publiziert. Eines der älteren Verfahren ist der in den 1980er Jahren entwickelte „Warmth monitor" (Aaker et al., 1986, S. 365), der erfassen soll, ob ein Werbespot bei den Rezipienten „warme", also *allgemein positive Gefühle* auslöst (vgl. auch Abeele & MacLachlan, 1994). Auf dieses Konstrukt beziehen sich auch Baumgartner et al. (1997), die jedoch noch *allgemeiner die Gefühle* der Zuschauer auf einer Skala von *negativ bis positiv* messen. Hughes (1992) lässt dagegen die Werbespots direkt bewerten und unterscheidet dabei zwischen den Dimensionen *Affekt* („unfavorable"–„favorable") und *Kognition* („useless"–„usefull", S. 66). Schließlich interessieren sich Woltman Elpers, Mukherjee und Hoyer (2004) für zwei spezifischere Dimensionen. Zwei Gruppen von Probanden bewerten Werbespots hinsichtlich ihrer *Überraschung* („not surprising at all"–„very surprising") bzw. ihres *Humors* („not funny at all"–„very funny", S. 594).

Zusammengefasst können wir feststellen, dass mit RTR-Messungen eine große Bandbreite unterschiedlicher Konstrukte rezeptionsbegleitend erfasst werden. Abstrahiert lassen sie sich hinsichtlich ihrer Bezugsobjekte in drei Gruppen einteilen. Zum einen werden die Eindrücke erfasst, die die Rezipienten im Moment vom *Stimulus insgesamt* haben. Zum zweiten gibt es RTR-Skalen, die den Eindruck von *bestimmten Bestandteilen des Stimulus* messen. In den berichteten Beispielen beziehen sich diese Skalen vor allem auf den Eindruck von Politikern, die im Stimulus vorkommen. Zum dritten werden die Rezipienten aufgefordert, mit der RTR-Technik Auskunft über ihr *eigenes Rezeptionserleben* zu geben, beispielsweise über ihre Emotionen.

Weiter sehen wir, dass sich die RTR-Messungen im Hinblick auf die Spezifität dessen, was genau von den Rezipienten angegeben werden soll, unterscheiden. Es finden sich sowohl bewusst sehr offen gehaltene Instruktionen, einen *allgemeinen Eindruck* bzw. ein *allgemeines Befinden* zu berichten, als auch recht *spezifische Vorgaben*. So fordern einige Forscher die Versuchsteilnehmer auf, ihren Eindruck vom Humor, vom Realismus des Dargestellten oder von der Verständlichkeit der Politiker wiederzugeben. Bezogen auf die Selbstwahrnehmungen finden sich z.B. Aufforderungen zum Bericht der subjektiven Informiertheit oder der emotionalen Betroffenheit.

Offen muss an dieser Stelle bleiben, ob die Vorgabe eines anderen Bezugsobjekts oder die Vorgabe eines allgemeinen bzw. spezifischen inhaltlichen Konstrukts von den Rezipienten tatsächlich valide umgesetzt wird. Biocca et al. (1994, S. 20) weisen darauf hin, dass die Skalengestaltung beschränkt wird durch „the respondent's ability to make the requested discrimination in real time". Können wir davon ausgehen, dass Rezipienten wirklich über einen längeren Zeitraum ausschließlich ihren Eindruck von der Kompetenz oder der Sympathie eines Politikers in Echtzeit wiedergeben? Leider liegt zu dieser Frage unseres Wissens bisher kaum systematische methodologische Forschung vor. Einzige Ausnahme ist die Evaluation des in der Werbeforschung eingesetzten Warmth-Konstrukts, die jedoch keine schlüssigen Befunde präsentieren kann (vgl. Kapitel 3.3). Die Ergebnisse von Wolf (2010), die durch zwei Gruppen entweder die Sympathie oder die Kompetenz der Kandidaten bewerten lässt, deuten darauf hin, dass dies zumindest schwierig ist. Im Zeitverlauf des Stimulus nähern sich die Bewertungen durch beide Gruppen immer weiter an, was darauf hindeutet, dass die Probanden eine solche konkrete Anweisung nur eine begrenzte Zeit lang befolgen (können).

Schließlich ist festzuhalten, dass entsprechend den vielfältigen Forschungsinteressen sehr unterschiedliche Typen von Stimuli zum Einsatz kommen. Die Bandbreite reicht von TV-Duellen über politische, kommerzielle oder informierende Spotformate bis hin zu (Ausschnitten aus) Informations- und

Unterhaltungssendungen. Die Stimuli unterscheiden sich damit deutlich hinsichtlich ihrer Inhalte, aber auch hinsichtlich ihrer formalen Charakteristika wie der Länge oder der formalen Strukturierung. Entsprechend finden sich sowohl Studien, in denen die Probanden einzelne (meist längere) Stimuli rezipieren, als auch Studien, in denen nacheinander mehrere (meist kürzere) Stimuli gezeigt werden. Wie die Präsentation der Stimuli im Design der RTR-Studien gehandhabt wird, um die Effekte der Stimulusinhalte auf die rezeptionsbegleitend gemessenen Konstrukte zu untersuchen, stellen wir im nächsten Abschnitt dar.

Designs von RTR-Studien

Ein wesentliches Ziel vieler RTR-Studien besteht darin, Effekte von Merkmalen der Stimuli auf die rezeptionsbegleitend gemessenen Konstrukte zu untersuchen. Ganz grundsätzlich können hierfür zwei Designs genutzt werden: experimentelle Designs und Messwiederholungsdesigns.

Experimentelle Designs In experimentellen RTR-Studien sieht jeder Teilnehmer eine von verschiedenen Versionen desselben Stimulus, die hinsichtlich der Merkmale manipuliert werden, deren Effekte auf das rezeptionsbegleitend erfasste Konstrukt von Interesse sind.[15] Unterschiede bzw. Zusammenhänge zwischen den RTR-Verläufen der Experimentalgruppen können in diesen Studien als Effekte der manipulierten Stimulusmerkmale interpretiert werden. Um direkte Vergleiche der RTR-Messungen zu ermöglichen, muss die zeitliche Struktur des Stimulus intakt bleiben. Typische Beispiele sind hier Studien, die den Modus der Stimuluspräsentation verändern, also den Experimentalgruppen das Video mit Ton, nur den Ton oder nur das Video präsentieren (z.B. Faas & Maier, 2004b; Kercher, 2013; M. Maier & Maier, 2009; Maurer, 2009; Roessing et al., 2009). Ergibt sich beispielsweise an einer Stelle des Stimulus ein Unterschied zwischen der Gruppe, die den Stimulus nur hört, und der Gruppe, die den Stimulus audiovisuell rezipiert, so kann dies als Effekt der visuell vermittelten Eindrücke interpretiert werden. Ähnlich ist auch die Stimulusvariation in der experimentellen TV-Duell-Studie von Wolf (2010) aufgebaut.

15 Der Vollständigkeit halber ist an dieser Stelle auf drei weitere experimentelle Variationen hinzuweisen, die sich häufiger in RTR-Studien finden: Manipulationen vor der Rezeption des Stimulus (z.B. durch das Priming einer Themenwichtigkeit (Tedesco & Ivory, 2009) oder eines Bewertungsmaßstabs (Früh & Fahr, 2006)); Manipulationen während der Rezeption des Stimulus, die sich jedoch nicht auf den Inhalt des Stimulus auswirken (z.B. die Variation der Rezeptionssituation (Ramanathan & McGill, 2007)); Vorgabe unterschiedlicher Bedeutungen der RTR-Skala (z.B. Sympathie bzw. Kompetenz der Kandidaten (Wolf, 2010) oder subjektive Informiertheit bzw. Unterhaltungserleben (Fahr, 2009)). Da in diesen Studien der Stimulus selbst unverändert bleibt, gehen wir auf diese Designs hier nicht weiter ein.

Hier werden künstlich erzeugte RTR-Verläufe, die zugunsten eines der beiden Kandidaten ausfallen und denen unterschiedliche Quellen zugeschrieben werden (Expertenurteil bzw. Publikumsurteil), in den Zusammenschnitt eines TV-Duells integriert. Unterschiede zwischen den Experimentalgruppen weisen so auf die Effekte der eingeblendeten Fremdurteile hin.

Eine vergleichbare analytische Logik liegt Studien zugrunde, die eine beschränkte Manipulation desselben Stimulus vornehmen, wobei zwar die exakte zeitliche Struktur des Stimulus verändert wird, die strukturelle Abfolge der wesentlichen Inhalte aber vergleichbar bleibt. So ist es beispielsweise bei der Manipulation der vokalen und durch Gestik ausgedrückten Betonungen eines Vortrags kaum zu vermeiden, dass die verbal identischen Inhalte des Vortrags um einige Sekunden verschoben werden (Jackob et al., 2008, S. 225). Die RTR-Verläufe der unterschiedlichen Experimentalbedingungen lassen sich jedoch zumindest inhaltlich, wenn auch nicht auf Sekundenbasis statistisch, vergleichen, um auf Effekte von rhetorischen Betonungen zu schließen. Hier deutet sich an, dass das notwendige Beibehalten der zeitlichen Struktur den Möglichkeiten der experimentellen Manipulation der Stimuli recht enge Grenzen setzt. Wenn die externe Validität der Stimuli zumindest halbwegs gewahrt bleiben soll, erfordert die Erstellung manipulierter Stimuli, die über die Ausblendung eines Präsentationsmodus oder die Einblendung grafischer Elemente hinausgeht, sehr großen Aufwand. Daher finden sich nur recht wenige RTR-Studien mit experimentellen Variationen von Merkmalen des Stimulus, obwohl nur diese einen (einfachen) kausalen Nachweis von Effekten der manipulierten Stimulusmerkmale ermöglichen.

Messwiederholungsdesigns Die zweite Möglichkeit, die Effekte von Merkmalen des Stimulus auf die Rezipienten zu untersuchen, bietet das Messwiederholungsdesign jeder RTR-Studie. Die kontinuierlichen Stimuli enthalten im Zeitverlauf variierende Merkmale, mit den RTR-Messungen werden kontinuierlich die Reaktionen der Rezipienten erfasst. Als Effekt eines Merkmals des Stimulus können Veränderungen in den RTR-Messungen interpretiert werden, die zeitlich auf eine Veränderung des Stimulus folgen. In einer anderen Betrachtungsweise können die RTR-Messungen infolge des Auftretens eines bestimmten Merkmals mit den RTR-Messungen zu anderen Zeitpunkten verglichen werden, um auf den Effekt des Merkmals zu schließen. Diese analytische Logik liegt jeder RTR-Studie, die an der längsschnittlichen Untersuchung der rezeptionsbegleitend gemessenen Konstrukte interessiert ist, zugrunde. Um eine solche Analyse möglich zu machen, muss bei der Planung des Forschungsdesigns einer RTR-Studie sichergestellt werden, dass die Merkmale, deren Effekte

untersucht werden sollen, im Verlauf des Stimulus ausreichend variieren. Dazu stehen zwei Vorgehensweisen zur Verfügung.

Zum einen können denselben Probanden nacheinander mehrere (meist kürzere) Stimuli präsentiert werden, die sich hinsichtlich der Merkmale, die für das Forschungsinteresse relevant sind, unterscheiden. In einem einfachen Messwiederholungsdesign ohne RTR-Messung könnte dann nach jedem Stimulus eine Befragung folgen. In der abschließenden Auswertung werden dann die Antworten aus jeder Befragungswelle intraindividuell miteinander verglichen. Wenn in einer RTR-Studie nacheinander mehrere Stimuli präsentiert werden, liegen in zweifacher Hinsicht Messwiederholungen vor: Von denselben Probanden werden zu jedem einzelnen Stimulus mehrere RTR-Messungen abgegeben, und von denselben Probanden werden nacheinander zu mehreren Stimuli RTR-Messungen abgegeben. Oder wie wir es in Kapitel 6 in der Begrifflichkeit der Mehrebenenmodelle ausdrücken: Die RTR-Messungen sind in einer Probanden-Stimulus-Kombination geschachtelt, die wiederum zu einem Probanden (ein Proband sieht nacheinander alle Stimuli) und zu einem Stimulus (jeder Stimulus wird von allen Probanden gesehen) gehören.

Besonders häufig findet sich diese Form der Messwiederholung in Studien, in denen kommerzielle, informierende oder politische Spots gezeigt werden (z.B. Aaker et al., 1986; Abeele & MacLachlan, 1994; Algie & Rossiter, 2010; Baggaley et al., 1992; Hughes, 1992; Kaid, 2009; J. Maier & Maier, 2007; Tedesco & Ivory, 2009). Die meisten Stimuli unter den hier gesichteten Studien setzen Baumgartner et al. (1997) sowie Woltman Elpers et al. (2004) ein, die ihren Probanden nacheinander 30 Werbespots präsentieren. Häufig geht es in diesen Studien darum, herauszufinden, welche Spots nach einem bestimmten Kriterium im direkten Vergleich zueinander am besten abschneiden – z.B. im Zeitverlauf möglichst schnell ein möglichst positives Gefühl erzeugen. In den Studien, die eine sehr große Zahl an Spots umfassen, kann zudem untersucht werden, welche Eigenschaften der Stimuli systematisch mit den Reaktionen der Rezipienten zusammenhängen.

Doch auch andere, kommunikationswissenschaftlich orientierte RTR-Studien setzen derartige Messwiederholungsdesigns ein. Früh (2010) zeigt den Probanden nacheinander fünf Nachrichtenbeiträge, die sich unter anderem in den darin vorkommenden Nachrichtenfaktoren unterscheiden. Die Stimuli von Kercher (2013) variieren in ihrer formalen Verständlichkeit. Bei Fahr (2009) sehen die Probanden nacheinander Ausschnitte aus sechs Talkshows, die hinsichtlich ihres Informationsgehalts, dem Sozialverhalten der Talkshowgäste und der Dynamik variieren. Im Vordergrund der Analysen stehen dann nicht mehr (nur) die Veränderungen der RTR-Messungen von Sekunde zu Sekunde, sondern der Vergleich der Verläufe zwischen den Stimuli. Für die Einsatz

solcher Messwiederholungsdesigns sprechen in erster Linie forschungsökonomische Gründe. Wenn die Forscher bereits den großen Aufwand geleistet haben, die technische Ausstattung zur RTR-Messung zu installieren und die Probanden in der Bedienung zu instruieren, und die Probanden ihrerseits den Aufwand der Teilnahme an einem Laborexperiment eingehen, so können auch gleich Messungen zu mehreren Stimuli durchgeführt werden. Darüber hinaus kann die Rezeption mehrerer (kürzerer) Medienstimuli auch der realen Rezeptionssituation entsprechen, z.B. beim Sehen von mehreren Werbespots in einem Werbeblock oder von mehreren Nachrichtenbeiträgen in einer Nachrichtensendung. Hier kann das Messwiederholungsdesign dazu beitragen, die externe Validität der Untersuchung zu erhöhen.

Die zweite Möglichkeit besteht darin, mit einem (meist längeren) natürlichen Stimulus zu arbeiten, von dem erwartet werden kann, dass ein ausreichendes Maß an Variation in den relevanten Merkmalen über die Zeit vorkommt. Dies trifft, wie in Kapitel 2 ausführlich erläutert, auf die TV-Duell-Studien zu, die in dieser Arbeit im Mittelpunkt stehen. Aber auch Studien, die längere Ausschnitte aus Spielfilmen als Stimuli verwenden (z.B. Wünsch, 2006b), lassen sich hier einordnen. Auf den ersten Blick geben die Rezipienten in diesen Studien kontinuierlich über einen längeren Zeitraum RTR-Messungen zu einem Stimulus ab. Doch die meisten längeren massenmedialen Stimuli lassen sich auch als eine Abfolge von voneinander abgrenzbaren Episoden auffassen. In TV-Duellen ist beispielsweise der Wechsel des Sprechers eine deutlich wahrnehmbare Grenze zwischen den Episoden. Nach dieser Logik können die RTR-Messungen in einer TV-Duell-Studie auch als eine Reihe von Reaktionen auf aufeinander folgende und sich aufeinander beziehende, aber auch voneinander abgrenzbare Aussagen von Politikern (und Moderatoren) betrachtet werden. Die Inhalte in jeder dieser Episoden können dann im Hinblick auf das Vorkommen der Merkmale von Interesse geordnet und die Reaktionen während Episoden, in denen bestimmte Merkmale vorkommen bzw. nicht vorkommen, miteinander verglichen werden. Dies entspricht dem analytischen Vorgehen, das auch in Studien mit mehreren, im Hinblick auf die Variation der relevanten Merkmale ausgewählten Stimuli zum Einsatz kommt. Wie die Episoden für diese Form des analytischen Zugangs abgegrenzt werden, bleibt zunächst den Forschern überlassen. Sowohl eine formale Abgrenzung – bei TV-Duellen z.B. nach Sprecherwechseln, in Spielfilmen z.B. nach Szenen – als auch eine inhaltliche Abgrenzung – z.B. nach thematisch geschlossenen Aussagen eines Akteurs – ist möglich. Die kleinste technisch mögliche Einteilung ist das Messintervall der RTR-Messung (z.B. sekündlich bei Nagel, 2012, siehe ausführlich in Kapitel 5.3.1). Voraussetzung ist jedoch, dass die für die Analyse gebildeten Episoden auch von Rezipienten als eigenständige Abschnitte des

Stimulus wahrgenommen werden. Nur dann ist eine valide Verknüpfung der RTR-Messungen mit den den Episoden zugeordneten Merkmalen des Stimulus möglich.

Zusammengefasst bestehen drei grundsätzliche Möglichkeiten, in RTR-Studien Effekte der Stimulusmerkmale auf die rezeptionsbegleitend gemessenen Konstrukte zu untersuchen. In experimentellen Designs werden die Stimuli hinsichtlich der Merkmale von Interesse manipuliert und dann die RTR-Messungen von Experimentalgruppen bei der Rezeption dieser Stimuli verglichen. In Messwiederholungsdesigns mit mehreren, in sich geschlossenen Stimuli werden solche Stimuli ausgesucht, die sich im Hinblick auf die relevanten Merkmale deutlich voneinander unterscheiden. Dann werden die RTR-Messungen derselben Rezipienten zu allen Stimuli in Abhängigkeit von den relevanten Merkmalen verglichen. Schließlich können (längere) Stimuli analytisch in Episoden zerlegt werden, denen dann die für die Forschungsfrage interessanten Merkmale zugeordnet werden. Die RTR-Messungen werden dann ebenfalls mit diesen Episoden verknüpft und können in Abhängigkeit von deren Merkmalen untersucht werden. Die drei Möglichkeiten schließen sich nicht gegenseitig aus. Es werden Studien durchgeführt, in denen die Probanden einen längeren episodisch unterteilbaren Stimulus (z.B. Faas & Maier, 2004b; Maurer, 2009) oder mehrere kürzere (z.B. M. Maier & Maier, 2009) Stimuli sehen und zudem die Präsentationmodi experimentell variiert werden. Ebenso kann in den Studien, die mehrere, in sich geschlossene Stimuli in Messwiederholung präsentieren, zu Analysezwecken eine weitere Unterteilung der Stimuli zu Episoden vorgenommen werden.

3.3 Methodologische Forschung zu RTR-Messungen

Seit der Einführung der RTR-Messungen in den 1930er Jahren, besonders aber seit ihrer Popularisierung durch technische Weiterentwicklungen in den 1980er Jahren, findet auch eine methodologische Auseinandersetzung mit diesem Verfahren statt. Auf zwei Schwerpunkte dieser methodologischen Forschung wollen wir im Folgenden eingehen: die Konsequenzen der RTR-Messungen für die externe Validität der Studien sowie die Güte des Messinstruments.

Externe Validität von RTR-Studien

Die Güte jeder empirischen Studie – und damit auch einer RTR-Studie – lässt sich anhand ihrer internen und externen Validität bemessen. Die externe Validität einer Studie bezieht sich ganz allgemein auf die Möglichkeit, die Ergebnisse

„auf andere Personen, Situationen oder Zeitpunkte“ (Bortz & Döring, 2006, S. 53) zu übertragen. Die interne Validität bezieht sich auf die Fähigkeit einer Untersuchung, „Veränderungen in den abhängigen Variablen eindeutig auf den Einfluss der unabhängigen Variablen zurückzuführen“ (ibid.). Mit Maßnahmen, welche die interne Validität einer Studie erhöhen, gehen in der Regel Einbußen bei der externen Validität einher. Um die Validität von RTR-Studien einschätzen zu können, müssen wir uns zuerst bewusst machen, dass die meisten von uns gesichteten RTR-Studien als (quasi-) experimentelle Rezeptionsstudien in einem Laborsetting durchgeführt wurden. Damit wird die interne Validität der Studien erhöht, da die Medienrezeption in einem für alle Probanden standardisierten Umfeld erfolgt und relativ gut kontrolliert werden kann. In der Folge gelten für die externe Validität der RTR-Studien aber auch dieselben Einschränkungen wie für andere Laborstudien (Maurer, 2013b, S. 231).

Zum einen ist die Übertragbarkeit der Befunde auf natürliche Rezeptionssituationen eingeschränkt. Es kann einerseits vermutet werden, dass die Probanden dem Medienstimulus im Labor eine größere Aufmerksamkeit schenken als bei der alltäglichen Medienrezeption, da sie genau zu diesem Zweck an der Studie teilnehmen. Andererseits ist es ebenso möglich, dass die Probanden sich bei einer gemeinsamen Rezeption in einer (größeren) Gruppe gegenseitig ablenken und beeinflussen, was eine Übertragbarkeit auf die Situation einer typischen Mediennutzung alleine oder in Kleingruppen einschränkt. Nach den Befunden von Ramanathan und McGill (2007) erscheint eine solche gegenseitige Beeinflussung durchaus wahrscheinlich. Papastefanou (2013, S. 7-9) kritisiert in diesem Kontext vor allem die Studien zu TV-Debatten, die – wenn sie die Datenerhebung zeitgleich zur öffentlichen Ausstrahlung durchführen wollen – eine Rezeption in sehr großen Gruppen von teils über 100 Personen kaum vermeiden können.

Zum anderen stellen Laborstudien recht hohe Anforderungen an die Versuchspersonen, was die Übertragbarkeit der Befunde auf andere Personen außerhalb der Stichprobe einschränkt. Selbst wenn es gelänge, eine repräsentative Stichprobe aus der Grundgesamtheit von Interesse zu ziehen, wäre die Teilnahmebereitschaft durch den erforderlichen Aufwand seitens der Probanden, an der Studie im Labor teilzunehmen, wohl äußerst gering. Zudem ist zu erwarten, dass die Teilnahmebereitschaft systematischen Einflüssen unterliegt, die einen unverzerrten Schluss auf die Grundgesamtheit beeinträchtigen. Wieder dürfte dies für Live-Studien zu TV-Debatten besonders problematisch sein, da hier neben der Anwesenheit am Durchführungsort zusätzlich die Einhaltung eines einzelnen Termins, meist recht spät am Abend, notwendig ist.

Diese Einschränkungen der externen Validität gelten für alle Laborstudien, lassen sich im Kontext von RTR-Studien jedoch besonders schwer beheben.

Zwar sind die meisten RTR-Systeme transportabel und ließen sich prinzipiell auch zuhause bei den Probanden installieren. Hiermit wäre aber ein sehr großer Aufwand verbunden. Speziell bei TV-Duell-Studien, an denen sehr viele Probanden gleichzeitig teilnehmen sollen, wäre dies kaum zu leisten. Mögliche Lösungen versprechen zum einen Online-RTR-Systeme, die von den Teilnehmern in einem Web-Browser ohne Anwesenheit eines Versuchsleiters ausgeführt werden können (z.B. Iyengar, 2011; Kercher et al., 2012). Auch Smartphone-Apps sind in diesem Kontext denkbar.[16] Zum anderen sind mittlerweile portable RTR-Geräte verfügbar, die autark von einem zentralen System arbeiten und gegebenenfalls auch wie eine postalische Befragung an die Teilnehmer versendet und von diesen nach der Studie wieder zurückgesendet werden können (Papastefanou, 2013). Mit solchen Systemen könnten zumindest Probanden erreicht werden, die vor allem wegen des notwendigen Erscheinens im Labor nicht an den RTR-Studien teilnehmen. Werden diese Geräte vor dem heimischen Fernseher genutzt, so könnte auch eine bessere Annäherung an eine natürliche Rezeptionssituation erreicht werden. Allerdings ist bei all diesen Varianten zu bedenken, dass die Kontrolle über die Rezeptionssituation aufgegeben wird und sich dadurch die interne Validität der Studie verschlechtert. Dies ist vor allem dann ein wesentlicher Nachteil, wenn die RTR-Messungen den Inhalten des Stimulus sehr genau – z.B. auf Sekundenbasis – zugeordnet werden sollen. Eine empirische Analyse zum Verhältnis von interner und externer Validität solcher „unkontrollierter“ RTR-Studien steht noch aus, da die methodologischen Arbeiten zu diesen RTR-Systemen (Kercher et al., 2012; Papastefanou, 2013) keinen Vergleich mit RTR-Studien im Labor vornehmen.

Neben diesen allgemeinen Nachteilen von Laborstudien für die externe Validität sind auch weitere Einschränkungen zu bedenken, die sich durch die Durchführung der RTR-Messung ergeben können. Durch den Einsatz einer rezeptionsbegleitenden Messung wird die interne Validität einer Studie weiter verbessert, da sich nun einzelne Reaktionen der Rezipienten zeitlich im Verlauf des Stimulus verorten lassen. Eine RTR-Messung ist aber, wie auch die Befragung, ein reaktives Verfahren der Datenerhebung, das die Versuchsteilnehmer beeinflussen kann (Maurer, 2013b, S. 227). Insbesondere ist zu befürchten, dass RTR-Messungen die Informationsverarbeitung und das Rezeptionserleben im Vergleich zu einer Rezeption ohne RTR-Messung verändern. Um einzuschätzen, ob die Probanden von den RTR-Messungen beeinflusst werden, können die Probanden selbst nach ihrer subjektiven Erfahrung mit dem Verfahren befragt

16 Zu den Debatten im US-Präsidentschaftswahlkampf 2012 wurde bereits eine Studie mit RTR-Messungen über eine Smartphone-App durchgeführt, an der fast 5000 Probanden in allen Staaten der USA teilnahmen. Die Befunde sind bisher noch nicht publiziert: `http://www.politicalcommunication.org/newsletter_23_1_react.html`.

werden. Aus Sicht der Probanden scheint die Abgabe der RTR-Messungen während der Rezeption recht unproblematisch zu sein. Sie berichten, dass das Bedienen der Geräte leicht fällt und nur wenig vom Stimulus ablenkt (z.B. Fahr, 2006, 2009; Fahr & Fahr, 2009; Kercher et al., 2012; J. Maier, 2013).

Einen besseren Nachweis etwaiger Reaktivität des Verfahrens können Methodenexperimente leisten, in denen Versuchsgruppen mit und ohne RTR-Messung verglichen werden. Eine Möglichkeit besteht darin, ebenfalls rezeptionsbegleitend erfasste physiologische Messungen in Abhängigkeit vom Einsatz von RTR-Messungen zu untersuchen. Fahr (2006, S. 210) berichtet von einem nicht im Detail publizierten Experiment, in dem die physiologisch gemessene Aktivierung von Probanden mit und ohne RTR-Messung verglichen werden. Die RTR-Messung führt demnach zu einer „Nivellierung der Aktivierung" (Fahr, 2006, S. 210): Die physiologischen Messungen der Probanden mit RTR-Messung zeigen im Vergleich zur Kontrollgruppe eine geringere Amplitude. An spannenden Stellen des Stimulus ist ihre Aktivierung geringer, an langweiligen Stellen dagegen höher als in der Kontrollgruppe. In einem weiteren Experiment mit dem gleichen Versuchsaufbau finden Fahr und Fahr (2009) eine komplexe Interaktion zwischen der Abgabe von RTR-Messungen, dem Zeitverlauf des Stimulus und einzelnen Merkmalen des Stimulus auf das physiologisch gemessene Arousal. Besonders groß ist der Unterschied zwischen den Gruppen mit und ohne RTR-Messung zu Beginn des Stimulus. Es ist plausibel, dass dieser Unterschied durch die zunächst noch ungewohnte Selbstauskunft mittels des RTR-Geräts verursacht wird. Daher sollte vor der RTR-Messung zum eigentlichen Stimulus immer eine Übungsphase stattfinden, in der sich die Probanden an diese Aufgabe gewöhnen können (Maurer, 2013b, S. 228). Aaker et al. (1986, S. 379-380) berichten im Gegensatz zu diesen Studien keinen relevanten Einfluss der Wiedergabe des emotionalen Erlebens mit dem Warmth Monitor auf die physiologisch erfasste Aktivierung.

Hutcherson et al. (2005) untersuchen schließlich den Einfluss der kontinuierlichen Selbstauskunft über das emotionale Erleben während der Rezeption von traurigen und heiteren Filmstimuli in einer Neuro-Imaging-Studie. Sie zeigen einerseits, dass Probanden, die ihre Emotionen mit RTR berichten, im Vergleich zu einer Kontrollgruppe ohne RTR keine geringere Aktivität in Hirnregionen aufweisen, die mit dem emotionalen Erleben assoziiert sind. Das emotionale Erleben selbst wird durch die Messung also nicht beeinträchtig. Andererseits finden sie in der RTR-Gruppe eine höhere Aktivierung der Regionen, denen die Verantwortung für die emotionale Introspektion zugeschrieben wird. Diese Probanden setzen sich also in stärkerem Maße bewusst mit ihren Emotionen während der Rezeption auseinander. Dies spricht für eine Abweichung von

einer nicht reflektierten Medienrezeption und könnte damit für einige Ansätze der Unterhaltungsforschung als problematisch gelten.

Weitere Studien untersuchen den Einfluss der RTR-Messung auf nach der Rezeption gemessene Variablen, um so auf die Reaktivität des Verfahrens zu schließen. Reinemann und Maurer (2009) sowie J. Maier (2011) widmen sich der Frage, ob die Durchführung einer RTR-Messung die Erinnerungsleistung der Probanden beeinflusst. Die erste Studie kommt zu dem Ergebnis, dass verbal vermittelte Informationen von Probanden mit und ohne RTR-Messung gleichermaßen erinnert werden. Visuell vermittelte Informationen werden von den Probanden, die RTR-Wertungen abgegeben haben, etwas schlechter erinnert. Die Autoren interpretieren dies als eine Folge von Blicken auf die RTR-Dials, die von den visuellen Inhalten des Stimulus ablenken. Die zweite Studie untersucht den Effekt einer RTR-Messung in Abhängigkeit vom Konfliktgehalt einer TV-Debatte. Während sich bei einem konfliktarmen Stimulus keine Unterschiede in der Erinnerungsleistung der Probanden mit und ohne RTR-Geräte finden, fällt die Erinnerungsleistung der Probanden mit RTR an eine konflikthaltige Debatte schwächer aus. Dies kann nach Sicht des Autors die externe Validität von RTR-Studien zu TV-Debatten gefährden, dürfte aber zumindest in Deutschland wegen der meist stark strukturierten und wenig hitzigen Debatten weniger stark ins Gewicht fallen.

Zu ähnlichen Befunden kommen auch Experimente, die den Effekt der RTR-Messung auf postrezeptiv gemessene subjektive Einschätzungen untersuchen. Aaker et al. (1986) finden bei der Bewertung von Werbespots nach der Rezeption bei 23 von 124 (nicht näher bezeichneten) Items signifikante Unterschiede zwischen der Experimentalgruppe mit RTR und der Kontrollgruppe. Rossiter und Thornton (2004) zeigen, dass die postrezeptiven Berichte der Anspannung während des Sehens von sieben Aufklärungsspots zu den Gefahren des zu schnellen Fahrens zwischen einer Gruppe, die ihre Anspannung während der Rezeption mit RTR berichtete, und einer Kontrollgruppe ohne RTR-Messung weitgehend übereinstimmen.

Insgesamt weisen die Befunde der vorliegenden Studien auf einen negativen Einfluss der rezeptionsbegleitenden Datenerhebung auf die externe Validität der RTR-Studien hin, der aber insgesamt relativ gering ausfällt. In Anbetracht des zusätzlichen Erkenntnisgewinns kommen die Autoren durchweg zu dem Schluss, dass dies durch die mit RTR-Messungen erreichte Verbesserung der internen Validität zu rechtfertigen ist. Durch eine gewissenhafte Instruktion der Probanden und ausführliche Übungsphasen sollten die Konsequenzen der Reaktivität beherrschbar sein (Reinemann & Maurer, 2009, S. 42). Es muss allerdings daran erinnert werden, dass diese Evaluationsstudien ebenfalls als Laborexperimente durchgeführt werden. Vergleichsmaßstab ist dadurch

nicht die natürliche Rezeptionssituation, sondern die Laborstudie ohne RTR-Messungen. Es kann also gefolgert werden, dass RTR-Messungen die externe Validität von im Labor durchgeführten Rezeptionsstudien nur unwesentlich weiter einschränken.

Reliabilität und Validität von RTR-Messungen

Abschließend fassen wir zusammen, welche Befunde zur Qualität des Messinstruments RTR vorliegen. Dazu ziehen wir die Kriterien Reliabilität und Validität heran.

Reliabilität Die Reliabilität oder Zuverlässigkeit eines Tests gibt an, wie präzise ein Instrument den wahren Wert eines Konstrukts bestimmt. Ist ein Test reliabel, so besteht der gemessene Wert zu einem sehr großen Anteil aus dem wahren Wert und zu einem sehr kleinen Anteil aus Messfehlern (Bortz & Döring, 2006, S. 196). Diese Definition gilt grundsätzlich auch für das Testinstrument der RTR-Messung. Allerdings ist die Anwendung der typischen Indikatoren, mit denen die Reliabilität eines Instruments bestimmt werden kann, auf die RTR-Messung aus verschiedenen Gründen problematisch. Das konzeptionell einfachste Vorgehen ist die Bestimmung der Retestreliabilität: Dasselbe Instrument wird durch dieselben Personen mehrmals auf dasselbe Bewertungsobjekt angewendet. Stimmen die Messungen in hohem Maße überein, so misst das Instrument zuverlässig (Bortz & Döring, 2006, S. 196-197). Dieses Vorgehen ist prinzipiell auch für RTR-Messungen durchführbar. Aaker et al. (1986) sowie Stayman und Aaker (1993) führen in ihren Evaluationsstudien zum Warmth Monitor klassische Tests der Retestreliabilität durch, indem sie das Instrument von denselben Probanden zweimal auf dieselben Werbespots anwenden lassen. In der ersten Studie finden die Autoren eine befriedigende durchschnittliche Korrelation von .81 zwischen den wiederholten Messungen. In der zweiten Studie schwanken die Korrelationen in Abhängigkeit von Charakteristika der Teststimuli und dem mit RTR wiederzugebenden Konstrukt deutlich zwischen .08 und .85. Wenn jedoch ein zu den jeweiligen Spots passendes Konstrukt gemessen wird (z.B. der Humor während eines lustigen Spots oder die emotionale Wärme während eines emotionalen Spots), liegt die minimale Reliabilität bei .65. Zu ähnlichen Befunden kommt eine Studie von Hughes (1992), der denselben Probanden zweimal dieselben Werbespots vorlegt, um die Abnutzung („wearout“, S. 61) der Spots zu untersuchen: Bei einigen Spots gleichen sich die über die Probanden aggregierten RTR-Verläufe in hohem Maße, bei anderen unterscheiden sie sich recht deutlich. Die Kontextabhängigkeit der Retestreliabilität zeigt sich auch bei Lambooij et al. (2011). Die

Probanden bewerten zuerst sechs kurze Sequenzen, die dann noch einmal im Kontext eines längeren Stimulus gezeigt werden. Zwischen den wiederholten Messungen ergeben sich zum größten Teil nur schwache und recht unsystematische Korrelationen. Diese insgesamt zunächst ernüchternden Befunde zur Retestreliabilität sind jedoch auch durch die Natur der RTR-Messungen gegeben. Maurer und Reinemann (2009) stellen daher den Nutzen des Retestverfahrens infrage: „As the goal of RTR is to measure spontaneous reactions, in some cases participants will react differently to a second presentation of the same stimulus“ (S. 9).

Weitere Reliabilitätsindikatoren – Paralleltestreliabilität, Testhalbierungsreliabilität und interne Konsistenz – können nur für Multi-Item-Skalen bestimmt werden (Bortz & Döring, 2006, S. 197-199). Da in RTR-Messungen jedoch nur ein (mit Dial- bzw. Slider-Eingabegeräten) oder sehr wenige (mit Push-Button-Geräten) Items von den Probanden bewertet werden, können diese Indikatoren nicht eins zu eins auf die Evaluation der Reliabilität von RTR-Messungen übertragen werden. Daher wurden Hybridmaße entwickelt, die einige Eigenschaften der Retestreliabilität und der Reliabilitätsindikatoren, die eigentlich Multi-Item-Skalen erfordern, vereinen. Bei der Interpretation dieser Befunde ist allerdings Vorsicht geboten, da die Tests häufig mit den in der klassischen Testtheorie üblichen Bezeichnungen versehen werden, jedoch im Detail eine andere Bedeutung haben.

Fenwick und Rice (1991) führen eine RTR-Evaluationsstudie mit vier Werbespots durch. Dabei werden dieselben Spots von „matched samples“ (S. 26), also mehreren unabhängigen Personenstichproben, bewertet, die sich in ihren wesentlichen Charakteristika gleichen. Da sich die RTR-Verläufe nicht zwischen den Stichproben unterscheiden, wird dem Verfahren eine gute Test-Retest-Reliabilität zugeschrieben. Die RTR-Messungen werden auch nicht davon beeinflusst, in welcher Reihenfolge die Probanden die Spots sehen. Da die Retestreliabilität in dieser Studie nicht durch die Anwendung des Instruments durch dieselben Personen, sondern durch mehrere strukturell identische Stichproben festgestellt wird, bezieht sich das Ergebnis jedoch nicht auf die intraindividuelle Stabilität der Messung. Vielmehr können wir folgern, dass das Verfahren stabile Messungen für unterschiedliche Stichproben aus derselben Grundgesamtheit ermöglicht. Eine ältere Studie von Schwerin (1940) kommt mit diesem Vorgehen ebenfalls auf sehr gute Reliabilitätswerte (zit. nach Biocca et al., 1994, S. 30).

Die ausführlichste methodologische Arbeit zu RTR-Messungen in TV-Duell-Studien legen Reinemann, Maier, Faas und Maurer (2005) (auch veröffentlicht in J. Maier, Maurer, Reinemann & Faas, 2007) vor. In dieser Publikation werden die Messungen von zwei unabhängigen Studien zum zweiten TV-Duell im Bun-

destagswahlkampf 2002 vergleichend evaluiert (Faas & Maier, 2004a; Maurer & Reinemann, 2003, vgl. zu den Studien im Detail Kapitel 3.4). Im Rahmen dieser Arbeit präsentieren die Autoren visuelle Vergleiche und Korrelationen der über das jeweilige Testpublikum aggregierten RTR-Zeitreihen. Da Faas und Maier ein Push-Button-System eingesetzt haben, Maurer und Reinemann aber ein Dial-System, ähnelt dieser Vergleich einem Verfahren zur Feststellung der Paralleltestreliabilität: Es wird untersucht, ob mit unterschiedlichen Instrumenten eine äquivalente Messung gelingt. Die Autoren finden beim visuellen Vergleich der beiden Zeitreihen nur unwesentliche Unterschiede. Die Zeitreihen weisen über die gesamte Debatte hinweg eine mittlere Korrelation von $r = .38$ auf. In Phasen, die von den Autoren als die Schlüsselstellen der Debatte identifiziert werden, erreichen die Korrelationen noch (etwas) höhere Beträge zwischen $r = .46$ und $r = .69$. Um diese Befunde einzuordnen, müssen wir berücksichtigen, dass die Messungen im Gegensatz zu einem klassischen Paralleltest aus zwei unabhängigen Stichproben stammen und die Zusammenhänge auf Basis der über die Stichproben aggregierten Zeitreihen bestimmt werden. Damit sagen die Befunde von Reinemann et al. (2005) aus, dass die unmittelbaren Kandidatenbewertungen durch das gesamte Publikum unabhängig von der Stichprobe und den eingesetzten RTR-Eingabegeräten (Push-Button vs. Dial) mit einer mäßigen bis zufriedenstellenden Reliabilität erfasst werden.

In einem ähnlich angelegten Methodenexperiment vergleichen wir die aggregierten RTR-Zeitreihen zur Bewertung von vier kurzen Videos mit Online-Umsetzungen von Push-Button-, Dial- und Slider-Eingabegeräten (Kercher et al., 2012, S. 18-19). Die aggregierten Messungen von Dreh- und Schiebereglern weisen bei allen Spots hohe Korrelationen von $r > .76$ auf. Die Korrelationen zwischen den Zeitreihen der Push-Button-Messungen und den Zeitreihen der beiden anderen Eingabegeräte zeigen für zwei Videos mittelstarke Zusammenhänge zwischen $r = .42$ und $r = .65$. Für die beiden anderen Videos fallen diese Korrelationen jedoch sehr schwach aus ($r \leq .11$). In der Diskussion der Funktionsweise der unterschiedlichen RTR-Eingabegeräte haben wir bereits auf die Gemeinsamkeiten von Dial- und Slider-Geräten und die Unterschiede zu Push-Button-Geräten hingewiesen (vgl. Kapitel 3.1). Die variablen Befunde zur (mit unabhängigen Stichproben und aggregierten Messungen ermittelten) Paralleltestreliabilität zeigen empirisch, dass die Äquivalenz der Messungen mit unterschiedlichen Geräten von Merkmalen der Stimuli abhängt. Da nach dem aktuellen Forschungsstand kein Eingabegerät als Goldstandard für RTR-Messungen im Allgemeinen gelten kann, bleibt vorläufig nur festzustellen, dass Dreh- und Schieberegler einerseits und Push-Button-Geräte andererseits die aggregierten unmittelbaren Publikumsreaktionen nur in manchen Kontexten mit einiger Gemeinsamkeit messen. Eine Antwort auf die Frage, welches

Instrument welchen Aspekt der unmittelbaren Reaktionen mit welcher Zuverlässigkeit misst, steht noch aus. Keine Aussagen können die Studien zudem über die Reliabilität der individuellen RTR-Messungen – egal mit welchem Gerät – machen.

Schließlich liegen einige Arbeiten vor, in denen die Reliabilität der individuellen RTR-Messungen in Anlehnung an die Testhalbierungsreliabilität bzw. die daraus abgeleiteten Maße der internen Konsistenz (Bortz & Döring, 2006, S. 198-199) bestimmt wird. Statt wie in den klassischen Tests die einzelnen Items einer Multi-Item-Skala in zwei Hälften zu teilen, werden hier die wiederholten Messungen des RTR-Items in zwei Hälften geteilt. Die Reliabilität wird dann über die Kovarianz der beiden Hälften bestimmt. Die Testhalbierungsreliabilität einer RTR-Messung hat daher auch Charakteristika der Retestreliabilität, da die wiederholten Messungen desselben Konstrukts verglichen werden – allerdings bei der Anwendung auf unterschiedliche Inhalte des Stimulus. Vereinfacht formuliert ist eine RTR-Messung nach diesem Prinzip reliabel, wenn die Messungen während eines Teils des Stimulus hoch mit den Messungen während eines anderen Teil des Stimulus korrelieren. Nach diesem Kriterium kann die Reliabilität von RTR-Messungen als zufriedenstellend bis sehr gut gelten: Abeele und MacLachlan (1994) ermitteln für den Warmth Monitor mit zu 110 dreisekündigen Segmente von zwölf Werbespots zusammengefassten RTR-Messungen Koeffizienten der Testhalbierungsreliabilität von mindestens .85. Biocca et al. (1994, S. 30) berichten aus älteren Evaluationsstudien ebenfalls Korrelationen mit Beträgen größer .8. Lambooij et al. (2011) sowie Papastefanou (2013) berechnen für Ausschnitte bzw. Segmente ihrer RTR-Messungen Cronbachs α als Koeffizient der internen Konsistenz. Sie erzielen dabei gute Ergebnisse von $\alpha > .77$ bzw. $\alpha > .81$ (mit einer Ausnahme). Diese Ergebnisse deuten darauf hin, dass RTR-Messungen die rezeptionsbegleitend gemessenen Konstrukte konsistent erfassen können. Allerdings ist dieser Befund nicht uneingeschränkt positiv zu interpretieren. Wir müssen bedenken, dass die Messungen während eines sich verändernden Stimulus erfasst werden. Die hohe Konsistenz kann daher auch als ein Zeichen dafür interpretiert werden, dass die individuellen RTR-Messungen nur wenig sensibel auf Veränderungen des Stimulus reagieren. Was dies im Kontext einer Studie bedeutet, deren Ziel es ist, gerade solche intraindividuellen Messunterschiede im Zeitverlauf zu erklären, diskutieren wir ausführlich bei der Evaluation unserer eigenen Messung in Kapitel 4.2.

Validität Mit der Validität eines Testinstruments wird dessen Fähigkeit bezeichnet, das zu messen, was es messen soll. In einer TV-Duell-Studie muss beispielsweise sichergestellt werden, dass die RTR-Messung der unmittelbaren

Kandidatenbewertung tatsächlich den Eindruck misst, den ein Proband gerade von einem Kandidaten hat. Die Validität als Gütekriterium der Testkonstruktion ist nicht zu verwechseln mit der externen und internen Validität eines Studiendesigns (Bortz & Döring, 2006, S. 200). Die Validität der Messung ist eine notwendige, aber noch keine hinreichende Bedingung für eine valide Studie. Wenn ein valides Instrument auf eine unpassende Stichprobe angewendet wird, können keine extern validen Schlüsse auf die Grundgesamtheit gezogen werden. Genauso kann ein valides Instrument keine validen Schlüsse auf Ursache-Wirkungs-Beziehungen sicherstellen, wenn beispielsweise in einem fehlerhaft aufgebauten Experimentaldesign die interne Validität der Studie beeinträchtigt ist.

Vergleichsweise ausführlich ist die Validität des in der Werbeforschung eingesetzten Warmth Monitor dokumentiert (Aaker et al., 1986; Abeele & MacLachlan, 1994; Stayman & Aaker, 1993). Die Studien weisen übereinstimmend nach, dass mit diesem Instrument ein positiv besetztes Konstrukt gemessen wird. Die RTR-Messungen korrelieren stark mit nach der Rezeption gemessenen Berichten zur empfundenen Wärme und zum allgemeinen Gefallen der Spots. Sie sind zudem mit der berichteten Kaufwahrscheinlichkeit des Produkts assoziiert. Auch die rezeptionsbegleitenden Messungen von Humor und Informationsgehalt korrelieren stark mit den postrezeptiven Messungen derselben Konstrukte. Schließlich sprechen alle drei Studien den RTR-Messungen übereinstimmend eine gute Augenscheinvalidität zu. Veränderungen im rezeptionsbegleitend gemessenen Konstrukt finden sich vor allem in Werbespots, die dieses Konstrukt ansprechen. So reagiert beispielsweise der Warmth Monitor stärker auf emotionale Spots, während sich in informationshaltigen Spots nur wenig Variation zeigt.

Bezüglich der diskriminanten Validität des Warmth-Konstrukts und seiner Verbindung zur physiologisch erfassten Aktivierung sind die Ergebnisse allerdings widersprüchlich: Aaker et al. (1986) sowie Stayman und Aaker (1993) interpretieren ihre Befunde dahingehend, dass das Konstrukt der empfundenen Wärme zwar mit dem allgemeinen Gefallen des Spots verwandt ist, sich jedoch auch von ihm abgrenzen lässt. Zudem zeigen die Autoren eine mittlere bis starke Assoziation der empfundenen Wärme mit der physiologischen Aktivierung. Abeele und MacLachlan (1994) bezweifeln diese Ergebnisse und kommen zu dem Schluss, dass der Warmth Monitor ein allgemeines positives Gefühl misst, das sich nicht klar von anderen positiv besetzen Items (z.B. „Joy, Surprise, Anticipation, Acceptance“, S. 599) abgrenzen lässt. Auch die Verbindung von Warmth mit dem physiologischen Arousal können diese Autoren nicht replizieren. Alles in allem lässt sich festhalten, dass die RTR-Messungen bei

einer allgemeineren Interpretation der Werte valide sind. Offen bleibt hingegen, ob spezifischere Konstrukte trennscharf erfasst werden können.

Auch in der Unterhaltungsforschung wird versucht, das mit RTR gemessene Rezeptionserleben durch Korrelationen mit postrezeptiv erhobenen Konstrukten zu validieren. Wünsch (2006a) findet größtenteils mittlere Korrelationen, teilweise aber auch widersprüchliche Befunde unter Verwendung der beiden Maße. Insgesamt spricht er den Messungen eine „akzeptable externe Validität“ (S. 198) zu. Ähnlich findet Fahr (2006) in seiner Evaluationsstudie „eine gewisse Korrespondenz zwischen Verlaufsmessungen und Fragebogen [...], allerdings auf deutlich reguliertem Niveau“ (S. 218). Während der Rezeption können einige recht deutliche Ausschläge gemessen werden, während in der Nachbefragung kaum bemerkenswerte Emotionen berichtet werden. Der Autor schließt dann auch mit einer grundsätzlichen Kritik an der Validierung des rezeptionsbegleitend gemessenen emotionalen Erlebens durch postrezeptive Befragungen. Erstens ist unklar, welche der beiden Messungen der Wahrheit entspricht, oder ob beide Verfahren unterschiedliche Dimensionen des Konstrukts messen. Zweitens ist davon auszugehen, dass das emotionale Erleben bei der postrezeptiven Introspektion „reflektiert und transformiert“ (S. 218) wird, was zu anderen Befunden in der Nachbefragung führt. Und drittens ist die Zusammenfassung der rezeptionsbegleitenden Messungen zu einem einfachen Durchschnittswert über den gesamten Verlauf des Stimulus eine unzulässige Vereinfachung des dynamischen Aspekts des mit RTR gemessenen Konstrukts.

In Studien zur unmittelbaren Bewertung von Kandidaten in TV-Debatten lässt sich die Validität der RTR-Messungen zum einen bestimmen, indem geprüft wird, ob sich die RTR-Messungen im Zeitverlauf und über die gesamte Debatte hinweg zusammengefasst nach den politischen Voreinstellungen der Probanden unterscheiden (Übereinstimmungsvalidität). Dieses Kriterium wird in beiden von Reinemann et al. (2005) vorgestellten Studien, allen drei Erhebungen von Papastefanou (2013) und in der TV-Debatten-Studie in Kercher et al. (2012) klar erfüllt. Auch die Befunde der nicht methodologisch ausgerichteten Studien zur unmittelbaren Kandidatenbewertung lassen sich in dieser Richtung interpretieren (vgl. Kapitel 3.4). Zum anderen gilt es als ein Indikator für die Prognosevalidität der RTR-Messungen, wenn sie zur Erklärung der nach der Debatte gemessenen Einschätzung des Debattensiegers oder der Debattenleistung der Kandidaten beitragen. Auch hier sprechen die Ergebnisse der genannten Studien für eine zufriedenstellende Validität der RTR-Messungen.

Abschließend lässt sich festhalten, dass die methodologische Literatur im Großen und Ganzen für eine Validität von RTR-Messungen spricht, die Spannweite der Befunde einzelner Studien jedoch von einer ausreichenden und bis

hin zu einer sehr guten Validität reichen. Dies bestätigt auch eine Sekundäranalyse von Bacherle, Schneider und Krause (2012), in der die Prognosevalidität in 15 RTR-Studien aus den Bereichen Unterhaltungsforschung, Organisationskommunikation und Politischer Kommunikation geprüft wird. Die Korrelationen zwischen während und nach der Rezeption gemessenen Konstrukten reichen von $r = .19$ bis $r = .72$, mit einem gewichteten Mittelwert von $\rho = .47$. Bei aller berechtigter Kritik an den Validitätsindikatoren dürfen wir zuversichtlich sein, dass RTR-Messungen hinreichend valide Messungen liefern. Mit Bortz und Döring (2006, S. 202) können wir schließen, dass der Einsatz eines Messverfahrens zu rechtfertigen ist,

> wenn die Entscheidungen und Vorhersagen, die auf der Basis des Tests getroffen werden, tauglicher sind als Entscheidungen und Vorhersagen, die ohne den Test möglich wären – es sei denn, der mit dem Test verbundene Aufwand steht in keinem Verhältnis zum Informationsgewinn.

Auch wenn zur Qualität und Leistungsfähigkeit von RTR-Messungen als Verfahren der Datenerhebung sicherlich noch einiger Forschungsbedarf besteht, lässt sich die Durchführung von RTR-Studien in Anbetracht ihrer besonderen Potenziale und der Hinweise auf ihre zumindest hinreichende Validität und Reliabilität rechtfertigen. Dies gilt jedoch nur, wenn die aufwändig erhobenen Daten tatsächlich längsschnittlich ausgewertet werden. Für andere Zwecke stehen in klassischen Befragungen bewährtere Instrumente zur Verfügung.

3.4 Empirische Befunde zur Erklärung der unmittelbaren Kandidatenbewertungen in TV-Debatten

Das folgende Teilkapitel stellt den empirischen Forschungsstand zur Erklärung der unmittelbaren Bewertung der Kandidaten während TV-Debatten vor. Dabei konzentrieren wir uns auf Arbeiten, in denen die unmittelbaren Kandidatenbewertungen rezeptionsbegleitend gemessen und längsschnittlich ausgewertet werden. In ihrem Forschungsüberblick für das *Handbook of Political Communication Research* fordern McKinney und Carlin (2004, S. 208):

> We hope that future investigation will attempt to link both content and rhetorical analyses of candidates' development of argument with viewer reactions to debate dialogue.

Auch Reinemann und Maurer (2008, S. 5062) sehen die Frage, „what types of arguments, statements, rhetorical means, and visual message elements resonate

with the audience", als eine der wesentlichen Herausforderungen für zukünftige TV-Duell-Studien. Erkenntnisse darüber, wie bestimmte Debatteninhalte unmittelbar nach ihrer Rezeption bewertet werden, vermitteln einen tieferen Einblick in die Mechanismen der Debattenwahrnehmung und -wirkung und können so auch dazu beitragen, die Verarbeitung politischer Medieninhalte im Allgemeinen besser zu verstehen. Folgen wir der Analogie der TV-Debatten als Miniaturwahlkämpfe (Faas & Maier, 2004a, S. 56, ausführlich Kapitel 2), so sind – sofern es die Zusammensetzung der Stichprobe zulässt – in Grenzen auch Rückschlüsse auf die Verarbeitung und Bewertung der zentralen Kampagnenbotschaften im gesamten Wahlkampf außerhalb der TV-Debatte möglich. Diese Erkenntnisse können jedoch nur gewonnen werden, wenn die Bewertungen einzelner Debatteninhalte überhaupt erfasst werden. Es müssen also TV-Duell-Studien durchgeführt werden, in denen die Kandidatenbewertungen *während* der Debatte erfasst werden. Darüber hinaus müssen verlässliche Verfahren zur Verfügung stehen, um die Verknüpfung zwischen Inhalten und unmittelbarer Bewertung herzustellen und dabei auch den für die Verarbeitung politischer Informationen wichtigen Einfluss der Voreinstellungen (Iyengar & Simon, 2000; Zaller, 1992) zu berücksichtigen.

Die Analysen, in denen die unmittelbaren Kandidatenbewertungen durch die Debatteninhalte erklärt werden, lassen sich hinsichtlich ihres Erkenntnisinteresses in induktive und deduktive Analysen unterteilen. *Induktive* Arbeiten beschreiben meist die Bewertung der Kandidaten im Zeitverlauf und/oder identifizieren ausgehend von den unmittelbaren Bewertungen die Debatteninhalte, die diese Reaktionen ausgelöst haben. *Deduktiv* ausgerichtete Arbeiten interessieren sich dagegen für die Effekte bestimmter, unabhängig von den RTR-Messungen identifizierter Merkmale des Debatteninhalts auf die unmittelbaren Kandidatenbewertungen. In den nächsten beiden Teilkapiteln stellen wir diese beiden Typen anhand ausgewählter Beispiele aus der Forschungsliteratur vor.[17] Die Darstellung der Befunde folgt dabei den Interpretationen der jeweiligen Autoren. Eine kritische Betrachtung, inwieweit einzelne Interpretationen mit den eingesetzten Auswertungsverfahren vereinbar sind, findet sich in Kapitel 5.

17 Wir beschränken uns in diesem Kapitel auf Arbeiten, die reale TV-Debatten in Rezeptionsstudien während ihrer Ausstrahlung untersuchen. Studien, die (manipulierte) Ausschnitte aus realen TV-Debatten in experimentellen Designs als Stimuli verwenden (z.B. Maurer, 2009; Wolf, 2010), werden hier nicht behandelt.

3.4.1 Induktive Analysen

Typische übergeordnete Forschungsfragen der induktiven Analysen lauten:

- Wie wurden die Kandidaten im Debattenverlauf bewertet?
- Welche Aussagen wurden besonders gut, welche besonders schlecht bewertet?
- An welchen Stellen der Debatte finden sich Unterschiede zwischen den Bewertungen durch Teilgruppen des Publikums?

Erstes Anliegen dieser Arbeiten ist es also, die Bewertung der Kandidaten durch das gesamte Publikum und/oder Teilgruppen des Publikums im Verlauf der Debatte zu beschreiben. Auf dieser Basis soll dann gezeigt werden, welche Aussagen der Kandidaten besonders positiv oder negativ bewertet werden, oder bei welchen Aussagen es eine besonders große Differenz zwischen den mittleren Urteilen zweier Gruppen gibt.

Innerhalb dieser Arbeiten kann noch einmal unterschieden werden zwischen Veröffentlichungen, die der Beschreibung der unmittelbaren Kandidatenbewertungen einen größeren Umfang einräumen, und Arbeiten, in denen Echtzeiturteile zu bestimmten Inhalten illustrativ zur Interpretation von Wirkungen der Debatte auf die Einstellungen der Zuschauer herangezogen werden. Zum zweiten Typ gehören unter anderem die Arbeiten von Delli Carpini, Keeter und Webb (1997), McKinnon und Tedesco (1993, 1999) sowie McKinney, Kaid und Robertson (2001). Diese Arbeiten nennen lediglich eine geringe Zahl besonders positiv bzw. negativ bewerteter Aussagen der Kandidaten, die sie dann hinsichtlich der Passung zu den Veränderungen der Kandidatenimages durch die Debatte interpretieren.

Mit einer ähnlichen Zielsetzung präsentieren wir in unserer Analyse zur Wirkung des bildungspolitischen Teils im TV-Duell zwischen Stefan Mappus und Nils Schmid die unmittelbaren Bewertungen der einzelnen Kandidatenaussagen zur Schulpolitik (Bachl & Vögele, 2013; Vögele & Schmalz, 2013). Mappus erfuhr für seine Aussagen, in denen er sich klar gegen Veränderungen des bestehenden Schulsystems aussprach, der Opposition jedoch derartige Pläne unterstellte, große Zustimmung im eigenen Lager und auch bei den Unentschiedenen. Sogar die Anhänger der Opposition stimmten einigen seiner Aussagen leicht zu und lehnten selbst seine Angriffe auf die schulpolitischen Pläne Schmids nicht stark ab. Dagegen polarisierten die Ausführungen Schmids zum Konzept des „Längeren gemeinsamen Lernens“ zwischen den Lagern: Sie wurden von den Anhängern der Regierungsparteien und teils auch von den Unentschiedenen negativ bewertet und erhielten nur im rot-grünen Lager

einige Zustimmung. Schmids Plan, eine Wahlmöglichkeit zwischen einem acht- und einem neunjährigen Weg zum Gymnasium einzuführen, wurde über die Lager hinweg begrüßt. Diese deskriptive Analyse der RTR-Wertungen hilft uns, die Veränderung der bildungspolitischen Kompetenzzuschreibungen, die sich durch die Debattenrezeption ergaben, besser zu verstehen. Mappus wurde nach dem Duell von den Zuschauern fast aller Lager besser bewertet, während Schmid von einem Großteil der Rezipienten als weniger kompetent eingeschätzt wurde. Die Echtzeit-Urteile während des schulpolitischen Debattenteils tragen zudem dazu bei, die Bewertung der bildungspolitischen Kompetenz beider Kandidaten nach der Debatte zu erklären. In Anbetracht der unmittelbaren Bewertungen der schulpolitischen Aussagen beider Kandidaten liegt der Schluss nahe, dass diese Veränderungen für einen Erfolg von Mappus' Strategie sprechen, sich als „Verteidiger" des von Schmid „bedrohten" Schulsystems zu positionieren. In diesem Sinne unterstützen die Interpretationen der deskriptiv dargestellten unmittelbaren Kandidatenbewertungen unsere Argumentation, auch wenn eine kausale Verknüpfung so natürlich nicht hergestellt werden kann.

Ausführlichere Analysen der Kandidatenbewertungen im Debattenverlauf liegen für die TV-Duelle vor den Bundestagswahlen 2002, 2005 und 2009 sowie für das baden-württembergische TV-Duell 2011 vor. Maier und Faas untersuchen die unmittelbare Bewertung von Kanzler Gerhard Schröder und Herausforderer Edmund Stoiber in den beiden Debatten vor der Bundestagswahl 2002 (J. Maier & Faas, 2003, 2004; Faas & Maier, 2004a). 32 (erstes Duell) bzw. 35 (zweites Duell) nach einem Quotenplan (Geschlecht, Alter, Bildung) rekrutierte Personen sahen die Debatten an der Universität Bamberg. In zwei Gruppen verfolgten die Probanden das Duell entweder wie im Fernsehen ausgestrahlt mit Bild und Ton oder nur mit Ton. Sie bewerteten die Kandidaten während der Rezeption im Gegensatz zu den übrigen hier vorgestellten Studien mit einem Push-Button-System: Vier Tasten einer Computertastatur waren mit den Bedeutungen „Schröder positiv", „Schröder negativ", „Stoiber positiv" und „Stoiber negativ" versehen und konnten von den Rezipienten jederzeit zur Abgabe eines Urteils gedrückt werden. Für die Analysen werden die Bewertungen für das gesamte Testpublikum sowie für Teilgruppen nach Kandidatenlager (Anhänger Schröder: Parteiidentifikation für SPD oder Grüne, $n_{1.\,\text{Duell}} = 11$, $n_{2.\,\text{Duell}} = 14$; Anhänger Stoiber: CDU/CSU oder FDP, $n_{1.\,\text{Duell}} = 15$, $n_{2.\,\text{Duell}} = 15$) betrachtet.

Die Autoren sammeln die Aussagen, in denen die Kandidaten besonders große Zustimmung erhielten, und ordnen diese thematisch ein: „So konnten die zwei Kontrahenten in den beiden Debatten mit bekannten Themen punkten: Stoiber mit den Bereichen Wirtschaft und Arbeit, Schröder mit der Irak-Frage,

aber auch mit seiner klaren Absage an die PDS“ (Faas & Maier, 2004a, S. 69). Dieser Befund lässt sich gut mit dem Ansatz des *Issue Ownership* (allgemein Petrocik, 1996; Petrocik, Benoit & Hansen, 2003; für eine Anwendung im deutschen Kontext Franzmann, 2006) erklären: Der Ansatz geht davon aus, dass die Parteien bei bestimmten Themen meist über einen längeren Zeitraum, teilweise aber auch im Kontext bestimmter Ereignisse, von den Wählerinnen und Wählern als kompetent wahrgenommen werden. Demzufolge ist es nicht überraschend, dass die Kandidaten in den TV-Debatten größere Zustimmung erhielten, wenn sie Aussagen zu Themen tätigten, in denen sie bzw. ihre Partei bereits vor der Debatte als Kompetenzführer galten.

Die zweite wichtige Erkenntnis der Autoren ist die Bestätigung des erwarteten großen Einflusses der Voreinstellungen auf die unmittelbare Bewertung der Kandidaten während der Duelle. Für beide Debatten zeigt sich, dass die Zuschauer eines Lagers die Aussagen ihres Kandidaten (fast) durchgängig positiv bewerteten. Trotz dieses starken Effekts verbleibt in den unmittelbaren Urteilen der eigenen Anhänger über die Zeit noch einige Varianz, die auf die unterschiedliche Bewertung der jeweiligen Debatteninhalte zurückgeführt werden kann. Kaum ausgeprägt war dagegen der Effekt der Lagerzugehörigkeit auf die unmittelbare Bewertung des gegnerischen Kandidaten. Zwar war der Saldo der Bewertungen jeweils zusammengefasst über den Verlauf der beiden Debatten negativ, für einzelne Passagen der Duelle erhielten beide Kandidaten aber auch von diesen Teilgruppen Zustimmung. Die Autoren folgern aus diesem Effektmuster, dass die Debatten vor allem bestehende Voreinstellungen bestärkten, sie aber auch (zumindest hinsichtlich einiger Themen) ein gewisses Persuasionspotenzial besaßen.

In einer weiteren Auswertung unterscheiden Faas und Maier (2004b) zwischen den Sehern und den Hörern der Debatten. Die Ergebnisse zeigen einige Unterschiede in den unmittelbaren Urteilen über die Kandidaten, am stärksten sind sie bei der Bewertung des jeweils gegnerischen Kandidaten ausgeprägt. Da die Gruppengrößen bei einer Aufteilung der Stichprobe nach Lager und Rezeptionsmodus jedoch äußerst gering ausfallen ($n_{min} = 4$; $n_{max} = 9$), sollten diese Ergebnisse vorsichtig interpretiert werden.

Dem zweiten TV-Duell zwischen Schröder und Stoiber widmen sich auch Maurer und Reinemann (Maurer & Reinemann, 2003; Reinemann & Maurer, 2005). Eine selbstselektive Stichprobe von 75 Personen sah das Duell an der Universität Mainz und bewertete die Kandidaten mit 7-stufigen RTR-Dials. Die RTR-Skala reichte von 1 = „Schröder ist gut/Stoiber ist schlecht“ bis 7 „Stoiber ist gut/Schröder ist schlecht“ (vgl. zu dieser Skalengestaltung auch Kapitel 4). In der ausführlichen Dokumentation der Studie (Maurer & Reinemann, 2003) identifizieren die Autoren zunächst die vom gesamten Publikum

am besten bzw. am schlechtesten bewerteten Passagen. Zudem werden die unmittelbaren Kandidatenbewertungen im Zeitverlauf getrennt nach Kandidatenlagern (Anhänger Schröder: Parteiidentifikation für SPD oder Grüne, n = 39; Anhänger Stoiber: CDU/CSU oder FDP, n = 23; Ungebundene, n = 13) und Geschlecht (n_{Frauen} = 32; $n_{Männer}$ = 43) untersucht. Dabei werden auch die Differenzen zwischen den mittleren Bewertungen durch Schröder- und Stoiber-Anhänger sowie durch Männer und Frauen im Duellverlauf analysiert, um besonders stark zwischen den Gruppen polarisierende Aussagen zu entdecken. In einer weiteren Veröffentlichung (Reinemann & Maurer, 2005) erfolgen alle Analysen auf Basis der nach Parteiidentifikation eingeteilten Gruppen. Gesucht wird zum einen nach Aussagen, die in allen drei Gruppen im Mittel überdurchschnittlich positiv bewertet wurden, zum anderen wiederum nach Aussagen, die überdurchschnittlich stark zwischen den Anhängern der beiden Kandidaten polarisierten. Als theoretischen Bezugsrahmen wählen Maurer und Reinemann (2003, S. 88) die „Rhetorik- und Persuasionsforschung“, die auch dabei helfen soll, aus den Befunden „allgemeine Aussagen darüber abzuleiten, wie Debatten gewonnen und verloren werden“.

Aussagen, die vom gesamten Publikum positiv bewertet wurden, thematisierten vor allem die eigenen Ansichten und Pläne. Besonders häufig fanden sich unter den sehr positiv beurteilten Aussagen „quasi zustimmungspflichtige Gemeinplätze“ (Maurer & Reinemann, 2003, S. 101), die so vage formuliert waren, dass niemand ihnen widersprechen konnte. Ihr Erfolg wurde noch wahrscheinlicher, wenn sie in Kombination mit emotionalen Appellen und in stringent aufgebauten Statements verwendet wurden. Konkrete Aussagen der Kandidaten über politische Maßnahmen oder Personalentscheidungen polarisierten dagegen das Publikum entlang der Parteilinien. Im Zuge des Gruppenvergleichs finden die Autoren – wie auch die zuvor dargestellten Arbeiten von Faas und Maier – einen deutlichen Einfluss der Lagerzugehörigkeit auf die Bewertung der Kandidaten während der Debatte, der sich wiederum darin zeigt, dass die Zustimmung zu den Aussagen des eigenen Kandidaten besonders stark war. Die Autoren interpretieren ihre Befunde dahingehend, dass in der Kommunikationssituation eines TV-Duells die Verwendung von emotionalen Appellen der Argumentation mit Evidenzen überlegen war, wenn es darum geht, die Zustimmung eines möglichst großen Teils des Publikums zu erlangen. Konkrete, mit Evidenzen gestützte Aussagen wurden naturgemäß von den Anhängern des Gegners abgelehnt, und sie halfen auch kaum dabei, die ungebundenen Zuschauer zu überzeugen.

Die umfassende Studie zum TV-Duell zwischen Bundeskanzler Gerhard Schröder und Herausforderin Angela Merkel vor der Bundestagswahl 2005 (Maurer et al., 2007) enthält auch eine ausführliche Analyse der unmittelbaren

Kandidatenbewertung im Duellverlauf (Reinemann & Maurer, 2007b, 2007c). Die Stichprobe verteilte sich auf zwei Standorte: die Universitäten in Jena (n = 49) und Mainz (n = 72). Innerhalb der Teilstichproben an den Standorten wurde auf eine Gleichverteilung der Merkmale Geschlecht, Alter, Bildung und politisches Lager (in Jena inkl. der Linkspartei.PDS als eigene Gruppe) geachtet. Zur Erfassung der unmittelbaren Kandidatenbewertungen wurden 7-stufige RTR-Dials entsprechend der Vorgängerstudie (Maurer & Reinemann, 2003) eingesetzt. Nach dem bereits für die Vorgängerstudie beschriebenen Prinzip werden die Bewertung der Kandidaten im Debattenverlauf in Hinblick auf besonders erfolgreiche bzw. zwischen den Lagern polarisierende Aussagen analysiert, wobei die Urteile der Probanden an den beiden Standorten immer getrennt behandelt werden. Untersucht werden die RTR-Bewertungen für das gesamte Publikum in Jena und Mainz und getrennt nach Kandidatenlagern (Anhänger Schröder: Parteiidentifikation für SPD oder Grüne, $n_{Jena} = 14$, $n_{Mainz} = 26$; Anhänger Merkel: CDU/CSU oder FDP, $n_{Jena} = 12$, $n_{Mainz} = 27$; Ungebundene: $n_{Jena} = 13$, $n_{Mainz} = 17$). Das analytische Vorgehen wird darüber hinaus gegenüber der Vorgängerstudie um zwei Komponenten ergänzt: Zum einen werden die mittleren RTR-Bewertungen der Kandidaten nach den im Duell diskutierten Themenblöcken zusammengefasst und vergleichend für das Publikum in Jena und Mainz betrachtet. Zum anderen werden die beim gesamten Publikum und bei den Teilgruppen besonders erfolgreichen sowie die stark zwischen den Kandidatenlagern polarisierenden Aussagen mit einer systematischen Inhaltsanalyse nach Aussagetypen und Themen kategorisiert.

Die Interpretation der Befunde lehnt sich stark an die Vorgängerstudie (Maurer & Reinemann, 2003) an. Die Autoren stellen heraus, dass wiederum allgemeine Aussagen über die Pläne und Ideen am häufigsten unter den beim gesamten Publikum sehr erfolgreichen Passagen zu finden waren. Polarisiert haben die Kandidaten zwischen den Lagern vor allem dann, wenn sie sich kritisch über den politischen Gegner äußerten – auch dann, wenn sie diese Angriffe mit belegbaren Fakten stützten. Neben diesen an der Rhetorik- und Persuasionsforschung orientierten Interpretationen rückt auch die Themenperspektive in dieser Analyse stärker in den Vordergrund. Die meisten als erfolgreich identifizierten Statements fielen in die Themenbereiche Steuerpolitik und Außenpolitik. Schließlich weisen die Autoren ein weiteres Mal den Einfluss der Lagerzugehörigkeit auf die unmittelbare Wahrnehmung der Kandidaten nach. Der Effekt der Parteiidentifikation fiel in Jena schwächer aus als in Mainz – neben einigen durch die unterschiedliche Zusammensetzung der Teilstichproben und inhaltlich spezifisch auf Ostdeutschland zugeschnittene Aussagen zu erklärenden Abweichungen der einzige systematischen Unterschied zwischen Ost und West.

Eine spezifische Auswertung der Daten zur Rezeption des TV-Duells Schröder gegen Merkel legen Reinemann und Maurer (2010) vor. Vor dem Hintergrund von Überlegungen zur individuellen Informationsverarbeitung in Abhängigkeit von den kognitiven Fähigkeiten und vom Involvement gehen die Autoren der Frage nach, ob die Kandidaten von interessierten und weniger interessierten Rezipienten unterschiedlich beurteilt wurden. Konkret stützen sie ihre Überlegungen auf das RAS-Modell (Zaller, 1992) und das Elaboration-Likelihood-Modell (Petty & Cacioppo, 1986). Ein Vergleich der mittleren RTR-Bewertungen durch Interessierte (n = 80) und weniger Interessierte (n = 21) im Debattenverlauf zeigt eine geringe Zahl von Abweichungen. Diese lassen sich jedoch nicht systematisch durch Eigenschaften der zuvor geäußerten Inhalte (z.B. Stärke der Argumente) klassifizieren. Auf Basis dieser Ergebnisse schließen die Autoren, dass bei der Verarbeitung eines zentralen Medienereignisses wie des TV-Duells möglicherweise das situationale Involvement so groß ist, dass die Unterschiede zwischen generell mehr und weniger interessierten Rezipienten aufgehoben werden. Sie geben allerdings auch zu bedenken, dass die künstliche Rezeptionssituation im Labor zu diesem Befund beigetragen haben könnte.

Schließlich vergleichen M. Maier und Strömbäck (2009) die unmittelbare Kandidatenbewertung während des TV-Duells zwischen Schröder und Merkel durch die Teilstichprobe in Jena mit einer ähnlich durchgeführten Studie zum TV-Duell zwischen Premierminister Göran Persson und seinem wichtigsten Herausforderer Fredrik Reinfeldt vor der Parlamentswahl 2006 in Schweden. Auch in der an der Universität Sundsvall durchgeführten Studie wurden die Probanden nach einem Quotenplan mit den Merkmalen Geschlecht, Alter, Bildung und Parteiidentifikation rekrutiert. Für die RTR-Messung in Sundsvall wurde eine 7-stufige, eindimensionale Skala verwendet, die von 1 = „sehr schlechter Eindruck vom Sprecher“ bis 7 = „sehr guter Eindruck vom Sprecher“ reichte. Für beide Debatten werden zuerst die vom gesamten Publikum sehr positiv bewerteten Aussagen der Kandidaten beschrieben. Es folgt eine detailliertere Analyse der Debattenausschnitte zum Thema Arbeitslosigkeit, deren Wahrnehmung in Abhängigkeit von der Parteiidentifikation untersucht wird. Die Autoren stellen heraus, dass die Kandidaten in beiden Debatten mit allgemeinen wie mit konkreten Aussagen beim gesamten Publikum erfolgreich waren. Zudem gab es in beiden Debatten kontextabhängige Interaktionen zwischen den Inhalten der Aussagen und der Parteiidentifikation, die sowohl zu zwischen den Kandidatenlagern polarisierenden wie auch zu in allen Lagern erfolgreichen Statements führten.

Die unmittelbare Bewertung von Bundeskanzlerin Angela Merkel und ihrem Herausforderer Frank-Walter Steinmeier im TV-Duell vor der Bundestagswahl

2009 wurde bislang lediglich in einem Konferenzbeitrag ausführlicher dargestellt (Faas et al., 2009). Steinmeier wurde vom gesamten Publikum am besten bewertet, wenn er sich für soziale Gerechtigkeit – traditionell ein Kernthema der SPD (Franzmann, 2006) – stark machte. Doch auch Merkel wurde am positivsten wahrgenommen, als sie den sozialen Aspekt der sozialen Marktwirtschaft und des deutschen Gesundheitssystems betonte. Die Diskussion über das Thema Atomkraft polarisierte zwischen den Lagern. Insgesamt waren die unmittelbaren Kandidatenbewertungen während der Debatte verhältnismäßig moderat ausgeprägt und bewegten sich im Rahmen dessen, was der Kontext des Wahlkampfs hatte erwarten lassen (vgl. auch Bachl & Brettschneider, 2011; Brettschneider & Bachl, 2009).

Auch in unserer Studie zum TV-Duell zwischen Stefan Mappus und Nils Schmid vor der baden-württembergischen Landtagswahl 2011 befassen wir uns ausführlich mit der Bewertung der Kandidaten im Duellverlauf (Bachl, 2013a; Bachl & Brettschneider, 2013, zu Stichprobe und RTR-Messung vgl. Kapitel 4). Bei den Auswertungen und Darstellungen der Ergebnisse orientieren wir uns am Vorgehen früherer Arbeiten, insbesondere an Reinemann und Maurer (2007b). Wir untersuchen die über das gesamte Duell und über die einzelnen Themenblöcke zusammengefassten Echtzeit-Urteile für alle Zuschauer und die Kandidatenlager (Anhänger Mappus: Wahlabsicht für CDU oder FDP, n = 48; Anhänger Schmid: SPD oder Grüne, n = 89; Unentschiedene, n = 39) und beschreiben die vom gesamten Publikum und den Teilgruppen nach Lager besonders positiv bewerteten Aussagen. Zudem werden die Stellen des Duells gesucht, in denen sich die mittleren Kandidatenbewertungen durch die Anhänger der beiden Kandidaten am stärksten voneinander unterschieden. Abschließend verknüpfen wir die Daten einer sekundengenauen Inhaltsanalyse des Duells (Bachl, Käfferlein & Spieker, 2013b) mit den als besonders positiv bzw. polarisierend identifizierten Passagen, um einen systematischen Überblick über ihre inhaltlichen Charakteristika zu erhalten.

Wichtigster Rahmen für die Interpretation unserer Befunde ist eine „Themenperspektive auf Wahlkämpfe und TV-Duelle“ (Bachl & Brettschneider, 2013, S. 96), die auf den bereits erläuterten Überlegungen zum Issue Ownership (Petrocik, 1996; Petrocik et al., 2003; Franzmann, 2006) und zum Themenmanagement in Wahlkämpfen (Brettschneider, 2005a, 2005b; Hinrichs, 2002) aufbaut. Außerdem ziehen wir den hier bzw. im folgenden Abschnitt 3.4.2 beschriebenen Forschungsstand zur unmittelbaren Bewertung von Debatteninhalten (u.a. Angriffe vs. Selbstpräsentation, Ambiguität vs. Konkretheit) zur Einordnung heran. In Übereinstimmung mit dem Forschungsstand weisen auch unsere Ergebnisse einen starken Einfluss der Lagerzugehörigkeit auf die unmittelbare Kandidatenbewertung nach. Wiederum zeigt sich dieser vor allem in einer

stärkeren durchschnittlichen Zustimmung der Lager zu ihrem Kandidaten und nur selten in einer im Gruppendurchschnitt negativen Bewertung des anderen Kandidaten. Die gewählte Themenperspektive eignet sich gut, um die erfolgreichen Aussagen der Kandidaten beim gesamten Publikum und in den einzelnen Teilgruppen zu erklären. Schmids Aussagen wurden häufig dann gut bewertet, wenn sie die Energiepolitik oder Fragen der sozialen Gerechtigkeit thematisierten. Mappus erhielt große Zustimmung mit Aussagen zu den traditionellen CDU-Themen Finanzen, Steuern und Wirtschaft. Unklar war der Zusammenhang zwischen Ambiguität bzw. Konkretheit der Aussagen und der Zustimmung: Unter den erfolgreichen Statements beider Kandidaten fanden sich sowohl sehr konkrete Ausführungen ihrer Pläne, als auch Gemeinplätze und Wertreferenzen. Mit Einschränkungen findet sich auch eine polarisierende Wirkung von Angriffen: Allerdings nur, wenn sie von Mappus stammten und wenn sie Themen behandelten, bei denen die Positionen der Lager klar unterscheidbar waren. Andere Angriffe der Kandidaten waren auch unter den Aussagen zu finden, die vom gesamten Publikum und/oder den Unentschiedenen sehr positiv bewertet wurden.

Schließlich befassen sich zwei Studien detaillierter mit der rezeptionsbegleitend gemessenen Bewertung von Kandidaten während TV-Duellen in US-Präsidentschaftswahlkämpfen. Jarman (2005) untersucht die unmittelbaren Reaktionen der Rezipienten auf die Debatte zwischen Präsident George W. Bush und seinem Herausforderer John Kerry vor der Präsidentschaftswahl 2004. 36 Probanden, die aus einem lokalen Wählerverzeichnis zufällig gezogen wurden, bewerteten die Kandidaten während des Duells mit RTR-Dials. Zur Bewertung diente eine eindimensionale, 11-stufige RTR-Skala von 0 = „extreme disagreement“ bis 10 = „extreme agreement“. Anhand der für Republikaner (n = 19) und Demokraten (n = 15) zusammengefassten RTR-Bewertungen werden die Aussagen identifiziert, in denen die beiden Kandidaten von den Gruppen jeweils am besten und am schlechtesten bewertet wurden. Die Einordnung der Befunde erfolgt auch hier in Bezug zum Ansatz des Issue Ownership, der sich größtenteils bewährt. Allerdings stellt Jarman (2005, S.240) auch fest: „There were times when Kerry scored highly on a traditionally Republican issue and when Bush scored well on a traditionally Democratic issue“. Als weitere Auswertung wird für jede der 56 Antworten der Kandidaten geprüft, ob sich die Bewertung von Republikanern und Demokraten signifikant unterschied. Dies war bei 52 Antworten der Fall, was abermals für einen starken Einfluss der Lagerzugehörigkeit spricht.

Mit der Bewertung der Kandidaten in den TV-Duellen vor den US-Präsidentschaftswahlen 2008 (Obama gegen McCain) und 2012 (Obama gegen Romney) durch noch unentschiedene Rezipienten beschäftigen sich Schill und

Kirk (2013). Gruppen von n_{min} = 25 bis n_{max} = 39 Unentschiedenen, die nach einer Kombination von Quoten- und Zufallsstichprobe aus lokalen Wählerverzeichnissen gezogen wurden, bewerteten die Kandidaten während eines der sechs TV-Duelle mit RTR-Dials auf einer eindimensionalen, 100-stufigen Skala (Endpunkte: negativ und positiv). Eine Zusammenfassung der am besten bewerteten Passagen aller Debatten identifiziert fünf Arten von Statements, die besonders große Zustimmung bei unentschiedenen Wählern erzielten:

> (a) making clear, broad, and affirmative policy statements, (b) affirmation of America's core values, (c) projection of candidate strength and directness in foreign policy, (d) appeals to middle-class voters and values, and (e) praising the military and our troops (Schill & Kirk, 2013, S. 11).

Kleinliche und übermäßig scharfe Angriffe führen dagegen zu negativen Bewertungen. Übertragen auf die USA bestätigen die Autoren damit zu großen Teilen die Befunde zu den deutschen TV-Debatten 2002 und 2005 (Reinemann & Maurer, 2005, 2007b).

Zusammenfassung

Das vorrangige Ziel der vorgestellten Studien ist es, die Bewertungen der Kandidaten im Verlauf der Debatten darzustellen. In den meisten Fällen werden die besonders positiv und/oder negativ bewerteten Aussagen identifiziert. Teilweise werden auch die Passagen der Debatte näher betrachtet, bei denen eine besonders große Differenz zwischen den Bewertungen durch zwei Gruppen auftrat. Den Studien liegt damit hinsichtlich der Wirkungen der Debatteninhalte eine explorative, induktive Analyselogik zugrunde: Ausgangspunkt der Analysen sind die mit RTR-Messungen beobachteten Reaktionen des Publikums im Zeitverlauf – und damit aus analytischer Sicht gesprochen die abhängige Variable. Mit ihrer Hilfe werden auffällige Bewertungen identifiziert. Von den Stellen, an denen diese auftreten, wird dann auf besonders wirksame Debatteninhalte – aus analytischer Sicht die unabhängige Variable – zurückgeschlossen.

Die Reichweite der so gewonnenen Befunde beschränkt sich nach der explorativen, induktiven Logik auf die Wahrnehmung der jeweils untersuchten Debatte durch das Testpublikum. Darüber hinaus werden die Ergebnisse teils als Indikatoren für die Bewertung der Aussagen durch andere Rezipienten der Debatten außerhalb der Stichprobe oder für Urteile über die als wirksam identifizierten Botschaften im jeweiligen Wahlkampf interpretiert. Folgerichtig

werden in diesen Analysen keine Hypothesen über die Beurteilung bestimmter Inhaltsmerkmale aufgestellt und getestet.[18] Teils werden jedoch auf Basis der Literatur Erwartungen formuliert, wie welche Gruppen des Publikums auf welche Inhalte reagieren könnten. Solche Überlegungen liefern auch die Basis für *ex-post*-Erklärungen, warum gerade diese Aussagen induktiv als wirksam identifiziert wurden.

Neben dem Rückgriff auf den Forschungsstand früherer Debattenstudien werden zwei Ansätze zur Formulierung der Erwartungen bzw. *ex-post*-Erklärung der Ergebnisse herangezogen. Über den Ansatz des Issue Ownership (Petrocik, 1996; Petrocik et al., 2003) wird erklärt, warum das Publikum (oder ein bestimmter Teil des Publikums) die Kandidaten bei bestimmten Themen besonders häufig überdurchschnittlich positiv bewertet. Dieser Ansatz ist im Kontext der deskriptiven Studien besonders hilfreich, da er keine Aussagen über die individuelle Informationsverarbeitung einzelner Rezipienten macht, sondern lediglich die Stärken und Schwächen der Parteien im Elektorat benennt. Dadurch ist eine Passung zu den größtenteils auf Basis von (Gruppen-) Aggregaten durchgeführten deskriptiven Analysen gegeben (vgl. im Detail Kapitel 5). Da der Ansatz außerdem recht wenige Randbedingungen setzt und einfach auf den größeren Kontext des Wahlkampfs, in dem eine Debatte stattfindet, zu beziehen ist, stellt er sich in den hier betrachteten Studien als gut geeignet für die plausible Erklärung der explorativ generierten Ergebnisse heraus.

Weitere Interpretationen der Befunde stützen sich auf die im weiteren Sinne rhetorische Gestaltung der als wirksam identifizierten Aussagen. Diese Arbeiten kommen zu dem Schluss, dass allgemein gehaltene Aussagen und emotionale Appelle auffällig häufig unter den über die Lager hinweg positiv bewerteten Passagen zu finden sind, da sie weniger Widerspruch hervorrufen. Konkrete Aussagen, mit Evidenzen gestützte Argumente sowie Angriffe auf den politischen Gegner finden sich dagegen häufiger unter den Passagen, die zwischen den Anhängern der Kandidaten polarisieren. Auch diese Befunde lassen sich im Nachhinein gut mit den Erkenntnissen der Persuasionsforschung plausibilisieren (vgl. dazu den ausführlichen Überblick bei Maurer & Reinemann, 2003, S. 121-133). Dass diese Interpretationen auf Basis der hier vorgestellten induktiven Studien jedoch nicht zu generalisieren sind, sondern es sich lediglich um theoriegestützte Interpretationen empirischer Indikatoren

18 Eine Ausnahme ist die Studie von McKinnon und Tedesco (1993). Die Autoren formulieren die sehr unspezifische Hypothese, dass sowohl die besten als auch die schlechtesten Bewertungen der Kandidaten in Diskussionen um spezifische Policy-Themen auftreten. Die Hypothese wird falsifiziert. Da die Analyselogik trotz des Hypothesentests eine induktive ist, wird die Studie hier und nicht in Abschnitt 3.4.2 eingeordnet.

handelt, zeigen auch die Widersprüchlichkeiten einiger Veröffentlichungen. So sehen Reinemann und Maurer (2007b) die Ergebnisse der Bewertung von Schröder und Merkel im TV-Duell 2005 als einen weiteren Hinweis darauf, dass die Verwendung von Allgemeinplätzen eine erfolgreiche rhetorische Strategie ist. In einer Interpretation eines Teildatensatzes derselben Studie nennen M. Maier und Strömbäck (2009) dagegen ausschließlich spezifische Policy-Aussagen als Beispiele für die erfolgreichsten Statements.

Ein übereinstimmender Befund sämtlicher gesichteter Studien ist der große Einfluss der politischen Voreinstellungen – hier meist operationalisiert als die Zugehörigkeit zum Lager eines der Kandidaten – auf die Bewertung der Kandidaten während der Debatten: Die Zuschauer nehmen die Duelle „durch ihre parteipolitische Brille“ (Maurer & Reinemann, 2007b, S. 232) wahr. Dies ist in Anbetracht der aus der politischen Wahl- und Einstellungsforschung bekannten zentralen Rolle des politischen Orientierungssystems, vor dessen Hintergrund politische Informationen eingeordnet und bewertet werden, für sich genommen noch nicht weiter überraschend (Greene, 2002; Zaller, 1992). Interessant sind für weitere Studien zur unmittelbaren Kandidatenbewertung in TV-Debatten jedoch zwei spezifische Details: *Erstens* wirkte sich die Voreinstellung zumeist nach einem bestimmten Muster aus. Der Kandidat, der gerade das Wort hatte, wurde von seinem eigenen Lager im Mittel fast immer positiv bewertet – Unterschiede ergaben sich hier nur im Ausmaß der Zustimmung. Die „Gegner“ des Sprechers bewerteten ihn im Aggregat aber *nicht* automatisch negativ. Oder wie es Faas und Maier (2004a, S. 61) prägnant auf den Punkt bringen: „[E]s wird eher ‚gecheert‘ als ‚geboot‘“. *Zweitens* übt die Lagerzugehörigkeit zwar einen starken Einfluss auf die unmittelbaren Bewertungen aus, sie determiniert sie jedoch nicht. Die Zustimmung des eigenen Lagers fällt im Verlauf der Debatte unterschiedlich stark aus, das gegnerische Lager bewertet meist neutral oder negativ, kann aber von Zeit zu Zeit zu einer Zustimmung bewegt werden. Es scheint also durchaus sinnvoll, nach Merkmalen des Debatteninhalts zu suchen, die diese Varianz über die Zeit erklären können. Dabei sollte allerdings beachtet werden, dass die Befunde der Arbeiten, die getrennte Auswertungen nach Lagern vornehmen, deutlich auf eine Interaktion zwischen Lagerzugehörigkeit und Debatteninhalten hinweisen. Eine ausschließlich auf die Inhalte fokussierende Analyselogik, in der die Voreinstellungen vernachlässigt werden, ist in Anbetracht ihres bereits in diesen einfachen Analysen deutlich sichtbaren Einflusses nicht ratsam.

Zusammenfassend kann festgehalten werden, dass der besondere Wert der dargestellten deskriptiven Studien darin liegt, detaillierte, empirisch fundierte Beschreibungen der unmittelbaren Kandidatenbewertungen während eines zentralen Ereignisses des Medienwahlkampfs zu liefern. So kann es die Ana-

lyse eines Wahlkampfs bereichern, zu wissen, mit welchen Aussagen welcher Kandidat sein eigenes Lager hinter sich vereinen konnte, oder welche Aussagen dazu geeignet waren, auch unter den noch Unentschiedenen oder gar im gegnerischen Lager Zustimmung zu generieren. Wie eingangs des Kapitels dargestellt, können diese Erkenntnisse der deskriptiven Studien zudem dazu genutzt werden, Debattenwirkungen auf die Einstellungen der Rezipienten besser zu verstehen. Nicht zuletzt können diese Analysen auch dabei helfen, die öffentliche und massenmediale Interpretation dieser Medienereignisse mit empirischen Befunden zu bereichern (Schill & Kirk, 2009). So konnten beispielsweise Reinemann und Maurer (2007b, S. 74, 76, 88) anhand ihrer RTR-Messungen zum TV-Duell 2005 zeigen, dass die medial viel diskutierte (Reinemann, 2007) „Liebeserklärung" Schröders an seine Frau zumindest bei den Testzuschauern keine bemerkenswerten Reaktionen auslöste.

Die *ex-post*-Erklärung der als „wirksam" identifizierten Debatteninhalte auf der Basis sinnvoll ausgewählter theoretischer Konzepte ist generell wünschenswert und kann zu einer weiteren kommunikationswissenschaftlichen Fundierung der Beurteilung von Kandidatenauftritten in Fernsehdebatten beitragen. Allerdings sollten bei der Generalisierung der Befunde immer auch die Grenzen des induktiven Vorgehens und der Fallbeispielcharakter einzelner TV-Duelle beachtet werden. Zudem sind die in den gesichteten Studien eingesetzten datenanalytischen Verfahren – wie im Folgenden noch zu zeigen sein wird – nur begrenzt dazu geeignet, Rückschlüsse auf die individuelle Informationsverarbeitung der Rezipienten zu ziehen. Die Befunde der dargestellten Studien eignen sich dazu, Hypothesen für zukünftige TV-Duell-Studien bzw. Studien zur Verarbeitung politischer Kommunikation im Allgemeinen zu generieren. Sie selbst sind aber nicht dazu angelegt, Annahmen über die Wirkung bestimmter Merkmale der Debatteninhalte zu prüfen.

3.4.2 Deduktive Analysen

Die im Folgenden vorgestellten Analysen zur Erklärung der unmittelbaren Kandidatenbewertung sind im Gegensatz zu den Arbeiten in Abschnitt 3.4.1 in ihrer Grundlogik deduktiv angelegt. Sie unterscheiden sich durch drei wesentliche Charakteristika von den induktiven Analysen:

- *Theoretisches Vorgehen*: Es werden *a priori* Annahmen über die Wirkung bestimmter Merkmale der Debatteninhalte getroffen.
- *Inhaltsanalytisches Vorgehen*: Die Merkmale, deren Wirkung überprüft werden soll, werden unabhängig von den Messungen der unmittelbaren

Kandidatenbewertung im gesamten zu untersuchenden Stimulusmaterial (das gesamte Duell oder Ausschnitte des Duells) mit systematischen Inhaltsanalysen identifiziert.

- *Datenanalytisches Vorgehen*: Um die Annahmen über die Effekte bestimmter Merkmale des Debatteninhalts auf die unmittelbaren Kandidatenbewertungen überprüfen zu können, werden inferenzstatistische Verfahren eingesetzt. Zudem muss, um der komplexen Datenstruktur der RTR-Messungen im Zeitverlauf gerecht zu werden, irgendeine Art der Datenreduktion gewählt werden. Fragen des datenanalytischen Vorgehens deduktiver Studien werden in Kapitel 5.3 diskutiert.

Insgesamt liegen uns vier Analysen (bzw. fünf Publikationen) vor, die sich mit der Wirkung bestimmter Merkmale des Debatteninhalts auf die unmittelbare Bewertung der Kandidaten durch die Rezipienten befassen. Gemeinsam haben alle Arbeiten, dass sie (neben anderen Aspekten) untersuchen, ob die Reaktionen des Publikums davon beeinflusst werden, ob ein Kandidat in einer Aussage über sich selbst oder über den politischen Gegner spricht. Da wir in der vorliegenden Arbeit ebenfalls dieser Frage nachgehen, soll das zugrundeliegende Konzept zur besseren Nachvollziehbarkeit der folgenden Ausführungen kurz beschrieben werden.

Exkurs: Relationale Strategien in TV-Duellen

Bereits Aristoteles (Übers. 2007, S. 201) unterscheidet in seinem Werk zur Rhetorik zwei grundsätzliche Strategien, mit denen sich ein Redner im Vergleich zu seinem Kontrahenten besser stellen kann: Entweder kann er sich selbst positiv darstellen und damit das Publikum davon überzeugen, dass er selbst der bessere Kandidat ist. Oder er kann seinen Gegner negativ darstellen und damit dem Publikum zeigen, dass der Gegner der schlechtere Kandidat ist. Der Effekt wäre im Erfolgsfall der gleiche: Relativ zu seinem Gegner stände der Redner besser da. In der Forschung zu TV-Debatten werden diese beiden Grundstrategien immer wieder aufgegriffen und dabei um verschiedene Punkte ergänzt und differenziert. Martel (1983, S. 62-74) nennt Angreifen, Verteidigen, Verkaufen, Ignorieren und ‚ich auch – aber besser'-Aussagen. W. L. Benoit (2007, S. 36-43) kategorisiert in seiner *functional analysis of campaign discourse*, mit der sich nicht nur die Kandidatenaussagen in TV-Debatten in verschiedensten Ländern,[19]

19 z.B. W. L. Benoit & Airne, 2005; W. L. Benoit & Benoit-Bryan, 2013; W. L. Benoit & Brazeal, 2002; W. L. Benoit & Harthcock, 1999; W. L. Benoit & Henson, 2007; W. L. Benoit & Klyukovski, 2006; W. L. Benoit et al., 2002; W. L. Benoit & Sheafer, 2006; W. L. Benoit, Wen & Yu, 2007; Choi & Benoit, 2013; C. Lee & Benoit, 2005.

sondern auch Reden (z.B. W. L. Benoit, Blaney & Pier, 2000), Wahlwerbespots (z.B. W. L. Benoit, 1999) oder gleich die gesamte Kampagnenkommunikation (z.B. W. L. Benoit, Blaney & Pier, 1998; W. L. Benoit, McHale, Hansen, Pier & McGuire, 2003) untersuchen lassen, nach Angriffen (attacks), Selbstpräsentationen (acclaims) und Verteidigungen (defences). Daran angelehnt unterscheidet Maurer (2007) zwischen einer Angriffsstrategie und zusätzlich nach der Zeitdimension der Selbstdarstellungen nach Leistungsbilanzstrategie und Handlungsankündigungen. Da es in allen Einteilungen im Wesentlichen um die Bezugsrichtung der Kandidatenaussagen geht, bezeichnen wir sie in Anlehnung an Martel (1983, S. 62) als *Relationen*, die relative Häufigkeit der unterschiedlichen Relationen in der Kommunikation eines Kandidaten als *relationale Strategie*. In unserer eigenen Studie orientieren wir uns an der Operationalisierung von Benoit und unterscheiden zwischen Selbstpräsentation, Angriff und Verteidigung (vgl. Kapitel 4 sowie ausführlicher Bachl, Käfferlein & Spieker, 2013a, 2013b). Auch die im Folgenden präsentierten Analysen beziehen sich entweder direkt auf diese Kategorisierung oder lassen sich zumindest einfach einordnen.

Ziel der Analyse von J. Maier (2009) ist es, die Wahrnehmung und Wirkung des wirtschaftspolitischen Teils des TV-Duells zwischen Schröder und Merkel genauer zu untersuchen. Die Datengrundlage seiner Analyse entspricht der induktiven Analyse von Reinemann und Maurer (2007b). Zusätzlich werden die Daten einer Inhaltsanalyse herangezogen, die bei Strömbäck, Maier und Maier (2009, siehe unten) näher beschrieben wird. Für die Inhaltsanalyse wird die Debatte in Aussagen aufgeteilt. Eine Aussage ist definiert als eine Einheit von Sprecher, Thema, Bezugsobjekt und Richtung der Bewertung. Als inhaltsanalytisch erfasste Merkmale, deren Wirkung auf die unmittelbare Bewertung der Aussagen untersucht wird, werden unter anderem herangezogen: Thema (Wirtschaft allgemein, Arbeitsmarkt, Steuern, Haushaltsdefizit), Zeitperspektive (Gegenwart, Vergangenheit, Zukunft), Bezugsobjekt (eigenes Lager, gegnerisches Lager), Strategie (Angriff, Verteidigung), Tenor (positiv, neutral, negativ). Analysiert wird, ob sich die mittlere Bewertung der Aussagen durch das gesamte Publikum und durch Teilgruppen nach Kandidatenlager (Anhänger Schröder, Anhänger Merkel, Ungebundene, PDS (nur in Jena)) in Abhängigkeit von den inhaltlichen Merkmalen der Aussagen unterscheiden. Alle Analysen werden getrennt für beide Standorte durchgeführt. Aus den Ergebnissen früherer Studien zur Kandidatenbewertung in Fernsehdebatten wird die Erwartung abgeleitet, dass Angriffe auf Reaktanz stoßen und daher negativ bewertet werden. Für die übrigen Merkmale des Debatteninhalts werden keine expliziten Annahmen getroffen.

Die Ergebnisinterpretation gestaltet sich vor allem explorativ. Insgesamt wurde Merkel vom gesamten Mainzer Publikum im wirtschaftspolitischen Debattenteil besser bewertet, in Jena kam dagegen Schröder besser an. Die bivariaten Analysen offenbaren, dass die Kandidaten jeweils bei Aussagen mit einigen bestimmten Charakteristika besser bewertet wurden: Merkel in Mainz mit Aussagen zu den Themen Arbeitsmarkt, Steuern und Haushaltsdefizit und mit Aussagen mit den Zeitbezügen Gegenwart und Zukunft sowie in Jena mit Aussagen zum Thema Haushaltsdefizit; Schröder in Jena mit Aussagen zum Thema allgemeine Wirtschaftspolitik und Steuerpolitik und mit Aussagen mit den Zeitbezügen Vergangenheit und Gegenwart. Nicht bestätigt werden in den bivariaten Analysen die Annahmen zur Wirkung von Angriffen: Es kann kein signifikanter Unterschied zwischen der Bewertung von Aussagen, die Angriffe enthielten, und den Bewertungen der übrigen Aussagen festgestellt werden. Es zeigt sich jedoch ein negativer Einfluss von Verteidigungen auf die Beurteilung von Merkels Aussagen.

Die multivariaten Analysen, die separat nach Standort und Lagerzugehörigkeit durchgeführt werden, ergeben nach der Interpretation des Autors für die Bewertungen der Aussagen Schröders keine klaren Muster. Die Aussagen Merkels werden dagegen von (fast) allen Teilgruppen negativ bewertet, wenn sie über die Haushaltspolitik sprach, und wenn sie sich gegen Angriffe Schröders verteidigte. Besonders das zweite Ergebnis wird als bemerkenswert herausgestellt, da es für einen Erfolg der Angriffsstrategie Schröders (vgl. hierzu auch Maurer, 2007) spricht: Die Angriffe führten nicht zu negativen Bewertungen seiner Aussagen, Merkel wurde dann aber schlechter bewertet, wenn sie sich gegen diese Angriffe verteidigte. Zusätzlich zu den Interpretationen des Autors deuten sich bei näherer Betrachtung der multivariaten Auswertungen für das Mainzer Publikum interessante Interaktionen zwischen Lagerzugehörigkeit und Relation der Aussagen an: Die Angriffe Schröders führen nicht zu einer Abwertung im eigenen Lager, werden von den Anhängern Merkels jedoch negativer bewertet. Merkels Verteidigungen kommen bei Schröders Anhängern und bei den Ungebundenen schlecht an, jedoch nicht in ihrem eigenen Lager. Schließlich zeigt die Betrachtung der Reaktionen der Kandidatenlager wiederholt den Einfluss der Voreinstellungen auf die unmittelbaren Urteile über die im Duell vorgebrachten Aussagen, unabhängig von deren Merkmalen.

Insgesamt ergeben sich einige interessante Variationen der unmittelbaren Kandidatenbewertungen in Abhängigkeit von den Aussagencharakteristika. Neben der Falsifikation der Erwartung, dass Angriffe einen negativen Effekt auf die Zustimmung des Publikums haben, bleiben die meisten Interpretationen allerdings im Kontext des vorliegenden Duells verhaftet. Neben der grundsätzlichen Frage nach der Eignung der gewählten Aggregations- und

Analyselogik (vgl. Kapitel 5.3.1) müssen zudem die für die multivariate Analyse gewählten Regressionsmodelle (J. Maier, 2009, S. 189-190) kritisch betrachtet werden. Die gemeinsame Aufnahme der Merkmale „Bezugsobjekt: Eigenes Lager", „Bezugsobjekt: Gegner", „Strategie: Angriff", „Strategie: Verteidigung" und „Tenor (dreistufig mit negativ (-1), neutral (0), positiv (1))" lässt auf eine hohe Multikollinearität schließen, die den Einfluss der einzelnen Merkmale verringert. Für diesen Verdacht sprechen auch die relativ hohen Varianzaufklärungen bei gleichzeitig keinen oder wenigen signifikanten Prädiktoren. In Hinblick auf die Wirkung der Relationen drängt sich beispielsweise die Frage auf, inwiefern sich eine Aussage mit dem Merkmal „Angriff" von einer Aussage mit „negativem Tenor" und „Bezugsobjekt: Gegner" unterscheidet, oder ob hier nicht zweimal dieselbe Information in das Modell aufgenommen wird.

Strömbäck et al. (2009) legen auf Basis der Rezeptionsstudie, die bereits für M. Maier und Strömbäck (2009) beschrieben wurde, und der Inhaltsanalyse, die für J. Maier (2009) beschrieben wurde, eine vergleichende Analyse des Einsatzes und der Wirkungen von negativen Aussagen in zwei TV-Debatten in Deutschland und Schweden vor. Bezüglich des Wirkungsteils, der für uns von größerer Bedeutung ist, formulieren Strömbäck et al. (2009, S. 11) die Forschungsfrage: „What are the effects of negativity as a rhetorical strategy with respect to the evaluation of the statements of the candidates?". Die Autoren stellen keine expliziten Hypothesen über die Wirkung von negativen Aussagen auf. Sie führen aber drei aus der Psychologie stammende Argumente bezüglich der Verarbeitung negativer Information an, die für ihre besondere Wirksamkeit sprechen (vgl. für eine ausführlichere Zusammenfassung solcher Befunde auch Skowronski & Carlston, 1989). Erstens werden negativen Informationen größere Handlungskonsequenzen zugeschrieben als positiven Informationen. Zweitens werden in Entscheidungssituationen Informationen über mögliche Verluste stärker gewichtet als Informationen über mögliche Gewinne (vgl. dazu auch Kahneman & Tversky, 1979). Drittens werden die Informationen, die direkt nach einem negativen Stimulus rezipiert werden, besser erinnert. Allerdings stellen die Autoren auch fest, dass die Befunde zur Wirkung von Negative Campaigning trotz dieser Überlegungen widersprüchlich sind und nicht generell für einen größeren Erfolg negativer Wahlkampfkommunikation sprechen. Auffällig ist, dass die Autoren weder definieren, was sie unter einer größeren „effectiveness" (S. 6) verstehen, noch eine Vermutung äußern, wie negative Aussagen in Abhängigkeit von den Voreinstellungen der Rezipienten wirken.

Abhängige Variable für alle Auswertungen ist die Veränderung der Bewertung in einer Aussage, dass heißt, die Differenz zwischen der Bewertung zu Beginn und am Ende der Aussage. Alle Auswertungen werden für beide un-

tersuchten Debatten in Deutschland und Schweden wie bei J. Maier (2009) getrennt für Teilgruppen des Publikums nach Kandidatenlager durchgeführt. Die Analyse geht in drei Schritten vor. Zuerst wird die Veränderung der Bewertung von positiven, negativen, vergleichenden und verteidigenden Aussagen verglichen. Für die Debatte in Deutschland finden sich außer dem Einfluss der Lagerzugehörigkeit keine relevanten Unterschiede. Die Autoren schließen gar: „On this level of analysis, the rhetorical strategy used simply does not appear to matter at all" (Strömbäck et al., 2009, S. 19). In der schwedischen Debatte finden sich leichte Unterschiede: Die negativen Aussagen von Amtsinhaber Persson wurden von den Unentschiedenen negativ bewertet, die positiven Aussagen dagegen positiv. Die negativen Aussagen von Herausforderer Reinfeldt führten unter den Unentschiedenen dagegen zu einem Anstieg seiner Bewertung – das Ergebnis ist allerdings nicht statistisch signifikant. Im zweiten Analyseschritt werden die Bewertungen der negativen Aussagen detaillierter in Hinblick auf weitere Merkmale (u.a. explizite vs. implizite Bewertung; Bezugsobjekt der Bewertung; Erwähnung des gegnerischen Kandidaten) verglichen. Wiederum finden sich für das TV-Duell in Deutschland kaum Unterschiede. Die Bewertung von Merkels Aussagen durch ihre eigenen Anhänger verbesserte sich stärker, wenn sie explizit negative Aussagen tätigte. In der Analyse des schwedischen Duells zeigen sich keine substantiellen Unterschiede zwischen den verglichenen Merkmalen. In einem letzten Schritt wird der Einfluss der Merkmale negativer Aussagen auf die Bewertungsveränderung multivariat geprüft. Dabei ergeben sich laut Strömbäck et al. (2009, S. 21) keine berichtenswerten Befunde. Insgesamt stellen die Autoren fest, dass negative Aussagen nicht wirksamer waren als positive Aussagen.

Eine dritte Analyse, die sich ausschließlich auf Effekte der Relationen konzentriert, befasst sich mit dem TV-Duell zwischen Bundeskanzlerin Angela Merkel und ihrem Herausforderer Frank-Walter Steinmeier vor der Bundestagswahl 2009. Spieker (2011) geht der Frage nach, ob die Selbstpräsentationen, Angriffe und Verteidigungen der Kandidaten in dieser Debatte unterschiedlich bewertet wurden.[20] Die Arbeit testet drei Hypothesen(gruppen). Erstens wird erwartet, dass die Zuschauer, die sich dem Lager eines Kandidaten zuordnen lassen, den eigenen Kandidaten positiver bewerten als den Gegenkandidaten. Zweitens wird aus der Literatur zur Wirkung von Negative Campaigning abgeleitet, dass Angriffe von den Anhängern des Angreifers positiv, von den Anhängern des Angegriffenen und den Unentschiedenen aber negativ bewertet werden. Es

20 Die Publikation basiert auf Arne Spiekers Diplomarbeit, die unter dem Titel „Rhetorische Strategien und deren Wirkungen im TV-Duell 2009" an der Universität Hohenheim eingereicht wurde und an deren Betreuung ich beteiligt war. Weiterführende Analysen haben wir gemeinsam auf zwei Fachtagungen präsentiert (Bachl & Spieker, 2010; Spieker & Bachl, 2010).

wird also eine Interaktion zwischen Voreinstellung und Angriffen erwartet. Drittens wird für Verteidigungen ein negativer Haupteffekt angenommen, sie werden von allen Zuschauern negativ bewertet. Die Effekte der Relationen Angriff und Verteidigung werden über einen Vergleich mit der Referenzkategorie Selbstdarstellung gemessen, da diese nach den Ergebnissen Benoits (vgl. den Exkurs zu Relationen) als Standardmodus der Kampagnenkommunikation gelten. Auch in der vorliegenden Debatte waren Selbstdarstellungen mit Abstand die häufigste Relation.

Die relationalen Strategien wurden mit einer standardisierten Inhaltsanalyse der Debatte auf Aussagenebene erhoben. Eine Aussage wird durch Sprecher, Thema, Bezugsobjekt und Relation bestimmt. Wechselt eines dieser Elemente, wird eine neue Aussage kodiert. Für den Test der Hypothesen werden die Bewertungen aller Aussagen mit einer der drei Strategien für jeden Probanden zu einem Mittelwert zusammengefasst. Die Auswertung erfolgt mit Varianzanalysen mit dem Zwischen-Subjekt-Faktor Lagerzugehörigkeit (Anhänger Merkel: Parteiidentifikation für CDU/CSU oder FDP, n = 75, Steinmeier: SPD oder Grüne, n = 75, Keine Parteiidentifikation: n = 30) und dem Inner-Subjekt-Faktor Relation der Aussage. Die erste Hypothese zum Einfluss der Lagerzugehörigkeit wird – wie es auch die übrigen hier vorgestellten Studien erwarten lassen – gestützt. Die Hypothesen zu Bewertungen von Angriffen werde nur für Steinmeier in Teilen gestützt: Seine Angriffe werden von den Anhängern Merkels und den Ungebundenen negativer bewertet als seine Selbstpräsentationen. Hypothesenkonform werden zudem die Verteidigungen beider Kandidaten durch alle Gruppen schlechter bewertet als die Selbstpräsentationen. Insgesamt fallen aber auch die statistisch signifikanten Unterschiede in ihrem Umfang nur gering aus. Spieker (2011) sieht den Grund hierfür vor allem im wenig konfrontativen Debattenstil der beiden Kandidaten, der kaum dazu geeignet war, starke Reaktionen bei den Rezipienten hervorzurufen.

Die bisher umfangreichste Analyse, in der die Wirkung bestimmter Debatteninhalte auf die unmittelbare Bewertung der Kandidaten getestet wird, legt Nagel (2012) in ihrer Dissertation vor. Eine weitere Publikation eines Teils der Befunde, in der zudem explizite Hypothesen zur Wirkung einiger Debattenmerkmale getestet werden, findet sich bei Nagel, Maurer und Reinemann (2012). Ziel der Arbeiten ist es, den relativen Einfluss verbaler, vokaler und visueller Merkmale des Debatteninhalts auf die Wahrnehmung der Kandidaten in einer TV-Debatte festzustellen und damit „den Mythos der starken Wirkung des Nonverbalen in politischen Diskussionssendungen auf den empirischen Prüfstand zu stellen“ (Nagel, 2012, S. 18). Die theoretischen Vorüberlegungen beruhen, neben einer ausführlichen Sichtung empirischer Literatur zum Einfluss der Kommunikationsmodalitäten, auf psychologischen Modellen

zur Informationsverarbeitung. Die Dual-Coding-Theorie (Paivio, 2007), das Elaboration-Likelihood-Modell (Petty & Cacioppo, 1986) und das Modell der heuristischen und systematischen Informationsverarbeitung (Chaiken, 1980) werden herangezogen, um zu erklären, warum verschiedene verbale und nonverbale Charakteristika der Kandidatenaussagen ihre Bewertung durch die Rezipienten in unterschiedlichem Ausmaß beeinflussen können. Nach den theoretischen Vorarbeiten wird von einem additiven Modell ausgegangen, das heißt, die Merkmale der Debatteninhalte beeinflussen gemeinsam die unmittelbare Bewertung durch die Rezipienten. Zudem wird vermutet, dass Interaktionen zwischen den Merkmalen vorkommen können. Spezifischer sind die Erwartungen der Autorin im Hinblick auf zwei für die Forschungsfrage zentrale Punkte: Sie geht davon aus, dass die nonverbalen Elemente keine größere Bedeutung haben als die verbalen Elemente, und dass die visuellen Elemente für die Erklärung der Bewertung durch weniger involvierte Zuschauer wichtiger sind als für die Bewertung durch die stärker Involvierten.

Untersuchungsgegenstand ist das TV-Duell zwischen Schröder und Merkel vor der Bundestagswahl 2005. Die Daten der unmittelbaren Kandidatenbewertung stammen aus der Mainzer Teilstichprobe der in Kapitel 3.4.1 beschrieben Studie von Reinemann und Maurer (2007b). Zusätzlich werden die Daten einer sehr detaillierten Inhaltsanalyse auf Sekundenbasis herangezogen. Erfasst werden unter anderem Thema, Bezugsobjekt, Tendenz, rhetorische Stilmittel, sichtbares Verhalten (Blick, Lächeln, Gestik) von Sprecher und Zuhörer sowie Merkmale der Parasprache (Stimmhöhe, Lautstärke, Sprechgeschwindigkeit). Mit einem eigens entwickelten Analyseverfahren (ausführlich diskutiert in Abschnitt 5.3.1) werden diese Merkmale des Debatteninhalts mit der unmittelbaren Bewertung der Kandidaten durch das gesamte Publikum, durch Teilgruppen getrennt nach Kandidatenlager (Anhänger Schröder: Parteiidentifikation für SPD oder Grüne, n = 26; Anhänger Merkel: CDU/CSU oder FDP, n = 27; Ungebundene: n = 17) und durch Teilgruppen getrennt nach Involvement (hoch: n = 18, mittel: n = 31, niedrig: n = 23) verknüpft. In Anbetracht der vielfältigen Ergebnisse der Dissertation müssen wir uns an dieser Stelle auf ausgewählte Befunde beschränken. Wir konzentrieren uns auf die zentralen Ergebnisse für Nagels (2012) Forschungsfrage und die für die vorliegende Arbeit interessanten Ergebnisse zu den Effekten der Relationen.

Hinsichtlich des relativen Einflusses verbaler, visueller und vokaler Merkmale des Debatteninhalts zeigt sich die größte Bedeutung für die verbalen Merkmale. Innerhalb dieses Blocks waren die politischen Themen am wichtigsten, es folgte die Argumentationsrichtung (u.a. Selbstdarstellung und Angriff) vor den rhetorischen Stilmitteln. Die populäre These einer „Übermacht des Visuellen und des Nonverbalen“ (Nagel, 2012, S. 43) kann, wie von der Autorin erwartet, verwor-

fen werden. Das sichtbare Verhalten der Kandidaten leistet aber stellenweise durchaus einen Beitrag zur Erklärung der Publikumsreaktionen, insbesondere, wenn spezifische verbale und visuelle Elemente in Interaktion miteinander wirkten. Die Parasprache hilft nur bei der Erklärung der Publikumsurteile über Merkel. Aussagen, die sie in einer höheren Stimmlage vortrug, wurden positiver bewertet. Der Vergleich der Effekte für die Gruppen der gering, mittel und hoch involvierten Zuschauer stimmt mit der Logik des Elaboration-Likelihood-Modells überein. Die mittleren Bewertungen der gering Involvierten können zu einem größeren Anteil durch die nonverbalen Merkmale des Debatteninhalts erklärt werden.[21]

Im Hinblick auf die Relationen finden sich positive Effekte für die Bewertungen durch das gesamte Publikum, wenn Schröder Selbstpräsentationen und wenn Merkel Angriffe nutzte. Merkel wurde dagegen schlechter bewertet, wenn sie ihr eigenes Lager positiv darstellte. Die Differenzierung nach den Bewertungen durch die Kandidatenlager zeigt die positiven Effekte von Schröders Selbstpräsentationen auch für die eigenen Anhänger und die Ungebundenen. Die positiven Effekte von Merkels Angriffen zeigen sich ebenfalls in der Bewertung durch die Ungebundenen. Ihre positive Darstellung des eignen Lagers bzw. der Bilanz des eigenen Lagers wurde dagegen von den Anhängern Schröders und den Ungebundenen unterdurchschnittlich bewertet. Nagel (2012, S. 258) ordnet diese Effektmuster als „klassische rollenabhängige Wirkungen“ ein. Ein Amtsinhaber werde positiv bewertet, wenn er die Erfolge seiner Regierung beschreibt, während eine Herausforderin gut ankäme, wenn sie die Regierung kritisiert.

In Nagel et al. (2012) wird mit dem Einfluss der Debatteninhalte auf die unmittelbare Bewertung durch alle Zuschauer ein Ausschnitt der empirischen Befunde präsentiert. Wesentlicher Unterschied ist, dass zu den erwarteten Einflüssen bestimmter Merkmale explizite Hypothesen formuliert werden. Auf Grundlage des Issue-Ownership-Ansatzes wird erwartet, dass die Kandidaten bei den Themen ihrer Partei besser bewertet werden. Die Hypothese wird im Großen und Ganzen gestützt, allerdings gibt es bei einigen Themen Abweichungen. Es wird erwartet, dass sowohl Selbstpräsentationen als auch Angriffe positiver bewertet werden. Erstes wird für Schröder, zweites für Merkel gestützt. Positive Effekte werden zudem für verschiedene im weitesten Sinne rhetorische Mittel (Evidenzen, emotionale Appelle, Gemeinplätze, rhetorische Figuren) erwartet. Hier fallen die Ergebnisse der Hypothesentests gemischt

21 Inwiefern ein Modell der individuellen Informationsverarbeitung jedoch anhand von Gruppendaten adäquat getestet werden kann, ist zweifelhaft (vgl. Kapitel 5.3.1). Das heißt nicht, dass dieses Ergebnis falsch sein muss – die individuellen Urteile über die Kandidatenaussagen wurden hier einfach nicht untersucht.

aus, es zeigen sich kandidatenspezifische Unterschiede. Dies gilt ebenso für die Hypothesen zu den Einflüssen visueller (Blick in die Kamera, Lächeln, Gesten) und vokaler (niedrigere Tonhöhe, größere Lautstärke, schnelleres Sprechtempo) Merkmale.

Insgesamt präsentieren die Autoren vielfältige Ergebnisse zur Wirkung des Debatteninhalts auf die unmittelbare Kandidatenbewertung, die sich nur schwer einheitlich zusammenfassen lassen. Zentrales Ergebnis aus Perspektive der vorliegenden Arbeit ist, auch wenn es kaum überraschend erscheint: Viele unterschiedliche Merkmale des Debatteninhalts haben das Potenzial, bei verschiedenen Teilgruppen des Publikums unterschiedliche Reaktionen hervorzurufen. Betrachtet man vor allem die Ergebnisse zur Bewertung der Kandidaten durch die Lager im Detail, so wird klar, dass sowohl Charakteristika des Inhalts als auch die Voreinstellungen der Rezipienten die unmittelbaren Kandidatenbewertungen beeinflussen. Diese Interaktion zwischen Inhalts- und Rezipientencharakteristika nimmt in den Publikationen jedoch eine zweitrangige Rolle ein. Der Fokus liegt staxttdessen stärker auf der Interaktion zwischen verschiedenen inhaltlichen Charakteristika sowie Eigenschaften der Kommunikatoren. Hierzu werden einige interessante Befunde vorgestellt, deren Reichweite allerdings noch vor dem Hintergrund der eingesetzten Analyselogik eingeordnet werden müssen (vgl. Kapitel 5.3.1).

Zusammenfassung

Zumindest implizit[22] ist es das Ziel der vorgestellten deduktiv angelegten Studien, allgemeine Annahmen über die Effekte bestimmter Charakteristika der Debatteninhalte auf die unmittelbare Bewertung der Kandidaten durch die Rezipienten zu überprüfen. Diese Annahmen werden zu einem großen Teil aus einem von experimentellen Studien dominierten Forschungsstand abgeleitet. Die Besonderheit der TV-Duell-Studien besteht folglich darin, diese Annahmen in einem Design zu testen, das eine größere externe Validität verspricht (vgl. dazu ausführlich Kapitel 2 und Nagel, 2012, S. 82-95).

Was das Erkenntnisinteresse der gesichteten Studien angeht, ist ein Fokus auf die Effekte bestimmter Charakteristika des Debatteninhalts auszumachen. Interaktionen zwischen den Inhalten und den Voreinstellungen der Rezipienten werden selten thematisiert und in den meisten Analysen nicht statistisch getes-

22 Nicht alle vorgestellten Studien formulieren explizite Hypothesen oder Erwartungen zu den Effekten der untersuchten Merkmale. Wir können allerdings davon ausgehen, dass die Autoren für jedes untersuchte Debattenmerkmal zumindest eine ungerichtete Unterschiedshypothese (Aussagen mit diesem Merkmal werden anders bewertet als Aussagen ohne dieses Merkmal) annehmen.

tet (Ausnahme: Spieker, 2011). Die Voreinstellungen der Rezipienten – deren wichtige Rolle auch in den vorgestellten Studien größtenteils thematisiert wird (Ausnahme: Nagel et al., 2012) – findet sich in den Analysen meist nur in der Form vergleichender Betrachtungen der Befunde für verschiedene Teilgruppen des Publikums wieder. Damit gehen die deduktiv angelegten Studien in dieser Hinsicht weniger weit als viele induktive Analysen, die immerhin die Differenz zwischen den Bewertungen durch die Anhänger verschiedener Lager als empirischen Indikator für die Polarisierung des Publikums und damit für eine Interaktion zwischen Inhalten und Voreinstellungen heranziehen. Die mangelnde Berücksichtigung dieser Interaktion in den vorliegenden deduktiven Studien ist, wie wir in Kapitel 5.3 zeigen werden, auch der eingesetzten Aggregations- und Analyselogik geschuldet.

Die Effekte von zwei Merkmalen des Debatteninhalts wurden in mehreren Analysen untersucht. Zum einen ist ein Forschungsinteresse an der Frage erkennbar, ob die Aussagen der Kandidaten in Abhängigkeit der Relation (vor allem Selbstpräsentation und Angriff, teils auch Verteidigung) unterschiedlich bewertet werden. Zum anderen haben zwei Arbeiten die Effekte des politischen Themas untersucht. In einem Vergleich sehr vieler Merkmale zeigt Nagel (2012), dass die politischen Themen den größten Einfluss auf die Bewertung der Kandidaten hatte. Die Zusammenfassung des Forschungsstands zu den Effekten von Relationen und Themen fällt trotz der geringen Zahl der vorliegenden Studien schwer. Die Effekte der Relationen können teils überhaupt nicht nachgewiesen werden, teils fallen sie sehr gering aus. Die Effekte des Themas sind sowohl bei J. Maier (2009) als auch bei Nagel (2012) statistisch signifikant und im Vergleich zu den übrigen berücksichtigten Merkmalen am stärksten. Auch in der vorliegenden Arbeit werden wir unsere Vorschläge zur Datenanalyse anhand dieser beiden Merkmale vorstellen und damit versuchen, etwas zu diesem noch lückenhaften Forschungsstand beizutragen.

4 Die TV-Duell-Studie Baden-Württemberg 2011

Das folgende Kapitel gibt einen Überblick über die Rezeptionsstudie zum TV-Duell zwischen Stefan Mappus und Nils Schmid vor der Landtagswahl 2011 in Baden-Württemberg.[23] Zuerst gehen wir knapp auf die Datenerhebung ein. Eine ausführliche Dokumentation hierzu haben wir bereits im Sammelband zu dieser Studie veröffentlicht (Bachl, Brettschneider & Ottler, 2013b). Ausführlicher dargestellt wird anschließend die Qualität der rezeptionsbegleitend gemessenen Kandidatenbewertungen, deren Erklärung in der vorliegenden Arbeit im Mittelpunkt steht.

4.1 Datenerhebung

Die Studie befasst sich mit den Inhalten, Wahrnehmungen und Wirkungen des TV-Duells zwischen Ministerpräsident Stefan Mappus (CDU) und dem Spitzenkandidaten der größten Oppositionsfraktion Nils Schmid (SPD). Unseres Wissens legen wir damit die erste Rezeptions- und Wirkungsstudie zu einem TV-Duell auf Landesebene vor. Das TV-Duell fand am 16. März 2011, eineinhalb Wochen vor der baden-württembergischen Landtagswahl am 27. März, statt und wurde live von etwa einer halben Million Zuschauer im SWR-Fernsehen verfolgt. Das TV-Duell dauerte etwas über eine Stunde und behandelte die wichtigsten Themen des Wahlkampfs (vgl. zur Ausgangslage vor dem TV-Duell ausführlich Vögele, 2013). Das Design der Studie folgt den etablierten Vorgängerstudien zu den TV-Duellen auf Bundesebene (z.B. Faas & Maier, 2004a; Faas et al., 2009; Maurer & Reinemann, 2003; Reinemann & Maurer, 2007a): In einer kontrollierten Rezeptionsstudie wurden nach einer quotierten Stichprobenziehung ausgewählte Rezipienten vor und nach dem TV-Duell befragt. Während des Duells bewerteten sie die Kandidaten mit RTR-Dials. Zusätzlich wurden Merkmale des Debatteninhalts mit einer standardisierten Inhaltsanalyse erhoben. In den folgenden Abschnitten geben wir einen Überblick über die Bestandteile des Studiendesigns.

23 Wir danken der Fritz Thyssen-Stiftung für die Ermöglichung dieser Studie durch eine Projektförderung.

Stichprobenziehung und realisierte Stichprobe

Die Rezeptionsstudie wurde an der Universität Hohenheim und der Dualen Hochschule Baden-Württemberg (DHBW) Ravensburg[24] durchgeführt. Die Rekrutierung der Teilnehmer und die Durchführung der Rezeptionsstudie erfolgten nach einem für beide Standorte standardisierten Vorgehen. Durch die Verteilung der Datenerhebung auf zwei Standorte kann natürlich keine repräsentative Abdeckung der regionalen Unterschiede in ganz Baden-Württemberg erreicht werden. Die Aufnahme eines zweiten Standorts neben der Landeshauptstadt Stuttgart stellt jedoch zumindest sicher, dass die Publikumsreaktionen auf hier besonders präsente Themen – zu denken ist in diesem Zusammenhang vor allem an „Stuttgart 21" – mit einer zweiten Teilstichprobe verglichen werden können. Neben der Verteilung der Stichprobe auf zwei Standorte wurde innerhalb der Standorte eine Gleichverteilung der Merkmale politisches Lager (dreistufig: Parteiidentifikation für die Regierungsparteien CDU und FDP bzw. die Oppositionsparteien SPD und Bündnis 90 / Die Grünen sowie Personen ohne Parteiidentifikation), formale Bildung, politisches Interesse und Alter (jeweils zweistufig) nach einem 24-zelligen Quotenplan angestrebt. In den Randverteilungen des Quotenplans sollte eine Gleichverteilung nach Geschlecht erreicht werden. Zudem mussten alle Probanden bei der anstehenden Landtagswahl in Baden-Württemberg wahlberechtigt sein. Vorgegeben durch die zur Verfügung stehende RTR-Ausstattung war für Stuttgart eine Stichprobengröße von 120 und für Ravensburg eine Stichprobengröße von 80 zu erreichen.

Über einen bestehenden Pool bereits registrierter Studieninteressierter, Flyer, Zeitungsartikel und Anzeigen wurde im Umfeld der Studienstandorte für eine Teilnahme an der Rezeptionsstudie geworben. In den Werbematerialien wurde auf das Thema der Studie und eine finanzielle Aufwandsentschädigung hingewiesen. Studieninteressierte füllten bevorzugt eine kurze Online-Befragung mit den zur Quotierung notwendigen Merkmalen und Kontaktdaten aus. Alternativ bestand zu einigen Terminen die Möglichkeit einer telefonischen Registrierung. Nach Abschluss der Registrierung wurden an beiden Standorten Teilnehmer nach der Quotenvorgabe eingeladen. Aus den überbesetzten Zellen des Quotenplans wurden zufällig ausgewählte Personen berücksichtigt. Zellen, deren Zielvorgaben durch die Rekrutierung nicht erfüllt werden konnten, wurden durch Personen aus benachbarten Zellen ausgeglichen. Die Tabellen 1 und 2 in Bachl, Brettschneider und Ottler (2013b, S. 13-14) informieren ausführlich über

24 Die Rekrutierung und Studiendurchführung in Ravensburg erfolgten durch Simon Ottler von der DHBW Ravensburg. Wir bedanken uns für die gute Zusammenarbeit.

das Verhältnis von registrierten und eingeladenen Studieninteressierten sowie Teilnehmern der Studie im Quotenplan.

Insgesamt nahmen in Stuttgart 119 und in Ravensburg 81 Personen an der Rezeptionsstudie teil. Tabelle 4.1 gibt einen Überblick über die wichtigsten Charakteristika der Stichprobe. Das Ziel einer Gleichverteilung der quotierten Stichprobenmerkmale wird recht gut erreicht. Allerdings sind die formal höher Gebildeten und politisch stärker Interessierten in der Stichprobe etwas stärker vertreten. Als problematisch stellt sich die Verteilung der politischen Voreinstellungen der Probanden, hier bemessen an der Zugehörigkeit zu den Lagern der Kandidaten, heraus. Obwohl die Gruppen nach der längerfristig wirksamen Parteiidentifikation eine ähnliche Größe haben, sind Personen, die eine Wahl für die Oppositionsparteien SPD und Bündnis 90 / Die Grünen beabsichtigen und damit dem Kandidaten Schmid nahestehen, im Vergleich zu den noch Unentschiedenen und den Personen mit einer Wahlabsicht für die Regierungsparteien CDU und FDP überrepräsentiert. Grund hierfür ist eine starke Präferenz der Personen ohne längerfristige Parteiidentifikation für eine Wahl der Grünen.

Aus dieser Verteilung folgt, dass sich aus einer Zusammenfassung der RTR-Messungen für die Gruppe der längerfristig Ungebundenen in dieser Stichprobe kein geeigneter Indikator für die Bewertung der Kandidaten durch die parteipolitisch neutralen Zuschauer ableiten lässt. Wir reagieren hierauf, indem wir in den folgenden Analysen, in denen die Lagerzugehörigkeit der Zuschauer als erklärende Variable berücksichtigt wird, eine Gruppenbildung nach der direkt vor dem TV-Duell erfassten Wahlabsicht vornehmen. Diese Entscheidung lässt sich auch inhaltlich gut begründen: Nach dem sozialpsychologischen Ansatz zur Erklärung des Wahlverhaltens kann die Wahlabsicht neben der längerfristig stabilen Parteiidentifikation durch die Einstellungen zu den Kandidaten und ihren Parteien erklärt werden (Schoen & Weins, 2005). Wenn wir nun die vor der Debatte gemessene Wahlabsicht als Rezipientenmerkmal zur Erklärung der unmittelbaren Kandidatenbewertung während des Duells heranziehen, nutzen wir eine Variable, die weitere wichtige Informationen über die politischen Voreinstellungen enthält.

Insgesamt erlaubt die Stichprobe der Rezeptionsstudie natürlich keinen direkten Inferenzschluss auf die Grundgesamtheit des gesamten Debattenpublikums oder seine Teilgruppen, und schon gar nicht auf die Grundgesamtheit aller Wahlberechtigten. Diese Repräsentativität ist in einer solchen Rezeptionsstudie kaum zu leisten, da durch die Standortgebundenheit der Technik zur rezeptionsbegleitenden Messung eine vollständige regionale Abdeckung praktisch nicht erreicht werden kann. Durch die erforderliche Bereitschaft der Teilnehmer, die Debatte an einem Abend live während der TV-Ausstrahlung in

Tabelle 4.1: Stichprobe der Rezeptionsstudie

	Stuttgart (n = 119)	Ravensburg (n = 81)	Gesamt (N = 200)
Geschlecht			
weiblich	52	51	51
männlich	48	49	49
Alter [$M(SD)$]	38.8(15.1)	43.0(19.4)	40.5(17.1)
Formale Bildung			
Schüler	8	3	6
Hauptschule	8	9	8
Realschule	28	22	26
Fachabitur	9	12	11
Abitur	19	27	22
Hochschulabschluss	29	22	26
Politisches Interesse[A] [$M(SD)$]	3.3(0.8)	3.5(0.8)	3.4(0.8)
Parteiidentifikation			
CDU/CSU	24	22	24
FDP	9	7	9
Lager Mappus	33	29	33
SPD	21	24	22
Bündnis 90 / Die Grünen	15	20	17
Lager Schmid	36	44	39
Andere	1	3	1
Keine	27	21	25
Wahlabsicht vor dem Duell			
CDU	20	22	20
FDP	7	4	6
Lager Mappus	27	26	26
SPD	16	17	17
Bündnis 90 / Die Grünen	33	25	30
Lager Schmid	49	42	47
Andere	2	1	2
Unentschieden	20	22	21

Anmerkungen
[A] Skala von 1 (geringes Interesse) bis 5 (sehr großes Interesse).
Alle Angaben außer Alter und politischem Interesse sind Spaltenprozente; die Angaben zu Lager Mappus bzw. Lager Schmid ergeben sich als Summe der Anteile von CDU und FDP bzw. SPD und Grüne; zu 100 fehlend: Befragte ohne Angabe, Rundungsfehler.
Übernommen aus Bachl, Brettschneider und Ottler (2013b, S. 17).

einem Hörsaal der Hochschulen in Hohenheim bzw. Ravensburg zu verfolgen, ist zudem von einer starken Selbstselektivität der Stichprobe auszugehen. In Hinblick auf die relevanten politischen Einstellungen ist uns jedoch eine ausgeglichene Zusammensetzung gelungen. Daher können die Befunde zumindest als geeignete Indikatoren dafür aufgefasst werden, wie die Kandidaten während der Debatte auch von anderen Zuschauern außerhalb unserer Stichprobe bewertetet wurden.

Untersuchungsablauf

Die Rezeptionsstudie wurde am 16. März 2013 während der Ausstrahlung des TV-Duells im SWR Fernsehen durchgeführt. Nach ihrem Eintreffen an der Universität Hohenheim bzw. der DHBW Ravensburg erhielten die Studienteilnehmer die Fragebögen für die Pre- und Post-Duell-Befragung, einen RTR-Regler sowie eine finanzielle Aufwandsentschädigung. Fragebögen und RTR-Regler waren mit anonymisierten ID-Nummern versehen, um die individuelle Verknüpfung der Daten zu ermöglichen. Die Teilnehmer wurden schriftlich über den wissenschaftlichen Zweck der Untersuchung und den Datenschutz aufgeklärt.

Nachdem sich die Probanden in den Hörsälen eingefunden hatten, in denen die Rezeptionsstudie durchgeführt wurde, füllten sie den ersten Fragebogen aus. Danach erfolgten eine Erläuterung zum Verhalten während der Rezeptionsstudie und eine Einweisung zur Bewertung der Kandidaten während der Debatte mittels der RTR-Dials. Die Bedienung des RTR-Reglers wurde anhand eines unpolitischen Probestimulus eingeübt.[25] Der Probestimulus wurde bereits im Vorfeld einer Rezeptionsstudie zum TV-Duell 2009 getestet und erwies sich als gut geeignet. Bei seiner Auswahl wurde darauf geachtet, möglichst alle Einflüsse auf die Bewertung der Kandidaten während des TV-Duells auszuschließen (vgl. zur Bedeutung des Probestimulus in RTR-Studien Schneider et al., 2011). Im Anschluss verfolgten die Probanden das TV-Duell und bewerteten währenddessen die Kandidaten Mappus und Schmid. Der Probestimulus und die TV-Debatte wurden auf einer Leinwand (Ravensburg) bzw. mehreren Leinwänden (Hohenheim) präsentiert. Durch ausführliche Tests haben wir im Vorfeld sichergestellt, dass für alle Versuchsteilnehmer eine gute Sicht auf den Stimulus sowie eine gute Akustik gegeben waren. Sofort nach Ende des TV-Duells und vor dem Wechsel in das Rahmenprogramm der Sendung wurde die Vorführung des Stimulus gestoppt. Die Probanden füllten dann den zweiten

25 Als Probestimulus diente das Streitgespräch zwischen zwei Eheleuten in Loriots „Das Frühstücksei“ (vgl. z.B. hier: `http://www.youtube.com/watch?v=bBQTBDQcfik`).

Fragebogen aus. Zuvor wurden Sie gebeten, sich nicht mit anderen Personen zu unterhalten, um eine Beeinflussung durch andere Meinungen auszuschließen.

Untersuchungsinstrumente

Befragung In den beiden schriftlichen Befragungen vor und nach dem TV-Duell haben wir zahlreiche Merkmale der Rezipienten zu politischen Einstellungen und politischem Wissen, psychologischen Charakteristika und Soziodemographie erhoben. Die vollständigen Fragebögen sind im Online-Anhang zum Sammelband dokumentiert (Bachl, Brettschneider, Kercher, Spieker & Vögele, 2013a, 2013b). Aus der Befragung direkt vor der Duellrezeption entnehmen wir einige Fragen, um die Voreinstellungen der Rezipienten zu operationalisieren:

- Alle Probanden, die in der Frage zuvor angegeben hatten, an der Landtagswahl in Baden-Württemberg teilzunehmen, wurden nach ihrer *Wahlabsicht* gefragt:

  > Und welcher Partei würden Sie dann Ihre Stimme geben? (Falls Sie bereits Ihre Stimme per Briefwahl abgegeben haben: Welcher Partei haben Sie Ihre Stimme gegeben?)

  Als Antwortoptionen standen die Parteien CDU, SPD, FDP, Bündnis 90 / Die Grünen und Die Linke zur Verfügung. Zudem war es möglich, in offener Angabe eine andere Partei zu nennen. Als letzte Antwortoption stand „Ich habe mich noch nicht entschieden“ bereit. Für die folgenden Analysen wird der Faktor *Lagerzugehörigkeit* gebildet, indem die Ausprägungen CDU und FDP zu Lager Mappus und die Ausprägungen SPD und Grüne zu Lager Schmid zusammengefasst werden. Die dritte Stufe des Faktors Lagerzugehörigkeit bildet die Ausprägung „noch nicht entschieden“. Alle Probanden mit anderen oder ungültigen Angaben werden aus der Analyse ausgeschlossen. Die Verteilungen der Ausgangsvariable und des Faktors Lagerzugehörigkeit in der Stichprobe kann aus Tabelle 4.1 entnommen werden.

- Die allgemeinen Einstellungen gegenüber den Kandidaten Mappus und Schmid sowie ihren Parteien CDU und SPD haben wir in Anlehnung an das bewährte Skalometer-Format (ZA & ZUMA, 2012) erfasst:

  > *Parteien*: Was halten Sie ganz allgemein von der Arbeit der folgenden Parteien in Baden-Württemberg?
  > *Kandidaten*: Und was halten Sie ganz allgemein von Stefan Mappus und Nils Schmid?

Die Probanden gaben ihre Einstellungen auf einer elfstufigen Skala von -5 bis $+5$ an. Die Endpunkte waren mit „Halte überhaupt nichts von dieser Partei / diesem Politiker“ und „Halte sehr viel von dieser Partei / diesem Politiker“ benannt. Die Variablen verteilten sich in der Stichprobe wie folgt: Skalometer Schmid: $M = 0.6$ $(SD = 1.7)$; Skalometer Mappus: $M = -1.4$ $(SD = 2.8)$; Skalometer SPD: $M = 1.0$ $(SD = 2.2)$; Skalometer CDU: $M = 0.4$ $(SD = 3.0)$. Auch hier zeigt sich die Tendenz der Gesamtstichprobe zu Gunsten von Schmid und der SPD bzw. zu Ungunsten von Mappus und der CDU. Besonders die mittlere Einstellung gegenüber dem Ministerpräsidenten ist vor dem TV-Duell negativ ausgeprägt.

Aus der Befragung nach der Debattenrezeption ziehen wir eine Frage heran, um in Teilkapitel 4.2 die prognostische Validität der RTR-Messungen zu prüfen:

- Die *ex post* erfassten Bewertungen der Kandidaten während des Duells erheben wir mit der Frage:

  Und wie beurteilen Sie den Auftritt der beiden Kandidaten im Einzelnen?

  Zur Bewertung dient eine benannte fünfstufige Skala von 1 bis 5 mit den Ausprägungen sehr schlecht, eher schlecht, mittelmäßig, eher gut und sehr gut. Aus den Variablen zu den beiden Kandidaten bilden wir zusätzlich ein Differential, indem wir die Bewertung der Leistung von Schmid von der Bewertung der Leistung von Mappus abziehen. Dieses Differential entspricht der Logik der RTR-Skala (siehe unten) und reicht von -4 (größter Vorteil Schmid) bis 4 (größter Vorteil Mappus). Die Variablen verteilen sich in der Stichprobe wie folgt: Debattenleistung Schmid: $M = 3.6$ $(SD = 0.8)$; Debattenleistung Mappus: $M = 3.5$ $(SD = 0.9)$; Differential: $M = -0.1$ $(SD = 1.4)$.

RTR-Messung Die unmittelbaren Bewertungen der Kandidaten während des TV-Duells wurden mit RTR-Reglern erfasst. Dabei haben wir Dials der Marke „Perception Analyzer“ (Modelle IV und V) des Herstellers Dialsmith[26] verwendet. Diese Dials arbeiten im Latched Mode, dass heißt, in einem vorgegebenen Zeitintervall wird für jedes Dial der Wert erfasst, auf den es gerade eingestellt ist. Ein automatisches Zurücksetzen auf den Ausgangswert der Skala findet nicht statt (vgl. Kapitel 3.1). Abgesehen von den durch die zur Verfügung stehende Technik vorgegeben Rahmenbedingungen können bei der Ausgestaltung des RTR-Messinstruments einige Entscheidungen frei getroffen

26 http://dialsmith.com/

werden. Dabei handelt es sich um die Vorgabe der Bewertungsobjekte und der Bewertungsdimension sowie die Zahl der Skalenpunkte und deren Präsentation gegenüber den Probanden. Schließlich kann bestimmt werden, in welcher zeitlichen Frequenz die Ausprägungen auf der RTR-Skala ausgelesen werden.

Bezüglich der *Bewertungsobjekte* haben sich in TV-Duell-Studien, in denen die Kandidaten mit Dials bewertet werden, zwei Vorgaben etabliert (vgl. auch die Angaben in den Kapiteln 3.1 und 3.4). In der ersten Variante werden die Probanden angewiesen, die Debatteninhalte, die sie gerade wahrnehmen, auf einer eindimensionalen Skala von negativ bis positiv zu bewerten. Diese Vorgabe ist sehr einfach zu verstehen und hat darüber hinaus den Vorteil, dass sie unabhängig von der Zahl der an der TV-Debatte teilnehmenden Kandidaten eingesetzt werden kann.

In der zweiten Variante, die bisher in allen Studien zu deutschen Kanzlerduellen eingesetzt wurde, wird eine Skala vorgegeben, auf der beide Kandidaten gleichzeitig bewertet werden können. Hier werden die Bedeutungen der Skalenendpunkte doppelt belegt, für beide Kandidaten mit der jeweils gegenläufigen Bewertung. An dieser Tradition haben wir uns in der Studie zum TV-Duell in Baden-Württemberg orientiert. Konkret ist die Skala folgendermaßen definiert: Eine Position des Reglers links der Skalenmitte bedeutet eine positive Bewertung von Schmid *oder* eine negative Bewertung von Mappus. Eine Position des Reglers rechts der Skalenmitte bedeutet umgekehrt eine negative Bewertung für Schmid *oder* eine positive Bewertung für Mappus. Neben der Anschlussfähigkeit an den Forschungsstand hat die doppelte Belegung der Skalenendpunkte den Vorteil, dass die Probanden relative Urteile über die beiden Kandidaten leichter abbilden können. Dies ist besonders für die Analyse der Effekte von Relationen relevant. Wie im Exkurs zu Relationen in Kapitel 3.4 (S. 78ff.) erläutert, hat ein Kandidat in einer Debatte das Ziel, im Verhältnis zum Gegenkandidaten besser bewertet zu werden. Der Erfolg eines Angriffs von Schmid könnte sich nach dieser Logik zum einen darin zeigen, dass Zuschauer diesen Angriff *per se* gut finden und Schmid daher besser bewerten. Zum anderen könnten die Zuschauer der Argumentation des Angriffs folgen und daher Mappus schlechter bewerten. Auf der hier eingesetzten RTR-Skala entsprechen beide Bewertungen – eine Aufwertung von Schmid oder eine Abwertung von Mappus – derselben Richtung. Damit sollte den Probanden eine relative Bewertung der Kandidaten leichter fallen. Wenn wir also die Prämisse akzeptieren, dass in der Situation des TV-Duells der Vorteil des einen Kandidaten gleichzeitig der Nachteil des anderen Kandidaten ist, so können wir die hier eingesetzte RTR-Skala als bipolare Differenzialskala begreifen, die von „größter Vorteil Schmid“ über den neutralen Skalenmittelpunkt bis hin zu „größter Vorteil Mappus“ reicht.

Bei der Anweisung, hinsichtlich welcher *Dimension* die Probanden die Kandidaten bewerten sollten, haben wir wiederum in Anlehnung an die Studien zu den Kanzlerduellen eine sehr offene Formulierung gewählt. Die Probanden sollten mit ihren RTR-Reglern angeben, ob sie gerade ganz subjektiv einen guten oder schlechten Eindruck von den Kandidaten haben. Auf was genau dieser Eindruck zurückgeht, haben wir ganz bewusst offen gelassen (Bachl, Brettschneider & Ottler, 2013b, S. 20; hier ist auch der gesamte Wortlaut der Erläuterung der Skala dokumentiert):

> Was genau ein guter oder ein schlechter Eindruck ist, wollen wir Ihnen dabei nicht vorschreiben. Sie können z.B. die Art bewerten, wie die Kandidaten auftreten, das, was sie sagen, oder das, was über sie gesagt wird. Kurz gesagt: Sie befinden darüber, wann Sie einen guten oder schlechten Eindruck von den Kandidaten haben, und nur Sie wissen, warum das so ist.

Durch diese offene Formulierung wollten wir die kognitive Belastung der Probanden reduzieren und einen zumindest etwas natürlicheren Rezeptionsmodus erreichen, da die Probanden sich nicht auf bestimmte Eigenschaften konzentrieren müssen. Zudem haben wir durch den Einschub „oder das, was über sie gesagt wird“ versucht, den Probanden den relativen Charakter der Messung ebenso bewusst zu machen. Schließlich ist die offene Vorgabe eines allgemeinen „Eindrucks“ als Bewertungsdimension auch sinnvoll, da wir nicht davon ausgehen, dass die Probanden über den gesamten Debattenverlauf eine spezifischere Bewertungsdimension konsistent anwenden.[27]

Bezüglich der *Zahl der Skalenpunkte* weichen wir von den bislang verbreiteten sieben- bis elfstufigen Ordinalskalen mit auf dem Regler verzeichneten Skalenpunkten ab. Stattdessen setzen wir eine aus Sicht der Probanden stufenlose, endpunktbenannte Skala ein. Abbildung 4.1 zeigt, wie wir diese Skala auf dem RTR-Regler visualisiert haben. Für die Probanden nicht sichtbar wurden die Ausprägungen auf einer 101-stufigen Skala von -50 bis 50 aufgezeichnet. Konzeptionell ähnelt unsere Skala damit einer visuell-analogen Skala mit einem zusätzlich eingezeichneten Skalenmittelpunkt, auf der die Probanden ihre Bewertung relativ zu den visuell und haptisch wahrnehmbaren Endpunkten und dem visuell wahrnehmbaren Skalenmittelpunkt verorten. Im Kontext von Online-Befragungen zeigen Methodenexperimente, dass visuell-analoge Skalen im Vergleich mit herkömmlichen ordinal gestuften Likert-Skalen und semantischen Differentialen häufig eine bessere Datenqualität erreichen, Messfehler

27 Vgl. hierzu auch die Befunde von Wolf (2010) zur asymptotischen Annäherung der RTR-Bewertungen unterschiedlicher Dimensionen des Kandidatenimages im Verlauf des Stimulus.

Anmerkung
Übernommen aus Bachl, Brettschneider und Ottler (2013b, S. 21).

Abbildung 4.1: RTR-Skala, RTR-Regler, Hörsaal in Hohenheim

verringern und in Mittelwerten und Streuungen äquivalente Ergebnisse liefern (z.B. Funke & Reips, 2012; Funke, 2010; Reips & Funke, 2008).

Aus unserer Sicht verspricht die Erweiterung der Spannweite der RTR-Skala gegenüber der in den Studien zu den Kanzlerduellen eingesetzten siebenstufigen, bipolaren Skala vor allem eine Vergrößerung der Optionen bei der Bewertung der eigenen Kandidaten. Die visuelle Inspektion der individuellen RTR-Verläufe aus der Studie zum Kanzlerduell 2009 (Faas et al., 2009) hat ergeben, dass die Anhänger eines Kandidaten häufig schon die erste positiv besetzte Ausprägung auswählen, sobald ihr Kandidat zu sprechen beginnt. Damit steht jedoch nur noch ein weiterer Skalenpunkt zur Verfügung, mit dem die Probanden eine ansteigende Zustimmung zu den folgenden Aussagen des Kandidaten ausdrücken können. Bedenken wir zusätzlich, dass eine negative Bewertung des eigenen Kandidaten nur sehr selten vorkommt, so spielt sich die Bewertung des eignen Kandidaten größtenteils auf einer gerade einmal dreistufigen Teilskala ab.[28] Zusätzlich wollen wir durch den Verzicht auf visualisierte Skalenpunkte erreichen, dass die Probanden ihre Aufmerksamkeit mehr dem Stimulus und weniger dem RTR-Regler zuwenden. Das von uns genutzte Dial-Modell besitzt keine haptisch zu erfassenden Skalenpunkte. Daher sind die Probanden gezwungen, häufig auf ihren Regler zu blicken, um

28 Aus diesem Grund verwirft Hughes (1992, S. 64) den Einsatz von fünfstufigen RTR-Skalen für die Bewertung von Werbespots in der angewandten Werbeforschung: „[T]here were problems with linking the response to the stimulus because subjects waited for a big change before using one of their few remaining scale positions. There were extreme floor and ceiling scale effects that either buried or distorted changes in evaluations“.

zu prüfen, ob sie tatsächlich den gewünschten Skalenpunkt „getroffen" haben. Diese Ablenkung lässt sich auch durch die stufenlose Skalierung nicht völlig vermeiden, sie sollte jedoch weniger stark auftreten.[29]

Als letzte Entscheidung ist die *zeitliche Frequenz* zu wählen, in der die Ausprägungen der RTR-Regler auf der Skala ausgelesen werden. Hier wählen wir die technisch kürzestmögliche Frequenz von einer Messung pro Sekunde. Wichtiger als diese Auswahl ist jedoch die Kommunikation der sekündlichen Datenerfassung an die Probanden. Während der Einweisung und des Probestimulus haben wir mehrmals darauf hingewiesen, dass die Bewertungen in jeder Sekunde aufgezeichnet werden und die Probanden daher jederzeit ihren RTR-Regler verwenden sollen, wenn sich ihre Bewertung verändert.

Aus technischer Perspektive verlief die rezeptionsbegleitende Messung der Kandidatenbewertung zufriedenstellend. Von den 199 Probanden, die mit RTR-Reglern ausgestattet waren, lieferten 191 gültige RTR-Verläufe. Acht Probanden wurden von der Analyse ausgeschlossen, da es zu gravierenderen technischen Ausfällen kam (mehr als 10 Prozent fehlende Messungen) oder die individuellen Verläufe klar darauf hinwiesen, dass die Probanden sich nicht an die Anweisungen zur RTR-Messung gehalten hatten. Da die RTR-Messungen als abhängige Variablen in dieser Arbeit im Zentrum des Interesses stehen, beschäftigen wir uns in Teilkapitel 4.2 noch eingehender mit der Reliabilität und Validität dieses Messinstruments.

Inhaltsanalyse Die Merkmale des Debatteninhalts haben wir in einer standardisierten Inhaltsanalyse auf Sekundenbasis erfasst. Zur technischen Umsetzung dieser speziellen Form der Inhaltsanalyse haben wir die Software ANVIL (Kipp, 2011, 2014) eingesetzt. Die Software erlaubt es, die Merkmale von audiovisuellem Material sekundengenau in voneinander unabhängigen sogenannten „tracks" zu annotieren. Ziel war es, eine Codierung nach der Rezeptionslogik der Zuschauer vorzunehmen. Ein Merkmal ist ab der Sekunde zu erfassen, ab der es für die Zuschauer erkennbar war. Als Kontext dürfen nur die vorangegangenen Inhalte herangezogen, das Wissen der Codierer um die folgenden

29 Leider haben es unsere Ressourcen bisher nicht erlaubt, die Konsequenzen unterschiedlicher Gestaltungen der RTR-Skala über einen in einer Lehrveranstaltung durchgeführten Versuch hinaus empirisch zu prüfen. Die Befunde zur Reliabilität und Validität der RTR-Messungen dieser Studie (vgl. Kapitel 4.2) liefern mit dem Forschungsstand konforme Befunde. Daher dürfte die Datenqualität durch diesen Wechsel auf eine neue Skalierung vermutlich nicht beeinträchtigt worden sein. Mehr methodologische Forschung zu RTR-Messungen als Erhebungsinstrument sind jedoch notwendig, um hier Sicherheit zu gewinnen (vgl. auch Kapitel 3.3).

Inhalte soll ausgeblendet werden.[30] Dadurch soll sichergestellt werden, dass bei der Verknüpfung der Debatteninhalte mit den ebenfalls sekundengenau aufgezeichneten RTR-Messungen zur Bewertung der Kandidaten nur Merkmale berücksichtigt werden, die in der zeitlichen Abfolge vor den Reaktionen der Rezipienten liegen.

Für diese Arbeit sind die Kategorien Thema und Relation relevant. Das *Thema* wurde auf sehr einfache Weise über die von den Moderatoren angekündigten Themenblöcke operationalisiert. Da das vorliegende TV-Duell sehr klar strukturiert war und die Moderatoren die Themenwechsel explizit kenntlich machten, ist die Einteilung sehr reliabel zu erfassen und sollte auch von den Rezipienten wahrgenommen worden sein. Um zu kontrollieren, ob diese Themenblöcke tatsächlich die diskutierten Themen widerspiegelten, haben wir zusätzlich die Passagen in den Aussagen der Kandidaten erfasst, die sich nach ihrem Inhalt auch einem anderen Themenblock zuordnen ließen. Dies war insgesamt in nur 276 Sekunden der Sprechzeiten der Kandidaten der Fall (Bachl, Käfferlein & Spieker, 2013b, Tabelle 3) und kann damit über das gesamte Duell hinweg betrachtet vernachlässigt werden. Die Kandidaten hielten sich größtenteils an die von den Moderatoren vorgegebenen Themen.

Die Operationalisierung der *Relationen* erfolgte in Anlehnung an die „functional analysis“ von W. L. Benoit (2007) und mit Rückgriff auf unsere Vorarbeiten zum Kanzlerduell 2009 (Spieker, 2011). Die Unterscheidung der Relationen baut auf dem im Exkurs zu Relationen und relationalen Strategien (S. 78ff.) erläuterten Grundverständnis der funktionalen Kommunikation zweier Kandidaten in der Situation eines TV-Duells auf. Grundsätzlich haben die Kandidaten zwei Möglichkeiten, um sich einen Vorteil gegenüber dem Kontrahenten zu verschaffen: Sie können ihre eigenen Leistungen, Positionen und Eigenschaften präsentieren, um sich selbst aufzuwerten. Oder sie können Informationen über den Kontrahenten thematisieren, von denen sie eine Abwertung des Gegners und in der Folge einen relativen Vorteil für sich selbst erwarten. Folgen wir der Logik, dass die Kandidaten die Relation ihrer Aussagen vor diesem Hintergrund strategisch auswählen, so haben alle Aussagen der Kandidaten über sich selbst und ihr politisches Lager die Intention, eine positive Bewertung durch die Rezipienten zu erzielen. Diese Aussagen bezeichnen wir mit der Ausprägung *Selbstpräsentation*.

Nach der relationalen Logik haben umgekehrt alle Aussagen eines Kandidaten über seinen Kontrahenten die funktionale Intention, den Gegner in der Wahrnehmung der Rezipienten abzuwerten. Dafür können explizit negative

30 Vgl. zu dieser Codierlogik ausführlich die allgemeinen Codieranweisungen in unserem Codebuch (Bachl, Käfferlein & Spieker, 2013a) sowie das Vorgehen von Reinemann und Maurer (2007a, S. 29) und deren in Nagel (2012) dokumentiertes Codebuch.

Formulierungen genutzt werden. Notwendig ist dies jedoch nicht, da wir davon ausgehen können, dass ein Kandidat in der Duellsituation nur Aussagen über den Kontrahenten tätigen würde, die diesem schaden sollen. Wir bezeichnen die Aussagen über den gegnerischen Kandidaten bzw. sein politisches Lager daher mit der Ausprägung *Angriff*. Bei der Interpretation dieser Ausprägung müssen wir darauf achten, dass es sich in den meisten Fällen um funktionale Angriffe mit dem Ziel der Abwertung des Gegenkandidaten handelt, nicht jedoch unbedingt um explizite Angriffe im Sinne der Alltagssprache.

Die dritte wichtige Relation ist die *Verteidigung*. Sie liegt vor, wenn ein Kandidat sich bzw. sein politisches Lager gegen vorangegangene Kritik oder Angriffe verteidigt. Notwendige Bedingung für die Codierung einer Verteidigung ist eine vorangegangene kritische Frage der Moderatoren oder ein Angriff des Gegenkandidaten. Eine Verteidigung wird jedoch nur dann erfasst, wenn der Kandidat explizit auf die in der Kritik angesprochenen Inhalte eingeht. Die Intention dieser Relation ist also das Entkräften der geäußerten Kritik. Reagiert ein Kandidat mit einem Gegenangriff oder indem er auf ein anderes Thema ausweicht, so wird dies nicht als Verteidigung erfasst.

Schließlich werden Aussagen der Kandidaten, die keinen Bezug zum eigenen oder gegnerischen Lager herstellen, sondern die Lage Dritter beschreiben, ohne die Verantwortung für diese Lage einem politischen Lager zuzuschreiben, mit der Ausprägung *Lagebewertung* erfasst. Zusätzlich wird erhoben, ob die Lagebeschreibung positiv oder negativ ist, und auf welches Objekt sich die Lagebeschreibung bezieht. Diese Relation kommt in den hier untersuchten Analyseeinheiten des Debatteninhalts allerdings nur selten vor und ist in dieser Arbeit nur in einer Teilauswertung in Kapitel 6.3.3 relevant.

Die Codierung der Relationen wurde von zwei Codierern durchgeführt. Dazu wurde das Duell in 24 Abschnitte unterteilt, die nach der Codiererschulung in ihrer Reihenfolge randomisiert auf die beiden Codierer verteilt wurden. Die Reliabilität der Kategorie ist zufriedenstellend: Krippendorffs $\alpha = .83$ (berechnet nach Hayes & Krippendorff, 2007). Die Verteilung der Relationen auf Kandidaten und Themenblöcke im gesamten TV-Duell berichten wir ausführlich in Bachl, Käfferlein und Spieker (2013b). Auf die Verteilung in den in Kapitel 6 verwendeten Analyseeinheiten des Debatteninhalts gehen wir jeweils zu Beginn der Auswertungen ein.

4.2 Qualität der RTR-Messungen

Da in dieser Arbeit die Analyse der rezeptionsbegleitend gemessenen Kandidatenbewertungen im Mittelpunkt steht, gehen wir im folgenden Teilkapitel näher

auf die Qualität der RTR-Messungen ein. Zuerst evaluieren wir die Reliabilität der Messungen, dann betrachten wir einige Indikatoren für ihre Validität.

4.2.1 Reliabilität

Eine „klassische“ Bestimmung der Reliabilität von RTR-Messungen als Zuverlässigkeit, bei der Anwendung desselben Instruments auf dieselben Bewertungsobjekte durch dieselben Versuchspersonen dieselben Ergebnisse zu erhalten, ist nicht unproblematisch (vgl. Kapitel 3.3). Ein echtes Retest-Verfahren ist schwierig zu realisieren, da davon ausgegangen werden muss, dass sich die spontane erste Reaktion auf einen Stimulus von der wiederholten Bewertung unterscheidet. Ebenso sind ein einfaches Split-Half-Verfahren bzw. die abgeleiteten Koeffizienten zur Ermittlung der internen Konsistenz der Messung problematisch, da die RTR-Skala nur aus einem Item besteht, das aber sehr häufig (einmal pro Sekunde) gemessen wird. Aus demselben Grund ist ein typischer Parallel-Test ebenfalls nicht möglich, auch wenn mit dem Vergleich der Messungen unterschiedlicher RTR-Techniken zu demselben Stimulus ein ähnlicher Test existiert (J. Maier et al., 2007; Reinemann et al., 2005). Da uns jedoch keine solche Vergleichsmessung mit einem anderen Instrument zur Verfügung steht, können wir auch diesen Test nicht durchführen.

Ein weiteres Problem betrifft die Unterscheidung der aggregierten und der individuellen RTR-Messungen – eine Unterscheidung, die in der Methodenliteratur in Hinblick auf die Reliabilität der Messungen unseres Wissens noch nicht explizit thematisiert wurde. Da wir in dieser Arbeit sowohl Auswertungen der aggregierten als auch der individuellen RTR-Messungen durchführen, wollen wir Indikatoren für die Reliabilität auf beiden Ebenen bestimmen.

Reliabilität der aggregierten RTR-Zeitreihen In Kapitel 5 analysieren wir vier RTR-Zeitreihen: die Bewertung der Kandidaten im Verlauf der Debatte durch das gesamte Publikum sowie durch die Anhänger der beiden Kandidaten und die Unentschiedenen.[31] Um die Zuverlässigkeit dieser Zeitreihen zu bestimmen, ziehen wir zwei Indikatoren heran. Zuerst prüfen wir, ob die Durchführung der Datenerhebung an zwei Standorten einen Einfluss auf die aggregierten RTR-Messungen hat. Dazu bilden wir die vier Zeitreihen für beide Standorte und betrachten die Korrelation zwischen den entsprechenden

31 Auf das Vorgehen zur Berechnung der Zeitreihen sowie die Konsequenzen der Aggregation der individuellen Messungen für Personengruppen gehen wir in Kaptitel 5 ausführlich ein. An dieser Stelle sei lediglich darauf hingewiesen, dass die aggregierten RTR-Zeitreihen durch die Bildung des (gewichteten) Mittelwerts der RTR-Bewertungen aller Rezipienten in einer Gruppe zu einer Sekunde berechnet werden.

Zeitreihen. Es zeigt sich eine große Parallelität der aggregierten RTR-Verläufe. Die Zeitreihen der gesamten Teilstichproben korrelieren mit $r = .88$. Auch die Bewertungen durch die Anhänger von Schmid ($r = .81$) und Mappus ($r = .80$) sowie durch die Unentschiedenen ($r = .79$) weisen im Zeitverlauf starke positive Korrelationen zwischen den Standorten auf. Die verbleibenden Unterschiede zwischen den Standorten können auch darauf zurückgeführt werden, dass die Zuschauer in Abhängigkeit vom Standort der Erhebung stellenweise unterschiedliche Bewertungen abgegeben haben. Zu denken ist in diesem Kontext beispielsweise an die Reaktionen auf die Diskussion um „Stuttgart 21", die in Stuttgart eine größere Volatilität zeigen als in Ravensburg. Insgesamt sprechen die starken Korrelationen für die Zuverlässigkeit des zur RTR-Messung eingesetzten Instruments. Unabhängig davon, ob es im Hohenheimer oder im Ravensburger Versuchssetting eingesetzt wurde, ergeben sich sehr ähnliche Verläufe der aggregierten RTR-Messungen.

Eine weitere Möglichkeit, die Reliabilität der aggregierten RTR-Verläufe einzuschätzen, ist der Vergleich der aggregierten RTR-Zeitreihen von zwei zufällig gebildeten Teilgruppen der Stichprobe.[32] Wenn die aggregierten Zeitreihen zuverlässig Auskunft über die Bewertung der Kandidaten durch die gesamte Gruppe geben, sollten zwei Zeitreihen, die jeweils aus den Bewertungen der Hälfte der Rezipienten gebildet wurden, sehr stark miteinander korrelieren. Wäre dies nicht der Fall, so wäre der Verlauf der aggregierten Zeitreihe in hohem Maße von der individuellen Zusammensetzung der Stichprobe abgängig. Damit wäre z.B. die aggregierte Bewertung der Kandidaten durch die Anhänger von Mappus in der Stichprobe kein zuverlässiger Indikator für die Bewertung durch das schwarz-gelbe Lager insgesamt.

Um die Zuverlässigkeit der aggregierten Zeitreihen nach diesem Prinzip zu bestimmen, wenden wir ein Resampling-Verfahren an. Das Resampling-Verfahren besteht aus drei Schritten:

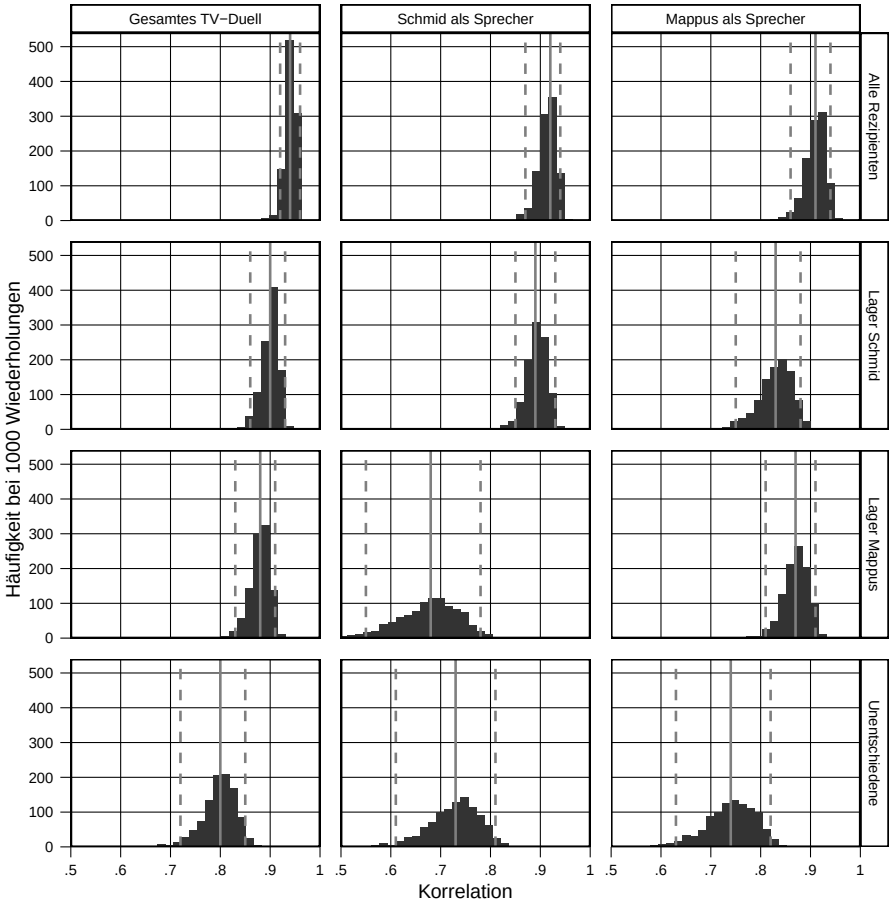
1. Die Stichprobe wird nach dem Zufallsprinzip in zwei Teilstichproben mit dem gleichen Umfang aufgeteilt.
2. Für beide Teilstichproben werden die aggregierten RTR-Zeitreihen berechnet.
3. Die Korrelation (Pearsons r) zwischen den Zeitreihen wird berechnet.

32 Maurer und Reinemann (2009, S. 9) bezeichnen dieses Verfahren zur Messung der Reliabilität als „split-half design", da die Probanden zufällig in zwei Gruppen aufgeteilt werden. Wir vermeiden diesen Begriff an dieser Stelle, um Verwechslungen mit der „klassischen" Split-half-Reliabilität oder Testhalbierungsreliabilität, bei der die Items eines Instruments in zwei Gruppen aufgeteilt und deren Korrelation in *einer* Stichprobe als Maß der Reliabilität herangezogen wird (Bortz & Döring, 2006, S. 198), zu vermeiden.

Die Schritte 1-3 werden sehr häufig wiederholt und die daraus resultierenden Korrelationen zwischen den RTR-Zeitreihen der Teilstichproben notiert. Aus den Verteilungen der Korrelationen bei vielen Durchgängen des Resampling-Verfahrens können wir darauf schließen, wie sehr der Verlauf der aggregierten RTR-Zeitreihen von der Zusammensetzung der Stichprobe beeinflusst wird. Viele starke Korrelationen sprechen für eine gute Reliabilität der aggregierten RTR-Bewertungen über die Zeit. Liegen die Korrelationen dagegen niedriger und schwanken sie stark zwischen den Wiederholungen, so hat die individuelle Zusammensetzung der Stichprobe einen größeren Einfluss auf die aggregierte RTR-Zeitreihe. Sie ist dann kein zuverlässiger Indikator für den Verlauf der Kandidatenbewertung über die Stichprobe hinaus.

Abbildung 4.2 stellt die Verteilung der Korrelationen in jeweils 1000 Wiederholungen des Resampling-Verfahrens in Form von Histogrammen dar. In den Spalten sind die Befunde für drei aggregierte RTR-Zeitreihen abgebildet. Links findet sich die Verteilung der Korrelationen für die Zeitreihen über den gesamten Verlauf des TV-Duells hinweg. In den beiden weiteren Spalten werden nur die Ausschnitte der Zeitreihen berücksichtigt, in denen Schmid bzw. Mappus das Wort hatte. Die Zeilen differenzieren die Befunde nach den Gruppen, deren aggregierte RTR-Verläufe wir in Kapitel 5 analysieren. In der ersten Zeile wird die Zuverlässigkeit der Kandidatenbewertungen durch das gesamte Publikum geprüft. Die folgenden Zeilen zeigen die Ergebnisse für die Anhänger der beiden Kandidaten und die Unentschiedenen. In allen Facetten sind zusätzlich der Median (durchgezogen) sowie das 2.5- und 97.5-Perzentil (gestrichelt) als vertikale Linien eingetragen. Der Bereich zwischen den roten Linien entspricht damit dem 95%-Resampling-Konfidenzintervall der Korrelation zwischen den RTR-Zeitreihen aus zwei zufällig gebildeten Hälften der jeweiligen (Teil-) Stichprobe.

In Abhängigkeit von der betrachteten Zeitreihe (Gesamtes Duell vs. Ausschnitte mit einem Sprecher) und den den Aggregaten zugrunde liegenden Rezipientengruppen zeigen sich drei Muster: *Erstens* fällt auf, dass die Zeitreihen, die auf größeren Fallzahlen basieren, zuverlässiger sind. Die Mediane der Verteilungen für die drei Zeitreihen auf Basis aller Rezipienten liegen über $r = .90$, die unteren Grenzen der Konfidenzintervalle über $r = .85$. Die im Vergleich niedrigsten mittleren Korrelationen und weitesten Konfidenzintervalle finden sich für die Zeitreihen der kleineren Gruppen der Unentschiedenen und der Anhänger von Mappus. Dieser Befund lässt sich direkt durch die Logik der Aggregatbildung erklären. Je größer die Gruppe ist, für die eine Mittelwert-Zeitreihe gebildet wird, desto geringer ist die statistische Unsicherheit um den Mittelwert, und desto zuverlässiger ist auch die Schätzung des Mittelwerts. Auf



Anmerkungen
Verteilungen der Korrelationen zwischen den Zeitreihen bei 1000 Wiederholungen des Resampling-Verfahrens. *Spalten*: Gesamte RTR-Zeitreihe bzw. nur die Ausschnitte der Zeitreihe, in denen Schmid bzw. Mappus sprechen. *Zeilen*: Zeitreihen auf Basis aller Rezipienten (n = 176), der Anhänger von Schmid (n = 89), der Anhänger von Mappus (n = 48) und der Unentschiedenen (n = 39). *Durchgezogene Linie*: Median; *Gestrichelte Linien*: 2.5- und 97.5-Perzentile.

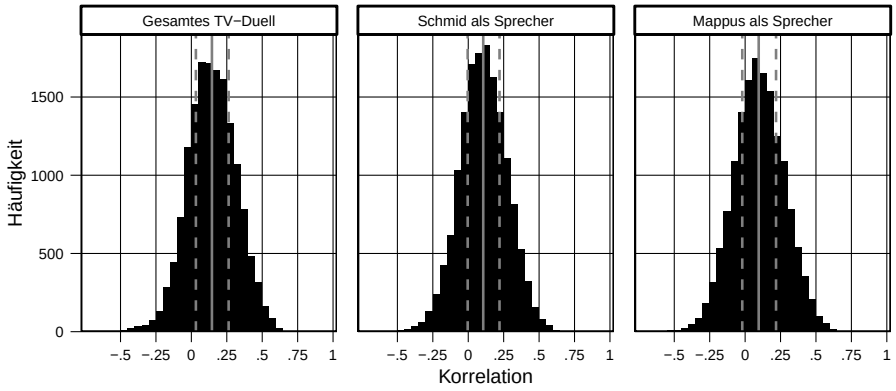
Abbildung 4.2: Verteilungen der Korrelationen im Resampling-Verfahren

diese Konsequenz der Aggregationslogik werden wir in Kapitel 5 ausführlich eingehen, wenn wir uns mit der Präzision der Zeitreihen beschäftigen.

Zweitens ist die Zuverlässigkeit der Zeitreihen über das gesamte Duell hinweg in allen Gruppen größer als die Zuverlässigkeit der Zeitreihen, die nur die Passagen eines Sprechers erfassen. Dies lässt sich zum Teil durch eine Regelmäßigkeit in den individuellen RTR-Bewertungen erklären. Während der Fragen der Moderatoren stellen die meisten Probanden ihren RTR-Regler auf die neutrale Ausgangsstellung der Skala, da sie gerade von keinem der Kandidaten einen positiven oder negativen Eindruck haben. Diese in zeitlicher Perspektive parallelen Regler-Bewegungen, die durch einen von den meisten Rezipienten einheitlich wahrgenommenen Reiz (ein Moderator spricht) ausgelöst werden, finden sich auch in den aggregierten Kurven wieder. Auch dieser Aspekt gehört zur Zuverlässigkeit der RTR-Messung, da er für eine interindividuell einheitliche Anwendung des Instruments spricht. In den Ausschnitten aus den Zeitreihen, in denen nur jeweils ein Kandidat das Wort hat, sind diese Passagen, in denen die Korrelationen besonders hoch sind, nicht enthalten. Die Gesamtkorrelation fällt daher (etwas) geringer aus.

Drittens zeigt sich eine interessante Interaktion zwischen der Zugehörigkeit zum Lager eines der Kandidaten und der Ausschnitte der Zeitreihen, in denen nur die Bewertungen während der Sprechzeiten eines der Kandidaten erfasst sind. Die Zeitreihen der Bewertung der Kandidaten durch die jeweils eigenen Anhänger (Schmid als Sprecher, Lager Schmid: $Mdn_r = .89, 95\%KI_r = [.85; .93]$; Mappus als Sprecher, Lager Mappus: $Mdn_r = .87, 95\%KI_r = [.81; .91]$) sind zuverlässiger als die Zeitreihen, die die Bewertung eines Kandidaten durch die Anhänger des Kontrahenten erfassen (Schmid als Sprecher, Lager Mappus: $Mdn_r = .68, 95\%KI_r = [.55; .78]$; Mappus als Sprecher, Lager Schmid: $Mdn_r = .83, 95\%KI_r = [.75; .88]$). Dies deutet darauf hin, dass die Bewertungen der eigenen Kandidaten relativ einheitlich verlaufen und damit über die individuellen Rezipienten hinweg zuverlässiger sind als die Bewertungen des gegnerischen Kandidaten. Zu diesem Befund passt auch, dass die aggregierten Zeitreihen der Unentschiedenen vergleichsweise uneinheitlich sind. Sie liegen in etwa auf dem Niveau der Bewertungen der Kandidaten durch das gegnerische Lager. In dieser Gruppe fehlt eine homogene Einstellung gegenüber den Kandidaten als der Faktor, der in den Bewertungen der Kandidaten durch ihre eigenen Anhänger für einheitliche RTR-Verläufe sorgt.

Insgesamt weisen die Verteilungen der Korrelationen auf eine zufriedenstellende bis sehr hohe Reliabilität der aggregierten RTR-Zeitreihen hin. Die insgesamt niedrigste Untergrenze eines 95%-Resampling-Konfidenzintervalls liegt bei $r = .55$ (Schmid als Sprecher, Lager Mappus), die niedrigste Untergren-



Anmerkungen
Verteilungen der Korrelationen (Pearsons r) zwischen den individuellen RTR-Zeitreihen aller 15400 Rezipienten-Paare. Die Facetten zeigen die Verteilungen der Korrelationen für die gesamte Zeitreihe des TV-Duells und die Ausschnitte der Zeitreihe, in denen Schmid bzw. Mappus sprechen. *Durchgezogene Linie*: Median; *Gestrichelte Linien*: 25- und 75-Perzentile, Interquartilsrange.

Abbildung 4.3: Verteilungen der Korrelationen zwischen den individuellen Zeitreihen

ze der für die Analysen in Kapitel 5 relevanten Zeitreihen über den gesamten Duellverlauf bei $r = .72$ (Unentschiedene).

Reliabilität der individuellen RTR-Messungen Bisher haben wir geprüft, ob die aggregierten RTR-Messungen robust gegen die Anwendung des Instruments an den beiden Standorten und gegen die individuellen Zusammensetzungen der (Teil-) Stichproben sind. In einem zweiten Schritt wollen wir zwei Indikatoren für die Reliabilität der individuellen RTR-Messungen heranziehen. Zuerst prüfen wir, wie sich die individuellen RTR-Verläufe im Verhältnis zueinander entwickeln. Dazu betrachten wir alle 15400 paarweisen Korrelationen zwischen den individuellen RTR-Zeitreihen der 176 Rezipienten, die sich einem der drei Kandidatenlager zuordnen lassen und für die gültige RTR-Messungen vorliegen. Die Verteilungen der Korrelationen zwischen den Zeitreihen des gesamten TV-Duells und den Ausschnitten, in denen Schmid bzw. Mappus sprechen, sind in Abbildung 4.3 als Histogramme dargestellt. Zusätzliche sind die Mitte (Median, durchgezogen) und die Grenzen der Quartile (gestrichelt) der Verteilungen verzeichnet.

Die mittlere Hälfte der Zeitreihen-Paare weist in allen drei Betrachtungen so gut wie keinen bis einen schwach positiven Zusammenhang auf. Das zweite Quartil beginnt jeweils ca. bei einer Korrelation von $r = 0$, das dritte Quartil endet bei Korrelationen von leicht über $r = .25$. Daraus können wir schließen, dass ein großer Teil der individuellen RTR-Verläufe bei einer genau zeitgleichen Betrachtung in keiner systematischen Beziehung zueinander steht. Eine Erklärung hierfür könnte sein, dass der Zusammenhang zwischen zwei individuellen Zeitreihen davon abhängt, wie ähnlich bzw. unähnlich sich die jeweiligen Rezipienten in ihren politischen Einstellungen sind. Es wäre zu vermuten, dass die Echtzeiturteile von zwei Probanden, deren politische Einstellungen in hohem Maße übereinstimmen, recht gleichförmig verlaufen. Die RTR-Messungen von zwei Probanden, deren politische Einstellungen sich stark widersprechen, sollten sich gegenläufig entwickeln. Um diese Vermutung zu prüfen, haben wir die Ähnlichkeit bzw. Unähnlichkeit jedes Rezipientenpaars bestimmt als ihre euklidische Distanz auf den Merkmalen Voreinstellungen zu den Kandidaten, Voreinstellung zu deren Parteien und Lagerzugehörigkeit. Ein Probandenpaar hat eine euklidische Distanz von 0, wenn die Probanden in diesen Variablen perfekt übereinstimmen. Je stärker sich die Werte hingegen unterscheiden, desto größer fällt die euklidische Distanz aus. Das Distanzmaß korreliert jedoch nur schwach bis mäßig negativ mit dem Zusammenhang zwischen ihren Zeitreihen: Gesamtes TV-Duell: $r = -.23$; Sprecher Schmid: $r = -.16$; Sprecher Mappus: $r = -.16$. So ergeben sich auch kaum größere Unterschiede in den Verteilungen der Zusammenhänge, wenn wir sie sortiert nach der Lagerzugehörigkeit der Rezipienten betrachten (vgl. Abbildungen A.1, A.2, A.3).

In dieser Betrachtungsweise erweisen sich die individuellen RTR-Messungen als wenig reliabel. Dieser Befund ist eigentlich wenig überraschend. Erstens reagieren die individuellen Rezipienten eben nicht deterministisch „wie auf Knopfdruck“ auf Variationen im Stimulus. Und zweitens bleiben die rezeptionsbegleitend abgegebenen Urteile individueller Probanden häufig über längere Zeit hinweg konstant, da die Rezipienten nicht ständig ihren Eindruck von den Kandidaten aktualisieren (vgl. dazu auch die Abbildungen der individuellen RTR-Verläufe in Kapitel 5.2). Wenn in einem Abschnitt einer Zeitreihe keine Varianz vorliegt, kann über diesen Abschnitt auch keine Kovarianz mit einer anderen Zeitreihe auftreten. Intuitiv liegt die Vermutung nahe, dass die größtenteils geringen paarweisen Korrelationen der sekundengenauen Zeitreihen durch eine zeitliche Verschiebung ähnlicher Reaktionen auf denselben Stimulus zustandekommen. So mögen Rezipienten durchaus auf eine Kandidatenaussage mit einer positiven Bewertung des Kandidaten reagieren. Die Veränderung des RTR-Werts tritt aber eben nicht zwingend in genau derselben Sekunde auf.

Die Vermutung, dass die im Mittel geringen paarweisen Korrelationen vor allem auf geringe zeitliche Verschiebungen ähnlicher Veränderungen zurückgehen, können wir empirisch prüfen. Dazu betrachten wir die paarweisen Zusammenhänge zwischen individuellen RTR-Zeitreihen, bei denen jeweils die RTR-Messungen in den aufeinanderfolgenden drei bzw. fünf Sekunden zusammengefasst sind. Doch auch in diesen Analysen fallen die paarweisen Korrelationen nicht wesentlich höher aus. Für die Zusammenfassung der gesamten Zeitreihe beider Sprecher zu drei- bzw. fünfsekündigen Segmenten liegt der Median fast unverändert bei $r = .150$ ($Q_{25} = .033$, $Q_{75} = .271$) bzw. $r = .154$ ($Q_{25} = .034$, $Q_{75} = .277$). Zum Vergleich: Die Verteilung der paarweisen Korrelationen auf Basis der sekundengenauen Zeitreihe hat einen Median von $r = .145$ mit einem Interquartilsrange von $Q_{25} = .032$ bis $Q_{75} = .263$. Auch die Zusammenhänge der Korrelationen mit der Distanz der politischen Einstellungen der Befragten ändert sich durch die Zusammenfassung der Messungen nicht. Der Befund der größtenteils schwachen paarweisen Korrelationen geht also offenbar kaum auf leichte zeitliche Verschiebungen der Reaktionen auf den Stimulus zurück. Die Reliabilität der sekundengenauen individuellen RTR-Messungen muss nach dieser Betrachtungsweise als gering gelten. Vor dem Hintergrund der zu erwartenden nicht-deterministischen Reaktionen individueller Rezipienten auf die Variationen im Stimulus ist dies nicht unbedingt überraschend. Die Befunde fordern aber Konsequenzen für Analysen der RTR-Messungen auf Individualebene. Solche Modelle müssen eine interindividuelle Flexibilität in der Definition der Zeiträume erlauben, in denen sich die Effekte von Stimulusmerkmalen in den Reaktionen der Rezipienten zeigen. Eine Möglichkeit hierfür stellen wir in dieser Arbeit vor, wenn wir in Kapitel 6 die RTR-Bewertungen in längeren zeitlichen Abschnitten gemeinsam betrachten. Die (kreuzklassifizierten) Wachstumskurvenmodelle (Kapitel 6.1 und 6.3) zeigen, wie sich innerhalb dieser längeren Abschnitte auch Veränderungen der RTR-Messungen modellieren lassen, ohne dass dafür zeitlich genau gleichförmige Verläufe der individuellen Messungen notwendig sind.

Ein zweiter Aspekt der Reliabilität, den wir für die individuellen RTR-Messungen bestimmen können, ist die interne Konsistenz, die üblicherweise mit dem Koeffizienten Cronbachs α quantifiziert wird (Bortz & Döring, 2006, S. 198-199). Da der Betrag des Koeffizienten auch mit der Zahl der getesteten Items steigt, ist es nicht zielführend, sämtliche individuellen RTR-Messungen zur Berechnung heranzuziehen. Papastefanou (2013, S. 16) schlägt daher vor, mehrere aufeinanderfolgende RTR-Messungen zu „quasi-items“ zusammenzufassen. In unseren Analysen der individuellen RTR-Messungen sind wir an den Bewertungen der Kandidaten während zweier zeitlich definierter Analyse-

einheiten interessiert: den Turns und den Antworten.[33] Daher liegt es nahe, die Quasi-Items zur Berechnung von Cronbachs α durch eine Zusammenfassung der RTR-Messungen in diesen Einheiten zu bilden. Entsprechend der Analysen in Kaptitel 6 ermitteln wir getrennte Koeffizienten für die Turns und Antworten von Schmid und Mappus.

Für alle vier betrachteten Einheit-Sprecher-Kombinationen ergeben sich durchweg sehr hohe Beträge des Koeffizienten Cronbachs α: Turns-Schmid: $\alpha = .95$; Turns-Mappus: $\alpha = .96$; Antworten-Schmid: $\alpha = .92$; Antworten-Mappus: $\alpha = .95$. Die beim Test von Multi-Item-Skalen übliche Interpretation des Koeffizienten ist auf diesen speziellen Kontext allerdings nicht direkt übertragbar. Streng genommen eignet sich der Koeffizient dazu, die interne Konsistenz eines eindimensionalen Tests zu quantifizieren. Oder wie Cortina (1993, S. 103) detaillierter ausführt:

> Coefficient alpha is useful for estimating reliability in a particular case: when item-specific variance in a unidimensional test is of interest. If a test has a large alpha, then it can be concluded that a large portion of the variance in the test is attributable to general and group factors. This is important information because it implies that there is very little item-specific variance.

Wir können argumentieren, dass mit Schmid und Mappus immer dieselben Objekte bewerten werden und es sich damit um eindimensionale Tests handelt. Ein Teil der gemeinsamen Varianz aller Quasi-Items lässt sich vermutlich durch die allgemeine Einstellung zu den Kandidaten erklären („general factor“). Außerdem äußern die Kandidaten in mehreren Antworten bzw. Turns ähnliche Positionen. Ein weiterer Teil der gemeinsamen Varianz kann daher durch die Einstellungen der Rezipienten zu diesen Positionen erklärt werden („group factors“). Es bleibt die Item-spezifische Varianz, in unserem Fall die Varianz, die den Eigenschaften der Turns bzw. der Antworten zuzurechnen ist.

Nach dieser Analogie sprechen die hohen α-Werte dafür, dass die Bewertung der Kandidaten in den einzelnen Turns bzw. Antworten eine sehr hohe interne Konsistenz aufweist. Durch die Bewertungen der Kandidaten in den vielen Turns bzw. Antworten wird zuverlässig gemessen, wie Schmid und Mappus von den Rezipienten während des TV-Duells bewertet werden. Dies ist sicherlich eine gute Nachricht, wenn wir das Ziel haben, aus allen RTR-Messungen

33 Ein Turn ist zeitlich definiert als der Redebeitrag eines Kandidaten, in dem er ununterbrochen das Wort hat. Er beginnt, wenn ein Kandidat das Wort ergreift, und endet, wenn der andere Kandidat oder ein Moderator spricht. Eine Antwort ist zeitlich definiert als die ersten 30 Sekunden eines Redebeitrag eines Kandidaten, der direkt auf eine Frage der Moderatoren folgt. Zu Beginn von Kaptitel 6 (ab S. 173) gehen wir detailliert auf diese Definitionen ein.

einen einzelnen Indikator für die Bewertung der Kandidaten zu bilden. Wenn wir allerdings – wie in der hier vorliegenden Arbeit – Unterschiede in den unmittelbaren Kandidatenbewertungen auch durch die Debatteninhalte erklären möchten, ist eine derart hohe interne Konsistenz wenig ermutigend. Sie deutet bereits vor der eigentlichen Analyse darauf hin, dass große, alleine durch den Debatteninhalt verursachte Unterschiede in den ummittelbaren Kandidatenbewertungen kaum zu erwarten sind. Oder anders ausgedrückt: Die einzelnen Messungen der Kandidatenbewertungen sind überaus konsistent, obwohl sie in der Folge unterschiedlicher Aussagen der Kandidaten abgegeben wurden. Der Einfluss der Aussageninhalte auf die unmittelbare Bewertung ist daher vermutlich recht gering.

4.2.2 Validität

Im folgenden Abschnitt prüfen wir die Validität der RTR-Messungen. Dabei beschränken wir uns an dieser Stelle auf die Validität der RTR-Messungen als Messinstrument (Bortz & Döring, 2006, S. 200). Ziel ist es damit, zu zeigen, dass die vorliegenden RTR-Messungen das erfasst haben, was sie erfassen sollen: den unmittelbaren Eindruck der Rezipienten von Schmid und Mappus während des TV-Duells. Dafür betrachten wir Indikatoren für die Kriteriumsvalidität. Auf eine Interpretation der Inhaltsvalidität (auch „Augenscheinvalidität“ Bortz & Döring, 2006, S. 200) verzichten wir an dieser Stelle, da sie in der Regel nur als *ex post* Plausibilisierung der Analyseergebnisse möglich ist und somit kaum über die inhaltliche Interpretation der Befunde hinausgeht. Ein ausführlicher Nachweis der Konstruktvalidität ist anhand der Daten dieser nicht als Test des RTR-Instruments angelegten Studie nicht möglich.

Übereinstimmungsvalidität Zuerst betrachten wir die Zusammenhänge zwischen den vor dem TV-Duell gemessenen Einstellungen der Rezipienten zu den Kandidaten und ihren Parteien sowie ihrer Lagerzugehörigkeit und den über den Duellverlauf hinweg zusammengefassten RTR-Messungen (Tabelle 4.2). Zu erwarten sind positive Korrelationen der Zugehörigkeit zum Lager Schmid sowie der Einstellung zu Schmid und zur SPD mit der RTR-Bewertung von Schmid (M_{Schmid}). Der Zusammenhang der Zugehörigkeit zum Lager Mappus sowie der Einstellung zu Mappus und zur CDU mit der RTR-Bewertung von Schmid sollte negativ sein. Für die RTR-Bewertung von Mappus (M_{Mappus}) ist von der umgekehrten Richtung der Zusammenhänge auszugehen. Da die RTR-Skala für das gesamte Duell (M_{Gesamt}) als Differential von „größter Vorteil Schmid“ bis „größter Vorteil Mappus“ reicht, sind mit den auf Schmid bezoge-

nen Voreinstellungen negative Korrelationen zu erwarten, mit den auf Mappus bezogenen Voreinstellungen dagegen positive Korrelationen.

Tabelle 4.2: Zusammenhänge zwischen Pre-Duell-Messungen und RTR-Messungen

	RTR-Messungen		
Voreinstellungen	M_{Gesamt}	M_{Schmid}	M_{Mappus}
Lager Schmid	−.60	.53	−.54
Lager Mappus	.54	−.44	.51
Skalometer Schmid	−.32	.27	−.30
Skalometer Mappus	.59	−.50	.55
Skalometer SPD	−.26	.23	−.23
Skalometer CDU	.56	−.46	.54

Anmerkungen
Korrelationen (Pearsons *r*) der über das TV-Duell zusammengefassten RTR-Messungen mit den vor dem TV-Duell gemessenen politischen Einstellungen. M_{Gesamt}: Zusammenfassung aller RTR-Bewertungen während der Sprechzeiten der beiden Kandidaten auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). M_{Schmid}, M_{Mappus}: Zusammenfassung aller RTR-Bewertungen während der Sprechzeiten des jeweiligen Kandidaten auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher).
n = 174-176; Korrelationen des Skalometer SPD mit den RTR-Messungen $p < .01$, alle anderen $p < .001$.

Im Hinblick auf die Richtung der Zusammenhänge entsprechen die Befunde den Erwartungen. Auch sprechen die Beträge der meisten Korrelationskoeffizienten für eine zumindest mittelmäßige (ab $r > .4$) bis hohe (ab $r > .6$) Validität der RTR-Messungen (Bortz & Döring, 2006, S. 202). Schwächere Korrelationen zeigen sich jedoch mit den Voreinstellungen zu Schmid und zur SPD. In Anbetracht der übrigen Ergebnisse spricht dies jedoch eher für eine schlechtere Indikator-Funktion dieser vor dem Duell gemessenen Einstellungen. Der Zusammenhang zwischen der Lagerzugehörigkeit der Rezipienten und den RTR-Messungen lässt sich auch varianzanalytisch darstellen. Alle drei Maße der RTR-Messungen unterscheiden sich signifikant nach der Lagerzugehörigkeit der Rezipienten (M_{Gesamt}: $F(2, 173) = 58.08$, $p < .001$; M_{Schmid}: $F(2, 173) = 36.80$, $p < .001$; M_{Mappus}: $F(2, 173) = 45.22$, $p < .001$). Diese Analyselogik entspricht der „Technik der bekannten Gruppen“ (Bortz & Döring, 2006, S. 201) zur Feststellung der Übereinstimmungsvalidität: Die gemessenen Konstrukte – hier die unmittelbaren Eindrücke von den Kandidaten im TV-Duell – unterscheiden sich signifikant und in erwarteter Richtung nach den Charakteristika der Versuchspersonen – hier den politischen Voreinstellungen der Rezipienten. Einschränkend müssen wir festhalten, dass wir an dieser Stelle nur die Vali-

dität der über das TV-Duell zusammengefassten RTR-Messungen untersucht haben. In den späteren Analysen dieser Arbeit werden wir zeigen, dass sich sowohl die aggregierten RTR-Zeitreihen (vgl. Kapitel 5, z.B. Abbildung 5.2) als auch die individuellen RTR-Verläufe (vgl. Kapitel 6, z.B. Abbildungen 6.8, 6.18, 6.39) deutlich nach den Voreinstellungen der Rezipienten unterscheiden. Auch diese Befunde können im Sinne einer Übereinstimmungsvalidität der RTR-Messungen interpretiert werden.

Als einen weiteren Indikator der Übereinstimmungsvalidität können wir die Korrelation zwischen den unmittelbaren Bewertungen von Schmid und Mappus während der Debatte heranziehen. Hier erwarten wir eine diskriminanten Zusammenhang, das heißt, mit einem positiven Eindruck von einem Kandidaten während des Duells sollte ein negativer Eindruck vom anderen Kandidaten einhergehen. Erwartungskonform findet sich eine starke negative Korrelation von $r = -.59$ $(p < .001)$ zwischen den zu beiden Sprechern zusammengefassten RTR-Messungen. Je besser ein Rezipient den einen Kandidaten während des Duells bewertet, desto schlechter bewertet er den anderen Kandidaten.

Prognostische Validität Um die prognostische Validität der RTR-Messungen zu bestimmen, untersuchen wir, ob die unmittelbaren Kandidatenbewertungen vorhersagen, wie die Debattenleistung der Kandidaten in der Befragung nach dem TV-Duell eingeschätzt wird. Tabelle 4.3 stellt dafür zunächst die Korrelationen zwischen den für das gesamte Duell und die Sprechzeiten der beiden Kandidaten zusammengefassten RTR-Messungen und den nach dem Duell gemessenen Bewertungen der Debattenleistungen dar.

Fast durchgängig zeigen sich mittlere bis hohe Korrelationen in den erwarteten Richtungen, was für eine zumindest mittelmäßige bis hohe prognostische Validität der zusammengefassten RTR-Messungen spricht. Dabei ist die Prognosekraft der unmittelbaren Bewertungen für die Einschätzung der Debattenleistung von Schmid durchweg etwas geringer als für die Debattenleistung von Mappus und die differentielle Betrachtung der beiden Kandidaten. Insgesamt können wir jedoch davon ausgehen, dass die mit dem RTR-Instrument gemessenen unmittelbaren Eindrücke zusammengenommen etwas ähnliches messen wie die Bilanz, die die Rezipienten nach dem TV-Duell zu ihrem Eindruck von den beiden Kandidaten ziehen. In einer ausführlicheren Wirkungsanalyse zum TV-Duell haben wir zudem gezeigt, dass die Erklärungskraft der RTR-Messungen auch unter Kontrolle der Voreinstellungen besteht, und dass die RTR-Messungen darüber hinaus auch zur Erklärung der Kandidatenbewertung nach der Debatte beitragen (Bachl, 2013b).

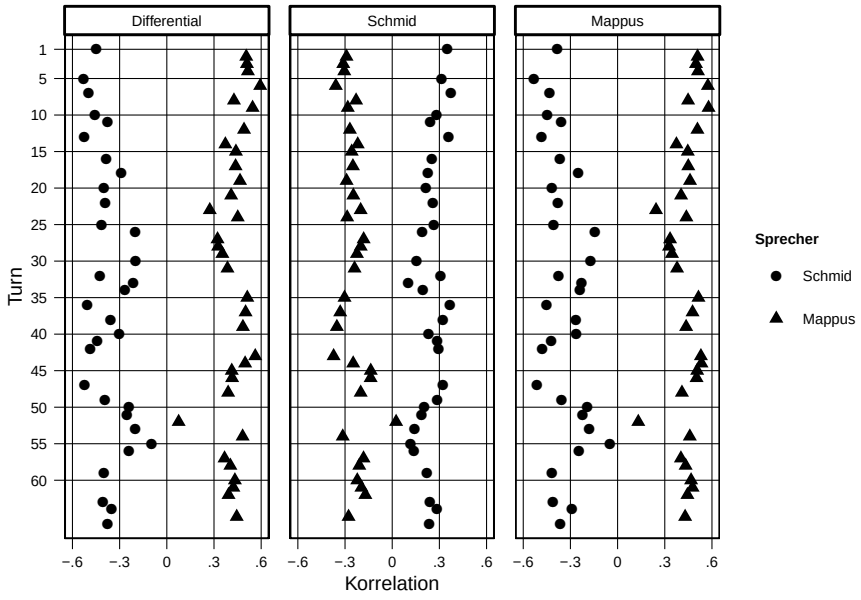
Tabelle 4.3: Zusammenhänge zwischen RTR-Messungen und Post-Duell-Messungen

	Debattenleistung		
RTR-Messungen	Differential	Schmid	Mappus
M_{Gesamt}	.69	−.44	.68
M_{Schmid}	−.57	.40	−.53
M_{Mappus}	.66	−.39	.67

Anmerkungen
Korrelationen (Pearsons *r*) der über das Duell zusammengefassten RTR-Messungen mit den nach dem TV-Duell gemessenen Bewertungen der Debattenleistung. M_{Gesamt}: Zusammenfassung aller RTR-Bewertungen während der Sprechzeiten der beiden Kandidaten auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). M_{Schmid}, M_{Mappus}: Zusammenfassung aller RTR-Bewertungen während der Sprechzeiten des jeweiligen Kandidaten auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Kandidatenleistung Differential: Leistung Mappus minus Leistung Schmid.
n = 175-176; Alle Korrelationen $p < .001$.

Neben dieser summarischen Betrachtung aller RTR-Messungen interessiert uns auch, wie stark die Bewertungen der Kandidaten während der einzelnen Turns mit der abschließenden Gesamtbewertung ihrer Leistung zusammenhängen. Dazu sind in Abbildung 4.4 die Korrelationen der mittleren RTR-Bewertungen während der Turns mit den drei Maßen der Debattenleistung aus der Befragung nach dem Duell dargestellt. Die Facetten der Abbildung differenzieren analog zu Tabelle 4.3 nach der differentiellen Verrechnung der Debattenleistung beider Kandidaten sowie ihren Einzelbewertungen.

Wie in der summarischen Betrachtung sind die Korrelationen zwischen den unmittelbaren Bewertungen während der einzelnen Turns und der Debattenleistung Schmids etwas geringer. Wenig überraschend fallen die Korrelationen in dieser sequentiellen Betrachtung etwas niedriger aus als in der summarischen Betrachtung in Tabelle 4.3. Im Zeitverlauf sind zudem durchaus einige Schwankungen festzustellen. Diese Befunde sind inhaltlich plausibel. Die Rezipienten gewichten bei der abschließenden Bewertung der Debattenleistung der Kandidaten nicht alle Eindrücke gleich. Manche Turns enthalten für das Gesamturteil wichtigere Informationen als andere. So zeigt sich beispielsweise, dass die Bewertungen während der Turns, in denen die Persönlichkeiten der Kandidaten diskutiert werden, vergleichsweise schwach mit den Gesamturteilen korrelieren. Einfache Recency- oder Primacy-Effekte, nach denen die letzten bzw. ersten Eindrücke besonders wichtig für die Urteilsbildung sind, finden sich zumindest in dieser einfachen Analyse nicht. Alles in allem zeigt sich auch für die unmittelbaren Eindrücke von den Kandidaten in den ein-



Anmerkungen
Korrelationen (Pearsons *r*) der über die Turns zusammengefassten RTR-Messungen mit den nach dem TV-Duell gemessenen Bewertungen der Kandidatenleistungen. Während des Duells: Bewertung des sprechenden Kandidaten auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Kandidatenleistung Differential: Leistung Mappus minus Leistung Schmid.

Abbildung 4.4: Zusammenhänge zwischen RTR-Messungen während einzelner Turns und Post-Duell-Messungen

zelnen Turns eine zufriedenstellende bis gute prognostische Validität. Wenn wir im Folgenden Einflüsse auf die Bewertung der Kandidaten in einzelnen Abschnitten der Debatte untersuchen, widmen wir uns also der Erklärung von Messungen, die auch mit dem Gesamturteil über den Auftritt der Kandidaten zusammenhängen.

5 Etablierte Verfahren zur Analyse der unmittelbaren Kandidatenbewertungen in TV-Duellen

Im folgenden Kapitel stellen wir die Verfahren vor, die sich zur Analyse des Zusammenhangs zwischen Debatteninhalten und unmittelbaren Kandidatenbewertungen etabliert haben. Um die Nachvollziehbarkeit der teils recht technischen Argumente zu erleichtern, erklären wir einführend die Struktur typischer RTR-Datensätze. Die Konsequenzen der Aggregations- und Analyseentscheidungen demonstrieren wir am Beispiel der unmittelbaren Kandidatenbewertungen im baden-württembergischen TV-Duell 2011. Auf dieser Basis diskutieren wir, welche Einschränkungen sich für die Gültigkeit und Interpretation der mit den etablierten Verfahren ermittelten Befunde ergeben. Außerdem machen wir Vorschläge, wie mit diesen Einschränkungen in der angewandten Datenanalyse und -interpretation umgegangen werden kann.

5.1 Struktur der Datensätze

In diesem Abschnitt stellen wir die Struktur typischer Datensätze vor, in denen die Informationen einer RTR-Studie zu einem TV-Duell erfasst sind. Dies soll vor allem Leserinnen und Lesern, die mit solchen Datensätzen nicht vertraut sind, das Nachvollziehen der folgenden Ausführungen zur Analyse der aggregierten Datensätze erleichtern.

Die Komplexität der Datenstruktur einer TV-Duell-Studie mit rezeptionsbegleitender RTR-Messung ist darin begründet, dass die Kandidatenbewertungen von vielen Personen in jeder Sekunde des Duells gemessen werden.[34] Allgemein gesprochen generiert ein solches Design viele Messwiederholungen bei vielen Untersuchungseinheiten. Damit unterscheidet sich die Datenstruktur von den üblicheren longitudinalen Datensätzen aus Panel- und Zeitreihenanalysen. In Panel-Designs werden in der Regel eine begrenzte Zahl von Messwiederholungen bei vielen Untersuchungseinheiten vorgenommen. So wurden z.B.

34 Gleiches gilt natürlich auch für andere rezeptionsbegleitende Messverfahren (z.B. physiologische Messungen), wenn sie in Studiendesigns mit größeren Stichproben und längeren Stimuli eingesetzt werden.

in der Panel-Studie der GLES zum Bundestagswahlkampf 2009 4552 Personen zu maximal sieben Messzeitpunkten befragt (Rattinger, Roßteutscher, Schmitt-Beck & Weßels, 2012). In diesem Datensatz ist die Person der Merkmalsträger, es gibt also 4552 Fälle mit bis zu sieben Variablensets, die zu jeweils einem Befragungszeitpunkt gemessen wurden. In Zeitreihen-Designs werden dagegen sehr viele Messungen bei wenigen Untersuchungseinheiten vorgenommen, wobei im Falle von Befragungen die „Messungen" häufig durch die Aggregation von Querschnittsbefragungen entstehen. Ein Beispiel aus der Kommunikationswissenschaft ist eine Zeitreihenanalyse zum Zusammenhang zwischen Thematisierung der Europäischen Einheit in den Fernsehnachrichten und der Relevanz des Themas in der deutschen Bevölkerung. Hier wurden die Berichterstattungshäufigkeit und die Relevanzzuweisung an 363 Tagen gemessen (Krause & Gehrau, 2007). In diesem Datensatz ist der Tag der Merkmalsträger, es gibt also 363 Fälle mit (für diesen einfachen bivariaten Zusammenhang) zwei Variablen.

Panel-Datensätze enthalten also wenige Messwiederholungen von vielen Untersuchungseinheiten. Zeitreihen-Datensätze enthalten viele Messwiederholungen von wenigen Untersuchungseinheiten. RTR-Datensätze enthalten als sogenannte „intensive longitudinal data" (Walls & Schafer, 2006, S. xi) im Vergleich zu Panel-Datensätzen viele Messwiederholungen und im Vergleich zu Zeitreihendatensätzen viele Untersuchungseinheiten. Der Personen-Datensatz unserer Rezeptionsstudie zum TV-Duell Mappus gegen Schmid enthält – je nach Art der Datenbereinigung – ca. 180 Probanden mit jeweils ca. 3900 RTR-Messungen der unmittelbaren Kandidatenbewertung.

Im Standard-Datenformat von SPSS, dem sogenannten „Wide-Format", entsteht so ein sehr breiter Datensatz, da jede Messwiederholung als Variable repräsentiert ist. Für jeden Probanden (Fall) sind in einer Zeile alle RTR-Messungen sowie die Personenvariablen gespeichert (vgl. Tabelle 5.1).

Das gängige SPSS-Format wird hier dargestellt, da es im für Personendatensätze (z.B. Datensätze aus Befragungen) typischen Format verdeutlicht, dass die RTR-Messungen bei den Personen gemessen werden, und dass die Personen zu einem Zeitpunkt unterschiedliche Urteile abgeben können. Zu jedem Zeitpunkt existiert eine Varianz zwischen den Probanden. Außerdem wird klar, dass die Informationen über den Inhalt der Debatte zu einem Zeitpunkt nicht in diesem Datensatz enthalten sind. Im SPSS-üblichen Wide-Format können den Variablen keine systematischen Informationen zugeordnet werden – dies ist nur für die Personen als Merkmalsträger möglich. Beziehungen können in diesem Datensatz nur zwischen den Personeneigenschaften und den RTR-Messungen hergestellt werden. Daher muss der Personendatensatz, der

Tabelle 5.1: Auszug aus einem Personen-Datensatz

ID	RTR_1	RTR_2	RTR_3	RTR_i	RTR_{3900}	lager	sk_map	sk_sch	...
1	0	0	14	...	2	−1	−4	2	...
2	1	29	42	...	1	−1	−5	3	...
3	−2	−2	−28	...	5	1	2	−1	...
4	2	2	2	...	4	1	3	0	...
...	...	...	...	...	...	...	...	...	...
180	−5	−5	−5	...	8	0	−3	1	...

Anmerkungen
ID: Personen-ID; RTR_x: RTR-Messungen; lager: Lagerzuordnung (−1: Schmid, 0: Unentschieden, 1: Mappus); sk_map, sk_sch: Skalometer Mappus, Schmid (−5: halte überhaupt nichts von [Kandidat] bis 5: halte sehr viel von [Kandidat]).

die RTR-Messungen enthält, umstrukturiert werden, um einen Bezug zu den Inhalten der Debatten zu einem bestimmten Zeitpunkt herstellen zu können.

Das in der Literatur prominenteste Vorgehen, eine Verknüpfung zwischen Debatteninhalten und Publikumsreaktionen zu einem bestimmten Zeitpunkt zu ermöglichen, ist die Zusammenfassung der RTR-Messungen aller (oder von Gruppen von) Rezipienten zu einem bestimmten Zeitpunkt. Üblicherweise werden nach diesem Prinzip die Messungen zu kontinuierlichen Zeitabständen (z.B. pro Sekunde) zusammengefasst. So entstehen RTR-Zeitreihen, die mit Peak-Spike-Analysen oder statistischen Verfahren der Zeitreihenanalyse untersucht werden können. Alternativ können auch die RTR-Messungen zu inhaltlich definierten Zeitabschnitten, z.B. in einer Inhaltsanalyse identifizierten Aussagen, über die Probanden hinweg zusammengefasst werden. Wir wenden uns im Folgenden zuerst ausführlich der Aggregation zu RTR-Zeitreihen zu, da fast alle vorliegenden Analysen auf dieser Verdichtung der Daten aufbauen.

Berechnen der RTR-Zeitreihen

Die einfachste Möglichkeit, einen Eindruck davon zu erhalten, wie die Kandidaten im Duellverlauf von allen Rezipienten bewertet werden, ist die Berechnung einer Zeitreihe als Mittelwert der RTR-Bewertungen durch alle Probanden in einer Sekunde. Formal ist die einfache Mittelwert-RTR-Zeitreihe M für alle Probanden damit definiert als (Biocca et al., 1994, S. 37-38):

$$M = m_1, m_2, \ldots, m_t \tag{5.1}$$

mit

$$m_t = \frac{\sum_{i=1}^{n} y_{it}}{n} = \frac{y_{1t} + y_{2t} + \cdots + y_{it}}{n} \tag{5.2}$$

wobei m_t der Mittelwert aller Probanden zum Zeitpunkt t, y_{it} der RTR-Wert des Probanden i zum Zeitpunkt t, und n die Gesamtzahl aller Probanden i ist.[35]

Wenn es das Ziel ist, einen Eindruck von der Bewertung der Kandidaten durch das gesamte Publikum unabhängig von den Voreinstellungen der Rezipienten zu gewinnen, wird häufig eine Gleichgewichtung der politischen Lager angestrebt.[36] Allerdings besteht in vielen Studien trotz einer Quotenvorgabe bei der Rekrutierung das Problem, dass die politischen Lager unterschiedlich stark vertreten sind. Da die Zugehörigkeit zu einem politischen Lager die unmittelbare Bewertung der Kandidaten deutlich beeinflusst (vgl. den Forschungsstand in Kapitel 3.4 und die Darstellung der RTR-Verläufe nach Lager in Abbildung 5.2), wird in einigen Studien eine auf Basis der Lagerzugehörigkeit gewichtete Mittelwert-RTR-Zeitreihe als Indikator für die Bewertung der Kandidaten durch das gesamte Publikum unabhängig von den Voreinstellungen berechnet (z.B. Bachl, 2013a; Maurer & Reinemann, 2003; Reinemann & Maurer, 2007b). Die Gewichte werden dabei so gewählt, dass die Anhängergruppen der beiden Kandidaten und die Ungebundenen in gleichem Ausmaß zur mittleren Bewertung der Kandidaten in einer Sekunde beitragen. In unserer Erhebung sind die Anhänger von Schmid (n = 89) deutlich stärker vertreten als die Anhänger von Mappus (n = 48) und die Unentschiedenen (n = 39). Um die gesamte Fallzahl nicht zu verändern, werden die Probanden in den Gruppen mit den Gewichten 0.66, 1.22 bzw. 1.50 versehen.

35 Zur Vereinfachung nehmen wir hier und in den folgenden Formeln an, dass die Zahl der Probanden n der Zahl der Messungen zu jedem Zeitpunkt t entspricht. Wenn es technisch bedingt zu vereinzelten fehlenden Werten in den individuellen Messreihen kommt, so verringert sich n für diese t entsprechend.

36 Für das Ziel einer Referenz auf die Grundgesamtheit aller Zuschauer des TV-Duells kann nach derselben Logik eine Gewichtung der politischen Lager nach der Zusammensetzung des TV-Duell-Publikums in repräsentativen Befragungen angestrebt werden.

Formal ist die gewichtete Mittelwert-RTR-Zeitreihe M_w für alle Probanden definiert als:

$$M_w = m_{w_1}, m_{w_2}, \ldots, m_{w_t} \tag{5.3}$$

mit

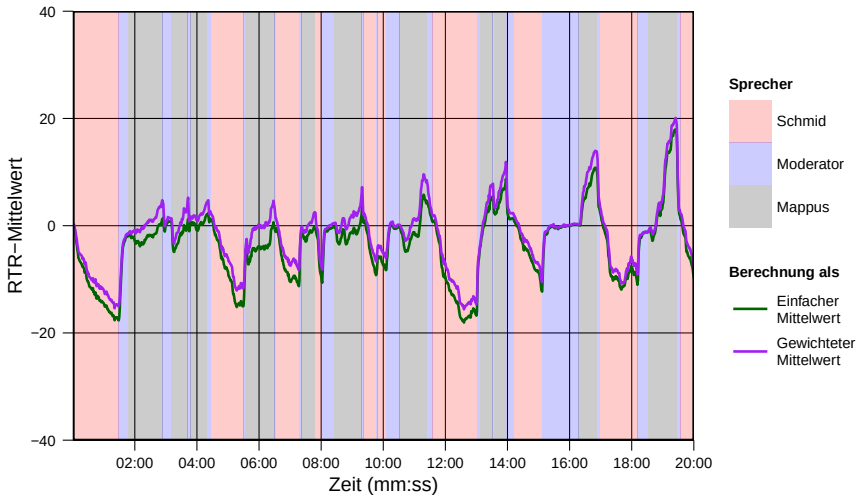
$$m_{w_t} = \frac{\sum_{i=1}^{n} w_i \cdot y_{it}}{\sum_{i=1}^{n} w_i} = \frac{w_1 \cdot y_{1t} + w_2 \cdot y_{2t} + \cdots + w_i \cdot y_{it}}{w_1 + w_2 + \cdots + w_i} \tag{5.4}$$

unter der Annahme, dass die Fallzahl n durch die Gewichtung nicht verändert werden soll, also $\sum_{i=1}^{n} w_i = n$ ist, ergibt sich

$$m_{w_t} = \frac{\sum_{i=1}^{n} w_i \cdot y_{it}}{n} = \frac{w_1 \cdot y_{1t} + w_2 \cdot y_{2t} + \cdots + w_i \cdot y_{it}}{n} \tag{5.5}$$

wobei m_{w_t} der gewichtete Mittelwert aller Probanden zum Zeitpunkt t, y_{it} der RTR-Wert des Probanden i zum Zeitpunkt t, w_i das Gewicht des Probanden i, und n die Gesamtzahl aller Probanden i ist.

Zur Interpretation und Ergebnisdarstellung wird die Mittelwert-Zeitreihe in der Regel grafisch über den Verlauf der Debatte dargestellt. Abbildung 5.1 zeigt beispielhaft die Bewertung von Mappus und Schmid in den ersten 20 Minuten der Debatte durch das gesamte Publikum, gemessen durch die einfache und die gewichtete Mittelwert-Zeitreihe der RTR-Messungen. Obwohl die beiden Zeitreihen im Zeitverlauf betrachtet kaum gravierende Unterschiede offenbaren, wird doch eine Verzerrung durch die Überrepräsentation der Anhänger von Schmid sichtbar. In einigen Passagen ist die Bewertung klar zugunsten von Schmid bzw. zu ungunsten von Mappus verschoben. Sind wir daran interessiert, wie das gesamte Publikum unter Kontrolle der Voreinstellungen die Kandidaten bewertet, so ist der gewichtete Mittelwert die bessere Wahl.



Anmerkungen
Mittlere Bewertung von Mappus und Schmid in den ersten 20 Minuten des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch n = 176 Rezipienten, ermittelt durch den einfachen und den gewichteten Mittelwert. Die Farbe im Hintergrund zeigt an, wer gerade das Wort hatte.
Diese und alle weiteren Grafiken wurden mit dem R Paket *ggplot2* (Wickham, 2009) erstellt.

Abbildung 5.1: Aggregierte RTR-Zeitreihen für das gesamte Publikum

Neben den Zeitreihen für das gesamte Publikum werden meist auch die Mittelwert-Zeitreihen für einzelne Gruppen verglichen.[37] Am häufigsten ist hier die Unterscheidung nach der Lagerzugehörigkeit der Rezipienten. Diese werden zum einen, wie in Abbildung 5.2 zu sehen, visuell verglichen. Zusätzlich wird in einigen Analysen die absolute Differenz zwischen den Mittelwerten der beiden Kandidatenlager als weitere Zeitreihe berechnet, um Passagen identifizieren zu können, die besonders zwischen den Lagern polarisieren. Die Differenz-Zeitreihe ist definiert als:

37 Maurer und Reinemann (2009, S. 8) nennen die Zusammenfassung von Gruppen nach in der Vorher-Befragung gemessenen Personeneigenschaften als ein Beispiel für die Analyse auf Individualniveau. Es ist zwar richtig, dass bei diesem Vorgehen individuell gemessene Informationen zur Bildung der Aggregate herangezogen werden. Da jedoch die Inner-Gruppen-Varianz bei der Aggregation verloren geht, handelt es sich nach der hier entwickelten Logik nicht um eine Analyse auf Individualniveau.

$$D = d_1, d_2, \ldots, d_t \tag{5.6}$$

mit

$$d_t = |m_{\text{Lager A}_t} - m_{\text{Lager B}_t}| \tag{5.7}$$

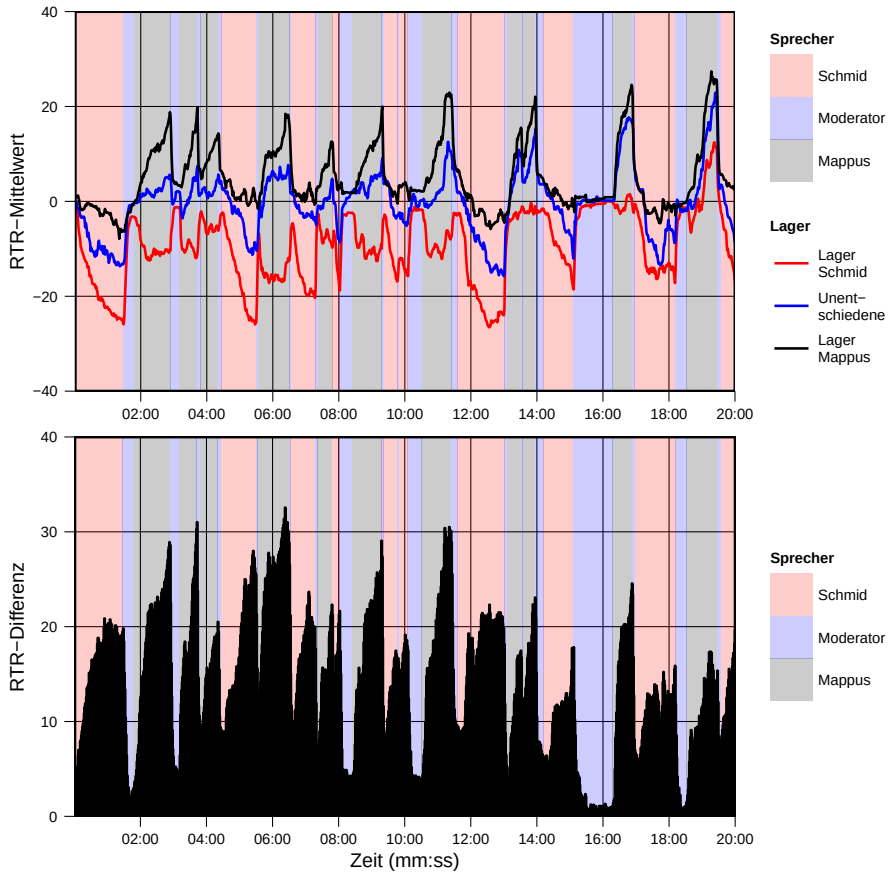
wobei $m_{\text{Lager A}_t}$ und $m_{\text{Lager B}_t}$ die Mittelwerte aller Probanden mit Zugehörigkeit zum Lager des Kandidaten A bzw. B zum Zeitpunkt t sind, wie sie sich nach Formel 5.2 ergeben.

Abbildung 5.2 zeigt im ersten Diagramm die mittlere Bewertung von Mappus und Schmid durch die Rezipienten in den drei Gruppen für die ersten 20 Minuten des TV-Duells. Im zweiten Diagramm ist die Differenz zwischen den Zeitreihen der Anhänger von Regierungs- und Oppositionsparteien dargestellt. Die teils sehr großen Differenzen zwischen den Bewertungen durch die Anhänger von Mappus und Schmid verdeutlichen erneut, dass die parteipolitische Prädisposition einen großen Einfluss auf die Bewertung der Kandidaten während des Duells hatte.

Zur weiteren Illustration der Datenreduktion, die durch die genannten Mittelwert-Aggregationen vorgenommen wird, sind in Tabelle 5.2 Auszüge der den Abbildungen zugrunde liegenden Zeitreihen dargestellt. Wie in Zeitreihen-Datensätzen üblich, sind die Zeiteinheiten (hier Sekunden) als Fälle in den Zeilen und die verschiedenen Aggregate als Variablen in den Spalten repräsentiert.

Der Vergleich mit dem Personendatensatz (Tabelle 5.1) zeigt, welche Informationen durch die Aggregation zu Zeitreihen gewonnen werden und welche verloren gehen. Neu hinzugekommen ist eine Variable, die angibt, zu welchem Zeitpunkt eine RTR-Bewertung im Aggregat gemessen wurde. Mithilfe dieser Information können nun Bezüge zum Inhalt der Debatte hergestellt werden – sei es bei der induktiven Interpretation in der Peak-Spike-Analyse oder durch eine systematische Zuordnung von Inhaltsanalyse-Daten in der Zeitreihen- oder Aussagenanalyse.

Nicht mehr im Datensatz enthalten sind nun die individuellen Probanden und ihre Eigenschaften. Lediglich die Information, welchem Lager die Probanden angehören, ist durch die Bildung der Gruppenzeitreihen nach Lagerzugehörigkeit erfasst. Alle anderen Unterschiede zwischen den Individuen – sowohl ihre unterschiedlichen unmittelbaren Kandidatenbewertungen zu einem Zeitpunkt als auch Unterschiede in anderen in der Befragung erhobenen



Anmerkungen
Oben: Mittlere Bewertung von Mappus und Schmid in den ersten 20 Minuten des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch die Anhänger Schmids (n = 89), Mappus' (n = 48) und Unentschiedene (n = 39).
Unten: Differenz zwischen der mittleren Bewertung der Kandidaten durch die Anhänger von Mappus und Schmid.
Die Farbe im Hintergrund zeigt an, wer gerade das Wort hatte.

Abbildung 5.2: Aggregierte RTR-Zeitreihen nach Lager

Tabelle 5.2: Auszug aus einem Zeitreihen-Datensatz

sec	time	mw_einf	mw_gew	mw_sch	mw_map	mw_unent	diff
1	00:00:01	0.2	0.3	−0.4	1.1	0.2	1.5
2	00:00:02	0.1	0.3	−0.4	1.2	0.1	1.6
3	00:00:03	−0.2	0.1	−1.0	1.2	0.0	2.1
4	00:00:04	−0.6	−0.4	−1.4	1.0	−0.9	2.4
5	00:00:05	−1.4	−1.0	−2.5	0.8	−1.5	3.3
6	00:00:06	−2.0	−1.4	−4.1	1.2	−1.5	5.3
7	00:00:07	−3.0	−2.2	−5.7	0.7	−1.6	6.4
8	00:00:08	−3.6	−2.6	−6.7	0.3	−1.6	7.0
9	00:00:09	−4.5	−3.6	−7.4	−0.8	−2.5	6.6
10	00:00:10	−5.1	−4.1	−8.1	−1.3	−2.9	6.9
...	...	...	...	...	...	...	...

Anmerkungen
Dargestellt sind die ersten zehn Sekunden des TV-Duells. Sekunde 1 markiert den Start von Schmids erster Aussage.
sec: Laufende Sekunde; time: Zeitcode (hh:mm:ss); mw_einf, mw_gew: einfacher, gewichteter Mittelwert aller Probanden; mw_sch, mw_map, mw_unent: Gruppenmittelwert der Anhänger von Schmid, Mappus, Unentschiedene; diff: Differenz zwischen den Gruppenmittelwerten Anhänger Schmid und Mappus.

Einstellungen wie z.B. die Bewertung der Kandidaten vor dem Duell – fehlen in den Zeitreihen. Nur die Aggregatsdifferenz der Bewertung durch die Anhänger beider Kandidaten bleibt durch eine weitere Zeitreihe erhalten. Auch eine Analyse, wie sich die Bewertungen eines individuellen Rezipienten im Zeitverlauf verändert haben, ist nicht mehr möglich. Schließlich kann dem Datensatz nicht mehr entnommen werden, auf wie vielen Probanden die mittleren unmittelbaren Bewertungen in einer Zeitreihe überhaupt basieren.

Die Erkenntnis, dass die Informationen über die Fallzahlen sowie die Varianzen zwischen und innerhalb der individuellen Rezipienten fehlen, ist in Anbetracht einer Aggregation über die Personen eigentlich offensichtlich. Für inferenzstatistische Verfahren, die Unterschiede zwischen Messungen oder Gruppen testen sollen, sind diese Informationen allerdings obligatorisch. Welche Konsequenzen ihr Fehlen für die verschiedenen Vorgehen bei der Datenanalyse hat, wird uns im Folgenden beschäftigen.

5.2 Verfahren zur induktiven Analyse: Peak-Spike-Analysen

Der erste Schritt der längsschnittlichen Analyse von RTR-Daten ist fast immer die visuelle Inspektion der nach den genannten Vorgehen generierten Verläufe (Biocca et al., 1994; J. Maier, 2013; Maurer & Reinemann, 2009). Ausgehend von den Stellen der RTR-Kurve, an denen sich besonders positive, negative oder polarisierte Bewertungen zeigen, wird über den Zeitcode ein Bezug zu der entsprechenden Passage des Stimulus hergestellt. Die Merkmale des Inhalts, die direkt vor diesen „Peaks“ oder „Spikes“ (Biocca et al., 1994, S. 38) liegen, werden als die Ursache für die auffällige Bewertung durch das Publikum interpretiert. Soll die Identifikation nicht alleine von den visuellen Eindrücken der Forscher abhängen, sondern systematisch und intersubjektiv nachvollziehbar erfolgen, so müssen zwei Entscheidungen getroffen werden: Nach welchen Kriterien werden die als bedeutsam zu interpretierenden Passagen ausgewählt? Wie wird bei der Interpretation und Systematisierung der so identifizierten Stimulusinhalte vorgegangen?

5.2.1 Vorgehen

Für die Auswahl der zu interpretierenden Peaks und Spikes finden sich in der Literatur zwei unterschiedliche Herangehensweisen. Zum einen werden auf Basis der beobachteten RTR-Zeitreihen empirisch *relative Grenzwerte* bestimmt. Die einfachste Variante dieses Vorgehens ist die Identifikation einer vorher festgelegten Anzahl von „Top-Aussagen“. So nennen beispielsweise McKinney et al. (2001) jeweils die zehn Aussagen, auf die die höchsten bzw. niedrigsten Mittelwerte in der RTR-Zeitreihe aller Zuschauer entfielen, die also am besten bzw. am schlechtesten bewertet wurden (ähnlich: McKinnon & Tedesco, 1993, 1999).

Ebenfalls auf der empirischen Verteilung der aggregierten RTR-Messungen beruht das von Biocca et al. (1994) vorgeschlagene Vorgehen: In einem ersten Schritt wird die RTR-Zeitreihe Z-transformiert. In einem zweiten Schritt werden die Werte der Z-Zeitreihe, die größer als 1.96 bzw. kleiner als -1.96 sind, als signifikant identifiziert. Das Vorgehen beruht darauf, dass unter Annahme einer ungefähren Normalverteilung aller aggregierten RTR-Werte etwa die fünf Prozent der Messungen erkannt werden, die sich am stärksten vom Mittelwert der gesamten Zeitreihe unterscheiden. Soll bei diesem Vorgehen die originale Skalierung der RTR-Daten beibehalten werden, so ist es auch möglich, alle Messungen als signifikante Abweichungen zu identifizieren, die ± 1.96 Standardabweichungen außerhalb des Mittelwerts der RTR-Zeitreihe liegen (Bachl, 2013a). Nach einem ähnlichen Prinzip identifizieren Reinemann und

Maurer (2005) Statements in einem TV-Duell, die die größte Zustimmung von allen Gruppen erhalten. Als Indikator für „unanimous support" (Reinemann & Maurer, 2005, S. 782) gilt eine gleichzeitige positive Abweichung vom jeweiligen Gruppenmittelwert um eine Standardabweichung in den RTR-Reihen der Anhänger beider Kandidaten und der Ungebundenen.

Der Vorteil dieser Verfahren liegt darin, dass die Grenzwerte, ab wann ein Peak als signifikant gilt und die ihm vorausgehenden Inhalte in der weiteren Interpretation verwendet werden, nicht willkürlich von den Forschern festgelegt wird. Damit ist das Verfahren in Grenzen auch robust gegen Verzerrungen, die sich aus einer ungleichen Verteilung relevanter Voreinstellungen in der Stichprobe ergeben können. Wenn die Stichprobe beispielsweise gegenüber der Botschaft eines Stimulus generell negativer eingestellt ist, so liegen die RTR-Zeitreihen konstant im negativeren Bereich der RTR-Skala. Die vorgestellten Verfahren können jedoch immer noch die Passagen identifizieren, in denen die Bewertung relativ gesehen am positivsten war. Identifiziert werden also immer die Passagen, die gemessen an der Bewertung des gesamten Stimulus durch dieses Publikum die größten Reaktionen verursachen. In dieser „Ergebnisgarantie" liegt aber gleichzeitig auch der Nachteil des Vorgehens: Per Definition werden diese Verfahren immer ein Ergebnis liefern. Wie wenig die Bewertung eines Stimulus im Zeitverlauf auch variiert haben mag, es gibt immer zehn Passagen, die relativ gesehen am besten oder am schlechtesten abgeschnitten haben. Ebenso wird es fast immer – sollte sich keine sehr ungewöhnliche Verteilung der aggregierten RTR-Messungen ergeben – ca. fünf Prozent Messungen geben, die im geforderten Umfang vom Gesamtmittelwert abweichen.

Neben den empirisch begründeten Auswahlverfahren werden in der Literatur verschiedene *absolute Grenzwerte* genannt, bei deren Über- bzw. Unterschreiten ein Peak als interpretierbar gewertet wird. Diese Werte lassen sich meistens über die Bedeutung der Skalenpunkte auf der RTR-Skala rechtfertigen. Dies kann beispielhaft anhand der Kategorisierung von Reinemann und Maurer (2007b) illustriert werden. Ausgehend von einer siebenstufigen RTR-Skala, die von 1 (größter Vorteil für Schröder) über den neutralen Mittelpunkt 4 bis zu 7 (größter Vorteil für Merkel) reicht, werden die folgenden Grenzwerte definiert: Für die Zeitreihen des gesamten Publikums und der Ungebundenen werden die Inhalte vor Abweichungen von ± 1 vom neutralen Skalenmittelpunkt interpretiert. Da die parteipolitisch gebundenen Zuschauer „ihren" Kandidaten durchgängig besser bewerten, wird ein Grenzwert von ± 2 Skalenpunkten um den Skalenmittelpunkt gewählt, um die Passagen zu finden, während derer die Zustimmung der eigenen Lager besonders groß war. Für die Identifikation der besonders polarisierenden Inhalte muss die absolute Differenz zwischen den Zeitreihen der beiden Kandidatenlager mindestens 2.5 Skalenpunkte betra-

gen. Alle diese Grenzen lassen sich vor dem Hintergrund der Spannweite der Skala gut begründen und sinnvoll interpretieren. Sie haben zudem den Vorteil, dass sie im Gegensatz zu den empirisch gezogenen Grenzwerten robuster gegen *false positives* sind. Wenn im Vorfeld begründet definiert wird, dass eine positive Bewertung durch das gesamte Publikum erst ab einem bestimmten Wert vorliegt und dieser Wert während des gesamten Stimulus nicht erreicht wird, so liegt eben das Ergebnis vor, dass es keine besonders positiven Reaktionen des Publikums gibt. Allerdings kann auch die *a priori* Festlegung der Grenzwerte eine Anpassung auf Basis der empirischen Befunde erfordern. So mussten Reinemann und Maurer (2007b, S. 77) die Grenze für die besonders polarisierenden Stellen von drei Skalenpunkten in der Analyse des Duells Schröder gegen Stoiber (Maurer & Reinemann, 2003, S. 106) auf zweieinhalb Skalenpunkte senken, da der höhere Grenzwert im 2005er Duell kein einziges Mal erreicht wurde. Hier zeigt sich, dass die Festlegung absoluter Grenzwerte immer mit einigen Freiheitsgraden seitens der Forscher verbunden ist.

Nachdem die relevanten Peaks in den RTR-Zeitreihen identifiziert wurden, muss ein Vorgehen für die Interpretation der Stimulusinhalte gewählt werden, die den Peaks zeitlich vorausgehen. In der Literatur findet sich in Hinblick auf die Systematik und Strukturiertheit des Vorgehens eine große Bandbreite. In einigen Analysen werden die Aussagen lediglich illustrativ aufgezählt und grob thematisch eingeordnet (z.B. Brettschneider & Bachl, 2012; Faas et al., 2009). In den meisten Studien findet eine theoriegeleitete Systematisierung statt, z.B. anhand des Issue-Ownership-Ansatzes oder mit Hilfe von Ansätzen aus der Persuasionsforschung (vgl. den Forschungsstand in Kapitel 3.4.1). Allerdings wird das konkrete Vorgehen bei der Systematisierung in diesen Arbeiten kaum thematisiert. Eine positive Ausnahme ist die Analyse von Reinemann und Maurer (2005) zur unmittelbaren Bewertung von Schröder und Stoiber in der zweiten Debatte vor der Bundestagswahl 2002. Die Autoren beschreiben ihr Vorgehen wie folgt:

> After identifying the two types of statements, we looked for their common features. [...] we wanted to be open for what we would encounter. Therefore, we did not develop a codebook in advance. Instead, in a first step of the analysis, each coauthor identified common features of the statements on his own. In a second step, the coauthors discussed their results until they arrived at a common interpretation of the most apparent and important common features of those statements (Reinemann & Maurer, 2005, S. 782-783).

In anderen Worten ausgedrückt haben die Autoren eine strukturierende qualitative Inhaltsanalyse (Mayring, 2010, S. 602) durchgeführt, mit dem Ziel,

gemeinsame Eigenschaften der Passagen zu finden und eine Typologie einender und polarisierender Aussagen zu entwickeln. Eine derartige systematische Vorgehensweise, die den explorativen Ansatz der induktiven Peak-und-Spike-Analysen konsequent berücksichtigt, ist gut dazu geeignet, auch über das untersuchte TV-Duell und den Wahlkampf hinaus neue Hypothesen zu generieren. Diese müssen dann natürlich in weiteren, deduktiv angelegten Studien auf ihre Gültigkeit hin geprüft werden.

Ein anderes Verfahren, die identifizierten Peaks systematisch einzuordnen, ist die Kombination mit einer standardisierten Inhaltsanalyse der gesamten Debatte. Dieses Verfahren haben wir bei der Analyse der Kandidatenbewertungen im baden-württembergischen TV-Duell eingesetzt (Bachl, 2013a; Bachl, Käfferlein & Spieker, 2013b). Die Kombination ermöglicht es, die Häufigkeit von *a priori* definierten Inhaltsmerkmalen im direkten Vorlauf der Peaks festzustellen, ohne vom Wissen beeinflusst zu werden, dass die vorliegenden Passagen bereits als bedeutsam identifiziert wurden. Dies verringert die Gefahr, dass einzelne Merkmale der Passagen angesichts des Wissens um ihre Bedeutung überinterpretiert werden. Während dieses Vorgehen einen sehr systematischen Überblick über die Charakteristika der Inhalte liefert, nach deren Rezeption ein Peak in einer RTR-Zeitreihe zu beobachten ist, so fehlt ihm doch der explorative Charakter, der eigentlich die Stärke einer induktiven Analyse ist. Gleichzeitig impliziert die Verwendung von Daten aus einer standardisierten Inhaltsanalyse, dass es sich um eine deduktive Analyse handeln könnte. Dies ist jedoch nicht der Fall, da die Publikumsreaktionen induktiv untersucht werden. Entsprechend vorsichtig ist bei diesem Vorgehen die Interpretation und Ergebnispräsentation zu handhaben, um keine falschen Erwartungen hinsichtlich der Reichweite der Befunde zu wecken.

5.2.2 Probleme

Das Verfahren der Peak-Spike-Analyse hat sich als Standardverfahren der induktiv ausgerichteten Studien zur Wahrnehmung des Debatteninhalts bewährt. Mit seiner Hilfe konnte ein wichtiger Beitrag zur detaillierten Beschreibung der Kandidatenbewertung in den TV-Duellen und zur Einordnung der TV-Duelle in den Kontext der Wahlkämpfe geleistet werden (vgl. den Forschungsstand in Kapitel 3.4.1). Darüber hinaus haben systematische Interpretationen auf Basis der Peak-Spike-Analysen dazu beigetragen, allgemeinere Hypothesen über die Wirkung bestimmter Merkmale von Kandidatenaussagen zu generieren. Trotz ihres vielfältigen Einsatzes und Nutzens ist die Peak-Spike-Analyse von RTR-Zeitreihen nicht unproblematisch. Kritisch zu beurteilen sind insbesondere zwei Punkte:

- *Die theoretisch angemessene Interpretation der Befunde*: Kann mit einer Peak-Spike-Analyse von über Personen aggregierten RTR-Zeitreihen auf individuelle Informationsverarbeitungsprozesse geschlossen werden, oder beschränken sich die zulässigen Interpretationen auf die Aufdeckung von Aggregatsphänomenen? Wäre zweites der Fall, so würde ein zentraler Nutzen der aufwändigen rezeptionsbegleitenden Messung vergeben werden (Biocca et al., 1994; Fahr, 2008).
- *Die Generalisierbarkeit der Befunde*: Inwiefern lassen sich mit Peak-Spike-Analysen verallgemeinerbare Aussagen über die Wirkung bestimmter Merkmale des Stimulusinhalts generieren bzw. Hypothesen darüber testen? Mit welcher Sicherheit lässt sich von den Befunden in der Stichprobe zumindest auf die Wahrnehmung derselben Debatte durch andere Rezipienten aus der Grundgesamtheit schließen? Die erste Frage ist vor allem in Hinblick auf Theorieentwicklung und Hypothesentests zur Wirkung bestimmter Debatteninhalte relevant, die zweite Frage für die Einordnung der Debattenwahrnehmung in den Kontext des Wahlkampfs und die Indikatorfunktion der Ergebnisse im Sinne der Analogie Debatten als Miniaturwahlkämpfe (vgl. zu den unterschiedlichen Zielen der TV-Duell-Studien ausführlich Kapitel 2).

Konsequenzen der induktiven Analyselogik für die Generalisierbarkeit der Befunde

Für Zweifel am korrekten Umgang mit den Befunden aus Peak-Spike-Analysen gibt es zwei wesentliche Argumentationslinien, die an ihrem analytischen Vorgehen ansetzen. Bereits häufiger problematisiert wurden die *Konsequenzen der induktiven Analyselogik*. Wie im letzten Abschnitt bereits ausführlich erläutert, werden die zu interpretierenden Inhalte des Stimulus auf Basis der Publikumsreaktionen identifiziert. Es wird also zuerst nach dem Auftreten besonderer Wirkungen gesucht, um sie in einem zweiten Schritt anhand von Merkmalen des Stimulus zu erklären. Nagel et al. (2012, S. 839) stellen hierzu fest: „This approach may be misleading because those message elements may appear to the same extent during ordinary parts of the debate not under examination“ (ähnlich auch Nagel, 2012; Spieker, 2011; Strömbäck et al., 2009). Diese Argumentation weckt Zweifel an der Generalisierbarkeit der Befunde.

Zwar können deskriptive Vergleiche der Charakteristika von als bedeutsam identifizierten Debatteninhalten mit Inhaltsanalysen der gesamten Debatte (z.B. Bachl, Käfferlein & Spieker, 2013b; Maurer, 2007) die Interpretation der Peaks argumentativ absichern, ein statistischer Vergleich ist allerdings komplexer

und stößt an anderer Stelle an Grenzen (vgl. dazu die Ausführungen zu den Zeitreihen- und Aussagenanalysen). In eine ähnliche Richtung geht der auf der Logik der Grounded Theory aufbauende Vorschlag von Rust (1985). Anhand der extremsten Peaks sollen theoretische Annahmen über die Wirkung bestimmter Stimulusinhalte entwickelt werden. Diese Annahmen sollen dann im nächsten Schritt für weitere Passagen des Stimulus überprüft werden. Für die Untersuchung kürzerer Stimuli – der Autor bezieht sich auf die Analyse von TV-Spots – scheint dieses Vorgehen gut geeignet. Ob es sich auf die Analyse längerer RTR-Zeitreihen, die wie in TV-Duell-Studien viele Peaks aufweisen, übertragen lässt, ist allerdings fraglich. Schließlich empfehlen Biocca et al. (1994) in ihrer systematischen Darstellung des Vorgehens bei einer Peak-Spike-Analyse, dass *vor* der Analyse Hypothesen formuliert werden sollten, an welchen Stellen des Stimulus Peaks auftreten werden. Der Test dieser Hypothesen solle eine Überinterpretation der induktiv identifizierten Peaks verhindern, da die Forscher ihren Interpretationsrahmen bereits bei der Formulierung der Hypothesen explizit machen müssen. Auch dieses Vorgehen hat jedoch die Schwäche, dass allgemeinere Hypothesen wie z.B. „Gemeinplätze werden überdurchschnittlich positiv bewertet“ nicht direkt getestet werden, da der Gegenvergleich der Peaks mit den übrigen Stellen ausbleibt.

Die Anmerkungen zu den Einschränkungen einer explorativ vorgehenden Analyse sind gerechtfertigt und werden vor allem von den Autoren ins Feld geführt, die Befunde deduktiv angelegter Studien zur Wirkung bestimmter Debatteninhalte präsentierten. Da diese Kritik grundsätzlich die Logik induktiver Analyseverfahren betrifft, lässt sie sich auch nicht durch die genannten Modifikationen oder Ergänzungen der Peak-Spike-Analyse vollständig entkräften. Hypothesen über die Wirkung bestimmter Debatteninhalte können nur mit deduktiv ausgerichteten Analysen getestet werden. Die Ergebnisse induktiver Analysen können vor dem Hintergrund dieser Überlegungen nur dazu dienen, die Bewertung der Stimuli im Zeitverlauf zu beschreiben und daraus mögliche Hypothesen über die Wirkung bestimmter Stimulusmerkmale abzuleiten.

Konsequenzen der Zeitreihen-Aggregation für die theoretisch angemessene Interpretation der Befunde

Kaum thematisiert wurden bisher dagegen die Einschränkungen für die theoretisch angemessene Interpretation und die Generalisierbarkeit der Befunde, die sich aus den bereits gezeigten Informationsverlusten infolge der Erstellung der RTR-Zeitreihen ergeben. Mit diesen Konsequenzen werden wir uns im Folgenden beschäftigen. Wir beginnen mit einer Betrachtung der individuellen RTR-Messungen im Vorlauf der in der Aggregatanalyse als bedeutsame Peaks

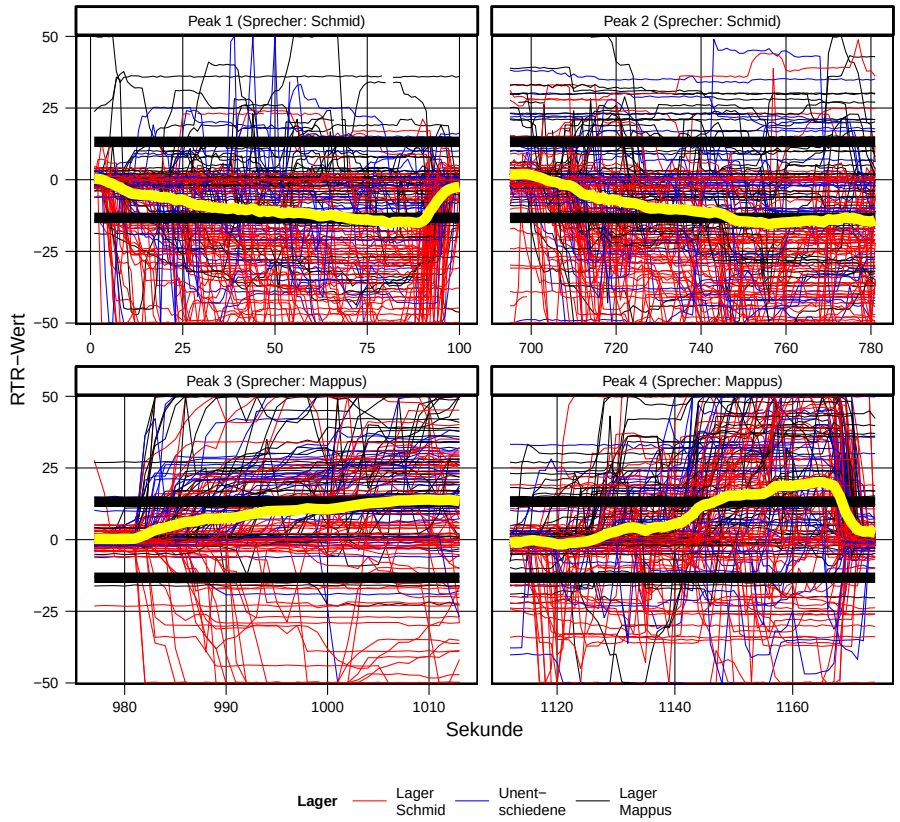
identifizierten Passagen. Damit wollen wir zur Klärung der Frage beitragen, ob es sich bei diesen Befunden um Individual- oder Aggregatphänomene handelt.

Als Beispiel ziehen wir wieder die Daten aus der TV-Duell-Studie vor der baden-württembergischen Landtagswahl 2011 heran. Um einen Eindruck davon zu gewinnen, wie die individuellen Probanden Schmid und Mappus im Umfeld der Aussagen bewerteten, die wir in einer Peak-Spike-Analyse als bedeutsam identifizieren (Bachl, 2013a), stellen wir die individuellen und die aggregierten RTR-Zeitreihen visuell vergleichend gegenüber. Dazu sind auf den folgenden Seiten die jeweils ersten Peaks in der Bewertung durch das gesamte Publikum (Abbildung 5.3), die Anhänger der Regierungsparteien (Abbildung 5.4), die Anhänger der Oppositionsparteien (Abbildung 5.5), die Unentschiedenen (Abbildung 5.6) und für die Differenz-Zeitreihe der Kandidatenlager (Abbildung 5.3) dargestellt.[38] Die Facetten der Diagramme zeigen jeweils einen Turn eines Kandidaten, in dem ein Peak in der aggregierten RTR-Kurve identifiziert wurde. Unter Turn verstehen wir im Folgenden den Zeitraum, ab dem ein Kandidat das Wort ergreift bis zu dem Zeitpunkt, an dem der Sprecher wechselt.[39] Da vor den meisten Turns der Kandidaten eine Frage der Moderatoren liegt, starten die meisten RTR-Kurven nahe des neutralen Mittelpunkts der RTR-Skala.

Zwei Hinweise zur Interpretation der Diagramme: Um die Übersichtlichkeit der gesamten Abbildungen zu steigern, haben wir für die Zeitachsen der Facetten unterschiedliche Skalierungen gewählt. Dies soll es erleichtern, in jeder Facette einen bestmöglichen Gesamteindruck von der Varianz der individuellen RTR-Bewertungen zu vermitteln. Vorsicht ist dadurch aber bei der Interpretation der zeitlichen Dynamik der Verläufe im Vergleich der Turns geboten, da visuelle Unterschiede hier auch durch eine anders skalierte x-Achse entstehen können. Weiter sind die Datenpunkte der individuellen Verläufe in der Horizontalen mit kleinen, aus einer uniformen Verteilung gezogenen Zufallszahlen verrechnet, um Überlappungen zu verringern und so einen möglichst vollständigen Überblick über die individuellen Verläufe zu ermöglichen. Durch dieses sogenannte „Jittering“ (Chambers, Cleveland, Kleiner & Tukey, 1983, S. 106-107) wird der visuell vermittelte Eindruck der intraindividuellen Varianz der einzelnen Verläufe etwas überschätzt. Dies nehmen wir in Kauf, um einen besseren Eindruck von der interindividuellen Varianz zu vermitteln.

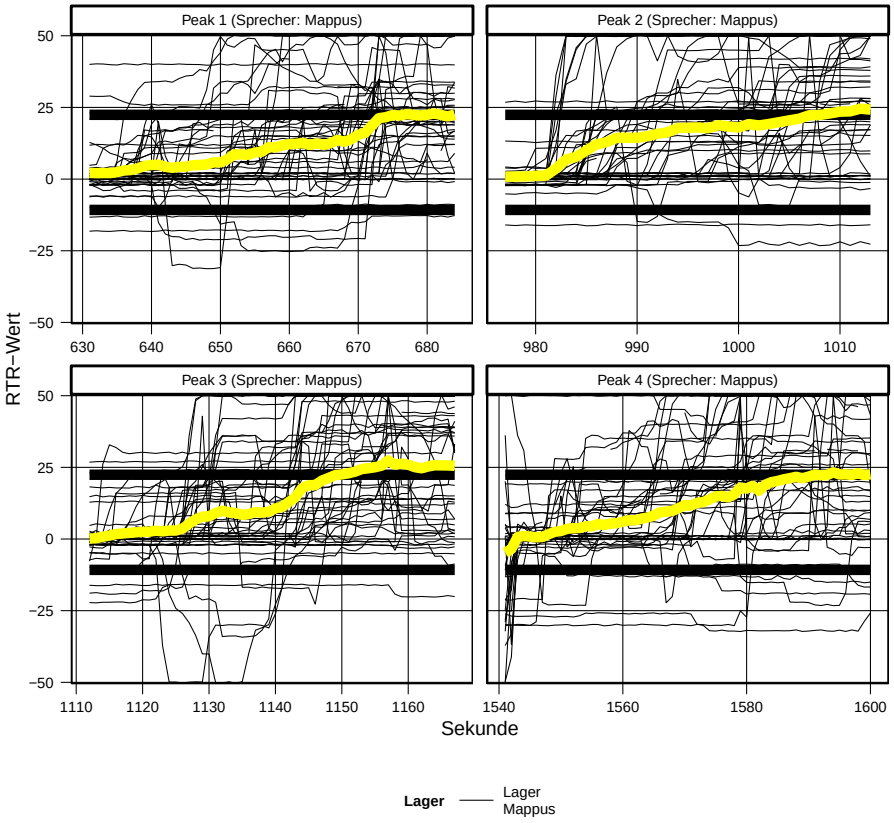
38 Minimale Differenzen zu den in Bachl (2013a) berichteten Werten können auftreten, da die Vorgehen bei der Datensatzbereinigung leicht voneinander abweichen.

39 Vgl. zu Definition und Bedeutung der Turns auch die Ausführungen auf S. 173.



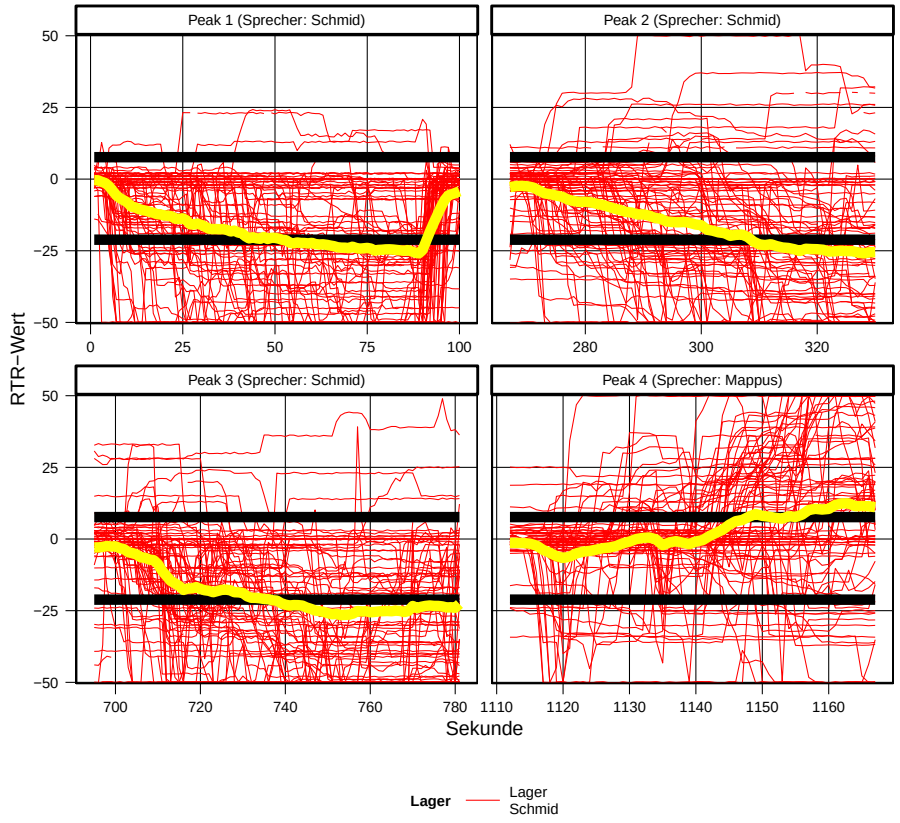
Anmerkungen
(Mittlere) Bewertung von Mappus und Schmid auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Individuelle RTR-Zeitreihen (farbig markiert nach Lagerzugehörigkeit), gewichtete Mittelwert-Zeitreihe aller Probanden (gelb) und Signifikanzgrenzen bei $M = -0.1 \pm 1.96 * 6.8SD$ für die gesamte Mittelwert-Zeitreihe während des Duells (schwarze horizontale Balken). Die individuellen Verläufe sind in der Horizontalen mit kleinen Zufallswerten verrechnet, um Überlappungen zu verringern.

Abbildung 5.3: Aggregierte und individuelle RTR-Zeitreihen für das gesamte Publikum



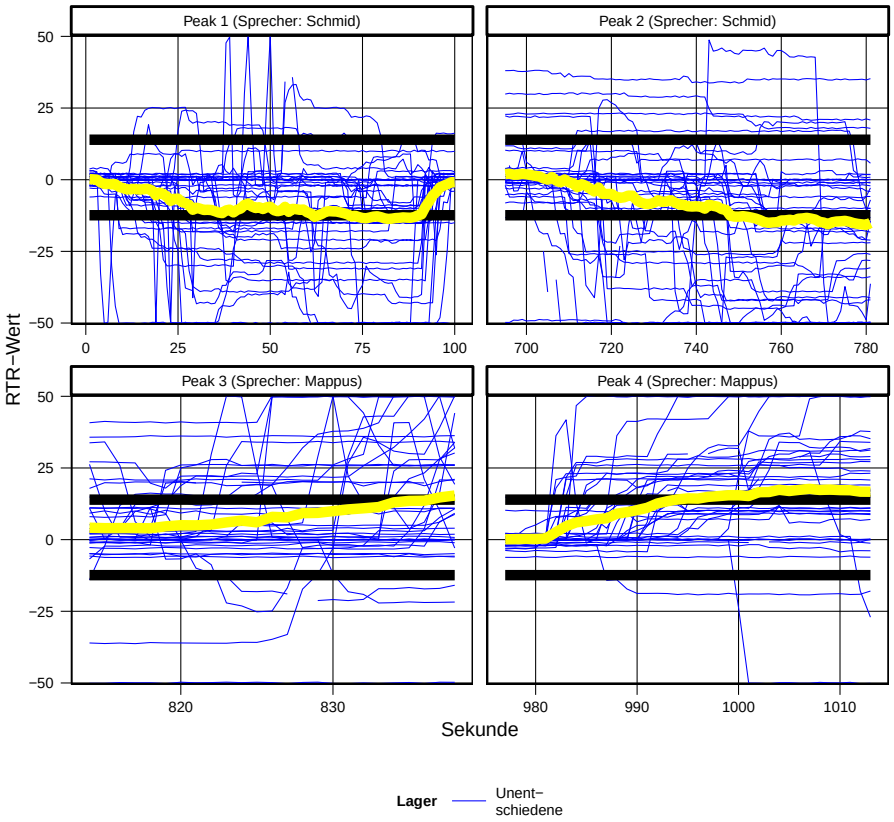
Anmerkungen
(Mittlere) Bewertung von Mappus und Schmid auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Individuelle RTR-Zeitreihen (farbig markiert nach Lagerzugehörigkeit), Mittelwert-Zeitreihe der Anhänger von Mappus (gelb) und Signifikanzgrenzen bei $M = 5.8 \pm 1.96 * 8.5SD$ für die gesamte Mittelwert-Zeitreihe während des Duells (schwarze horizontale Balken). Die individuellen Verläufe sind in der Horizontalen mit kleinen Zufallswerten verrechnet, um Überlappungen zu verringern.

Abbildung 5.4: Aggregierte und individuelle RTR-Zeitreihen für die Anhänger von Mappus



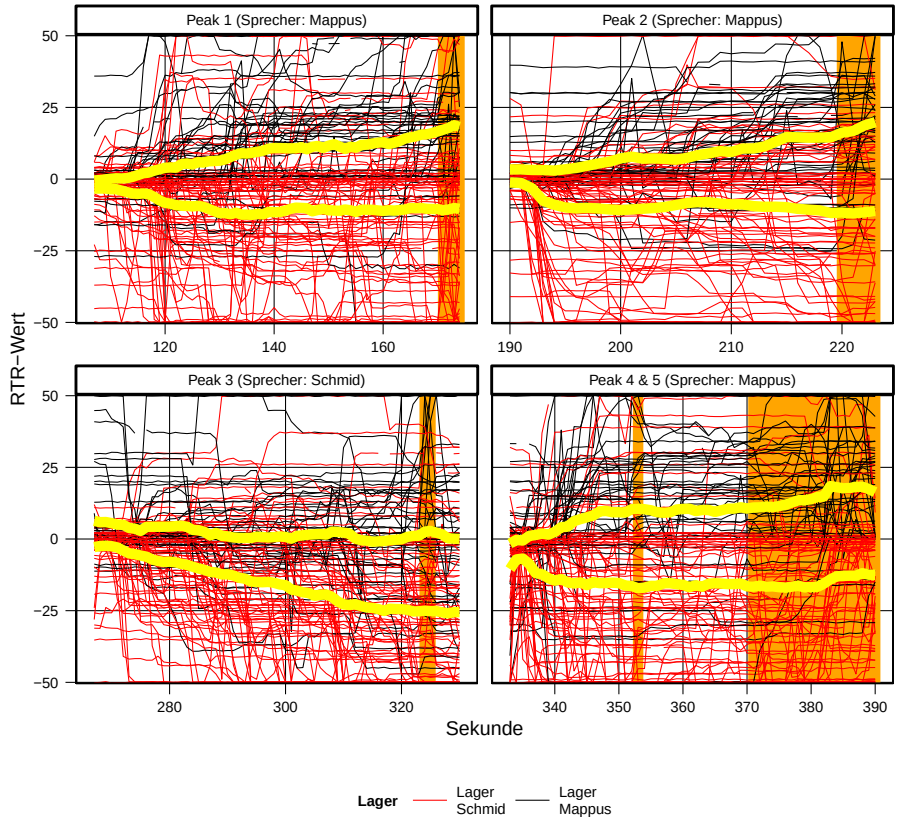
Anmerkungen
(Mittlere) Bewertung von Mappus und Schmid auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Individuelle RTR-Zeitreihen (farbig markiert nach Lagerzugehörigkeit), Mittelwert-Zeitreihe der Anhänger von Schmid (gelb) und Signifikanzgrenzen bei $M = -6.8 \pm 1.96 * 7.3SD$ für die gesamte Mittelwert-Zeitreihe während des Duells (schwarze horizontale Balken). Die individuellen Verläufe sind in der Horizontalen mit kleinen Zufallswerten verrechnet, um Überlappungen zu verringern.

Abbildung 5.5: Aggregierte und individuelle RTR-Zeitreihen für die Anhänger von Schmid



Anmerkungen
(Mittlere) Bewertung von Mappus und Schmid auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Individuelle RTR-Zeitreihen (farbig markiert nach Lagerzugehörigkeit), Mittelwert-Zeitreihe der Unentschiedenen (gelb) und Signifikanzgrenzen bei $M = 0.8 \pm 1.96 * 6.7SD$ für die gesamte Mittelwert-Zeitreihe während des Duells (schwarze horizontale Balken). Die individuellen Verläufe sind in der Horizontalen mit kleinen Zufallswerten verrechnet, um Überlappungen zu verringern.

Abbildung 5.6: Aggregierte und individuelle RTR-Zeitreihen für die Unentschiedenen



Anmerkungen
(Mittlere) Bewertung von Mappus und Schmid auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Individuelle RTR-Zeitreihen (farbig markiert nach Lagerzugehörigkeit), Mittelwert-Zeitreihe der Anhänger von Mappus (gelb, oben) und Schmid (gelb, unten). Signifikanzgrenzen bei $M = 12.6 + 1.96 * 7.6SD$ für die gesamte Differenz-Zeitreihe der Anhänger beider Lager während des Duells (orange vertikale Balken). Die individuellen Verläufe sind in der Horizontalen mit kleinen Zufallswerten verrechnet, um Überlappungen zu verringern.

Abbildung 5.7: Aggregierte und individuelle RTR-Zeitreihen für die Anhänger von Mappus und Schmid

Die individuellen Verläufe sind zu einem großen Teil gekennzeichnet durch einzelne sprunghafte Wechsel auf neue Bewertungen, die dann meist längere Zeit beibehalten werden. Weiter fällt auf, dass ein substanzieller Anteil der Teilnehmer die Bewertung während einzelner hier dargestellter Turns überhaupt nicht verändert, sondern den Drehregler auf dem neutralen Mittelpunkt der Skala lässt. Bereits eine oberflächliche Betrachtung der Abbildungen macht klar, dass eine beträchtliche Varianz zwischen den RTR-Bewertungen durch die individuellen Probanden besteht.

Wie auf Basis der Analysen zu den nach Lagern getrennten Zeitreihen und ihrer Differenz-Zeitreihe zu erwarten, ist die Divergenz um den Mittelwert des gesamten Publikums besonders groß. Welche Fehlschlüsse daraus entstehen können, lässt sich besonders gut am vierten Peak in Abbildung 5.3 illustrieren. Diese Aussage, in der Mappus die Fort- und Weiterbildung der Menschen in Baden-Württemberg dem Anwerben von ausländischen Fachkräften als Maßnahme zur Behebung des Fachkräftemangels den Vorzug gab, haben wir auf Basis unserer Peak-Spike-Analyse als seine beim gesamten Publikum erfolgreichste Aussage identifiziert (Bachl, 2013a). Dies ist ohne Frage insofern richtig, als dass an dieser Stelle der Debatte der gewichtete Mittelwert aller Echtzeiturteile den für Mappus positivsten Wert erreichte. Aber können wir daraus schließen, dass er mit dieser Aussage die einhellige Zustimmung des Publikums erhielt? Erst die individuellen Verläufe offenbaren, dass dies keineswegs der Fall war. Im Gegenteil: Das Statement wirkte zwar äußert aktivierend auf einen großen Teil der Zuschauer, und es brachte viele Zuschauer dazu, eine positive Bewertung von Mappus abzugeben. Allerdings beurteilten ein bedeutsamer Teil der Anhänger Schmids und auch einige Unentschiedene die Aussage klar negativ. Der Erfolg im Aggregat des gesamten Publikums erklärt sich also nicht durch eine hohe Zustimmung über alle Rezipienten hinweg, sondern dadurch, dass Mappus deutlich mehr Personen zu einer positiven Bewertung aktivieren konnte, als er mit ihr verschreckte. Abstrakt formuliert zeigt das Beispiel, dass es sich also keineswegs um einen Individualeffekt handelt. Dies wäre der Fall, wenn alle Zuschauer diesem Statement stärker als anderen Aussagen zugestimmt hätten. Auch wenn wir hier keinen intraindividuellen Vergleich mit anderen Aussagen durchgeführt haben, so ist es doch sehr wahrscheinlich, dass die Zuschauer, die Mappus hier (sehr) negativ bewerteten, ihn an anderer Stelle besser bewerteten. Es handelt sich um einen Aggregatseffekt, der sich darin zeigt, dass relativ viele Personen eine positive und relativ wenige Personen eine negative Wertung abgegeben haben.

Auch an anderen Stellen können wir innerhalb der Lager einige Varianz in den individuellen Zeitreihen feststellen, auch wenn sie naturgemäß weniger groß ist als innerhalb des gesamten Publikums. Detailliert zeigen dies die

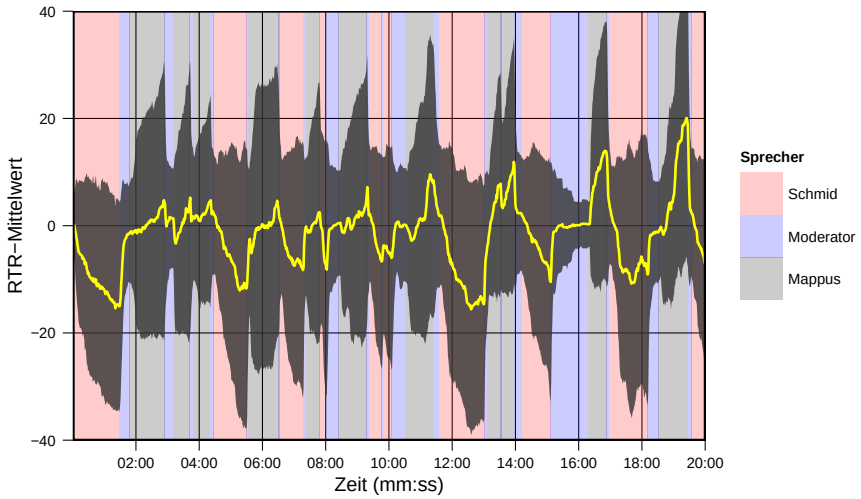
Abbildungen 5.4 bis 5.6. Es sei noch einmal daran erinnert, dass wir hier lediglich Turns betrachten, die von den jeweiligen Gruppen im Aggregat am besten bewertet wurden, wozu relativ einheitliche Urteile aller Probanden in der Gruppe notwendig sind. Selbst in diesen Aussagen sehen wir einen gewissen Anteil abweichender RTR-Bewertungen und eine relativ große Zahl neutraler Urteile. Letzteres gilt auch für die in Abbildung 5.7 dargestellten Turns, in denen Peaks der Differenz-Reihe liegen, die also von den Anhängern der beiden Kandidaten (an einigen Stellen) im Mittel besonders unterschiedlich bewertet wurden. Es ist zudem festzustellen, dass selbst im sehr polarisierenden Turn, in dem Peak 4 und 5 liegen, eine nicht zu vernachlässigende Zahl an gegenläufigen Bewertungen innerhalb der Lager vorliegt.

In Anbetracht der Relevanz, welche die Varianz der individuellen Messungen um die aggregierten Zeitreihen, die alleine die Datengrundlage für die Peak-Spike-Analysen sind, für die theoretische Einordnung der Befunde hat, überrascht es, dass dies in keiner der TV-Duell-Studien thematisiert wird. Selbstverständlich wird in diesen Analysen nicht behauptet, dass die Mittelwert-Zeitreihen sich direkt in einen Individualeffekt übersetzen lassen. Ein weitergehender Bezug auf die Varianz der individuellen Messungen bei der Interpretation – und sei es auch nur wie hier dargestellt visueller Art – bleibt völlig aus. Dies ist umso bemerkenswerter, als dass die Anleitung von Biocca et al. (1994) zur Peak-Spike-Analyse, die in fast jeder Publikation zitiert wird, explizit auf die Notwendigkeit der Berücksichtigung der Streuung um die Mittelwert-aggregierten Zeitreihen hinweist. Zwar findet sich auch hier kein Hinweis auf die Analyse der individuellen RTR-Verläufe, jedoch wird die Betrachtung der Standardabweichungen um die Mittelwerte der Zeitreihen herum dringlich empfohlen:[40]

> In any case, the standard deviation series *should generally be plotted alongside the mean response series* [Hervorhebung M.B.], because the mean series can sometimes disguise radical shifts of signaled mental states in the audience. For example, the mean may not change from t_1 to t_2, but during that interval the audience may have moved from a convergence of response to radical polarization (Biocca et al., 1994, S. 44).

Abbildung 5.8 veranschaulicht anhand der bereits in Abbildung 5.1 eingeführten Zeitreihe der mittleren Bewertung von Mappus und Schmid durch das

40 Eine Präsentation von Zeitreihen der Standardabweichung findet sich auch außerhalb der TV-Duell-Studien nur selten. Zu den Ausnahmen gehören u.a. Kercher (2013) und Lambooij et al. (2011).



Anmerkungen
Mittlere Bewertung von Mappus und Schmid in den ersten 20 Minuten des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch n = 176 Rezipienten. Gewichtete Mittelwert-Zeitreihe M_w (gelb) und $\pm 1 SD_w$ um den Mittelwert (graue Fläche). Die Farbe im Hintergrund zeigt an, wer gerade das Wort hatte.

Abbildung 5.8: Aggregierte RTR-Zeitreihe mit Standardabweichungen für das gesamte Publikum

gesamte Publikum, wie die Mittelwert-Zeitreihe mit der Zeitreihe der Standardabweichung zusammenhängt. Dazu ist das Intervall von $\pm 1 SD_w$ um die Mittelwert-Zeitreihe eingezeichnet.[41] Dieses Intervall ist ein typisches Maß, um einen Eindruck von der Streuung um das arithmetische Mittel zu gewinnen, da unter der Annahme einer Normalverteilung etwa zwei Drittel der Messungen in diesem Bereich liegen.

Das Intervall der Standardabweichung verdeutlicht noch einmal, dass die Streuung um die mittlere Bewertung der Kandidaten durch das gesamte Publi-

41 Die Berechnung der Varianz bzw. Standardabweichung um einen gewichteten Mittelwert ist nicht trivial. Für die vorliegende Darstellung wird die Standardabweichung als Wurzel der mit der R Funktion *wtd.var* (Harrell Jr, 2012) geschätzten gewichteten Varianz ermittelt. Es wird eine Methode nach Cochran verwendet, die sich im Vergleich verschiedener Methoden bewährt hat (Gatz & Smith, 1995a).

kum enorm ist. Relativ homogene Messungen liegen nur zu den Zeitpunkten vor, in denen die Moderatoren sprachen. Dies ist eine Konsequenz des bekannten Verhaltens der meisten Probanden, während dieser Passagen zum neutralen Skalenmittelpunkt zurückzukehren. Ergriff einer der Kandidaten das Wort, so steigt die Streuung an. Dabei lassen sich zwei Muster ausmachen: Zum einen ist z.B. während des ersten Turns von Mappus das von Biocca et al. (1994) beschriebene Phänomen zu erkennen: Die Mittelwert-Zeitreihe verändert sich kaum, die Streuung steigt jedoch an. Positive und negative Bewertungen gleichen sich hier im Mittel aus. Zum anderen ist auch in Passagen, in denen der Gesamtmittelwert sich beträchtlich verändert (z.B. während des ersten Turns von Schmid) ein deutlicher Anstieg der Varianz zu erkennen. Die Analyse der individuellen Verläufe (vgl. Abbildung 5.3, Peak 1) offenbart, dass der Anstieg vor allem durch die Differenz zwischen den Zuschauern, die bei ihrer neutralen Bewertung bleiben, und denjenigen, die der Aussage (stark) zustimmen, entsteht.

Die Zeitreihe der Standardabweichungen ist natürlich ebenfalls ein Aggregatsmaß, mit dessen Hilfe keine Aussagen über die Streuung der individuellen RTR-Verläufe, die zu ihrem Anstieg führt, getroffen werden können. Die vergleichende Betrachtung der Mittelwert- und der Standardabweichung-Reihen kann jedoch zumindest dabei helfen, einen Eindruck davon zu erhalten, wie typisch der mittlere RTR-Verlauf ist. In Anbetracht der visualisierten Befunde können wir uns der Forderung von Biocca et al. (1994) anschließen, bei der Präsentation der Befunde zumindest die Streuung der Messungen zu jedem Zeitpunkt zu berichten.

Zwischenfazit: Konsequenzen der Zeitreihen-Aggregation für die theoretisch angemessene Interpretation der Befunde In Bezug auf die Frage, ob sich die RTR-Mittelwert-Zeitreihen und die in ihnen identifizierten Peaks nur auf Aggregats- oder auch auf Individualebene interpretieren lassen, fallen die hier präsentierten Analysen eindeutig aus. Die aus vielen Publikationen bekannten RTR-Verläufe, die während der Aussagen der Kandidaten kontinuierlich fallen oder steigen, bis sie womöglich einen substantiellen Peak erreichen, sind eindeutig Aggregatphänomene, die auf Individualebene in dieser Form nicht existieren. Eine im Verlauf einer Kandidatenaussage ansteigende aggregierte RTR-Zeitreihe bedeutet nicht, dass der Eindruck, den die individuellen Rezipienten von dem Kandidaten haben, kontinuierlich besser wird. Vielmehr steigt die aggregierte Kurve an, wenn sich im Zeitverlauf mehr und mehr individuelle Probanden dafür entscheiden, ihre Bewertung von „neutral“ auf einen positiven Wert zu verändern, oder wenn andere Probanden von einer

zuvor negativen Wertung abrücken. Auch in der Perspektive der Messwiederholung ist es nicht klar, ob Ausschläge in der mittleren Bewertung immer auf dieselben Probanden zurückgehen, oder ob jeweils unterschiedliche Probanden (in derselben Gruppe) dafür verantwortlich sind.

Dementsprechend sollten auch die Interpretationen der mittleren RTR-Verläufe und deren Peaks auf Aggregatebene erfolgen. Peaks sind demnach Passagen der Debatte, in dem die Zustimmung eines großen Teils der zusammengefassten Probanden groß war und gleichzeitig vergleichsweise wenige Probanden Ablehnung äußerten. Solche Interpretationen lassen sich gut mit einer Deskription der Kandidatenbewertung durch die politischen Lager vereinbaren. So könnte z.B. im Kontext des Issue-Ownership-Ansatzes interpretiert werden: Große Teile der eigenen Anhänger haben die Aussagen zu bestimmten Themen positiv bewertet, während sie sich bei anderen Themen weniger einig waren oder größtenteils neutral urteilten. Für eine solche Interpretation ist die individuelle Zusammensetzung der RTR-Messungen zunächst einmal unerheblich, da sie darauf abzielt, zu erklären, warum es dem Kandidaten bei einem Thema gelang, die Gruppe seiner Anhänger zu mobilisieren. Das Ergebnis könnten wir dann z.B. auch als einen Indikator dafür heranziehen, dass dieses Thema auch im Wahlkampf ein „Gewinnerthema" für den Kandidaten und seine Partei war.

Weniger gut passen dagegen Interpretationsversuche auf Basis von auf das Individuum bezogenen Erklärungsansätzen zu diesen Aggregatbefunden. Dafür gibt es zwei Gründe, die sich am Beispiel der Analyse nach politischem Interesse gruppierter RTR-Zeitreihen (Reinemann & Maurer, 2010) erläutern lassen: *Erstens* stellt sich die Frage, welche inhaltliche Bedeutung es hat, wenn sich die Gruppen der Interessierten und der weniger Interessierten bei der Bewertung verschiedener Debatteninhalte unterscheiden. Während die relative Zustimmung zu der Aussage eines Kandidaten innerhalb eines politischen Lagers eine inhaltlich plausibel zu interpretierende Bezugsgröße hat („Große Teile der Anhänger Schröders stimmten seiner Aussage zum Irakkrieg zu"), bleibt die inhaltliche Tragweite des Ergebnisses „Große Teile der politisch Interessierten stimmten Schröders Aussage zum Irakkrieg zu" unklar.[42] Aus dem ersten Ergebnis könnte z.B. abgeleitet werden, dass Schröder das Thema Irakkrieg dazu nutzen kann, viele seiner Anhänger zu mobilisieren. Diese Interpretation bezieht sich nicht auf einzelne seiner Anhänger, da sie keine Erklärung dafür zu finden versucht, warum individuelle Anhänger Schröders hier zustimmen. Für die „Gruppe" der politisch Interessierten kann eine ähnliche

42 Dies sind zur Illustration erdachte Beispiele. Die Untersuchung von Reinemann und Maurer (2010) konnte keine systematischen Unterschiede zwischen den RTR-Zeitreihen der beiden Gruppen feststellen.

Aggregatinterpretation nicht erfolgen, da das politische Interesse ein Merkmal der Individuen ist.[43] Damit kann *zweitens* eine Interpretation eines solchen Ergebnisses nur mit Rückbezug auf die individuelle Informationsverarbeitung erfolgen, wie sie z.B. im Elaboration-Likelihood-Modell beschrieben wird. Der Schluss von Gruppendaten auf solche Prozesse innerhalb einzelner Individuen in dieser Gruppe ist aber nicht möglich, ohne die Gefahr eines ökologischen Fehlschlusses einzugehen (Yanovitzky & Greene, 2009). Denn es wird in dieser Analyse weder untersucht, ob das Involvement tatsächlich bei einzelnen Rezipienten zu einer unterschiedlichen Bewertung bestimmter Debatteninhalte führt, noch, ob dieser Effekt in der Messwiederholung konsistent bei denselben Individuen auftritt.

Konsequenzen der Zeitreihen-Aggregation für die Generalisierbarkeit der Befunde

Die mit der Erstellung der RTR-Zeitreihen verbundenen Informationsverluste sind auch in Hinblick auf die Generalisierbarkeit der Befunde relevant. Um diese Konsequenzen aufzeigen zu können, müssen wir zunächst zwischen zwei typischen Zielen der Generalisierung unterscheiden, die auf Basis einer Betrachtung der Verläufe und Peaks der aggregierten RTR-Zeitreihen angestrebt werden können. Zum einen geht es dabei um die Fähigkeit der Peak-Spike-Analyse, die „richtigen" Ausschnitte der Zeitreihe zu identifizieren. Da nach der Logik der Peak-Spike-Analyse oft nur die Ausschnitte der Debatte inhaltlich näher betrachtet werden, bei denen die RTR-Zeitreihe einen (relativ oder absolut) definierten Grenzwert überschreitet, ist die Präzision dieses Tests von großer Bedeutung. Würde eine große Unsicherheit darüber bestehen, welche Passagen für die weitere Interpretation ausgewählt werden sollen, so stände auch die Grundlage für die weiterführenden Interpretationen unter dem Vorbehalt dieser Unsicherheit. In letzter Konsequenz würde dadurch die Möglichkeit beeinträchtigt, aus den Peak-Spike-Analysen Annahmen über die Wirkung bestimmter Inhalte abzuleiten.

Zum anderen sollen die RTR-Zeitreihen und ihre Peaks als Indikator dafür dienen, wie die Kandidaten von der Grundgesamtheit aller Debattenzuschauer bewertet werden. Es soll also ein klassischer Inferenzschluss von der Personen-

43 Es könnte natürlich argumentiert werden, dass auch die Lagerzugehörigkeit einer Person lediglich ein individuelles Merkmal ist. Gegen dieses Argument spricht aber nicht zuletzt, dass die relative Stimmabgabe im Aggregat der Wahlbevölkerung – das Wahlergebnis – eindeutig eine real bedeutsame Aggregatgröße ist. Daher ist nicht nur die Erklärung der individuellen Stimmabgabe, sondern auch die Erklärung des aggregierten Wahlergebnisses ein relevantes Feld der empirischen Wahlforschung.

Stichprobe auf die Grundgesamtheit aller Zuschauer vorgenommen werden – auch wenn die statistische Repräsentativität der realisierten Stichproben aus praktischen Einschränkungen des Designs heraus kaum gewährleistet werden kann (vgl. Kapitel 3.3). Unter der Annahme, dass wir den Ergebnissen der TV-Duell-Studien eine solche Indikatorfunktion für die Wahrnehmung der untersuchten Debatte durch alle Zuschauer zuschreiben können, wollen wir im Folgenden auch prüfen, mit welcher Sicherheit wir die durch die RTR-Zeitreihen erfassten unmittelbaren Kandidatenbewertungen auf die Grundgesamtheit übertragen können.

Die Präzision des Mittelwert-Schätzers wird durch seinen Standardfehler ausgedrückt. Da die RTR-Zeitreihen als Mittelwerte über die RTR-Werte aller Probanden in einer Sekunde gebildet werden (vgl. Formel 5.2), lässt sich der Standardfehler der RTR-Reihe in jeder Sekunde berechnen als

$$SE_{m_t} = \frac{SD_t}{\sqrt{n}} \tag{5.8}$$

wobei SE_{m_t} der Standardfehler des Mittelwerts m_t aller Probanden zum Zeitpunkt t, SD_t die geschätzte Standardabweichung um den wahren Mittelwert in der Grundgesamtheit und n die Gesamtzahl aller Probanden ist.

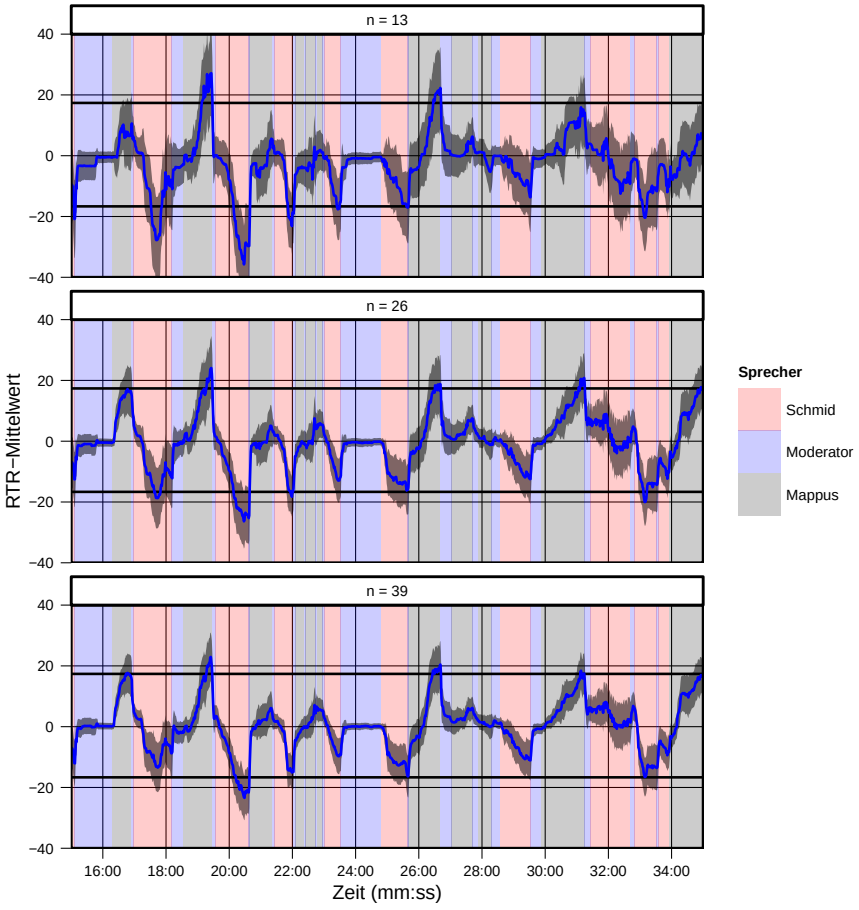
Unter der Annahme des zentralen Grenzwerttheorems, dass die Schätzer m_t um den wahren Mittelwert in der Grundgesamtheit normalverteilt sind, können wir mithilfe des Standardfehlers ein Konfidenzintervall um die RTR-Mittelwert-Zeitreihen konstruieren. Allerdings fällt auf, dass beide hierfür notwendigen Informationen – die Streuung um den Mittelwert sowie die Fallzahl der Gruppe, deren RTR-Messungen zusammengefasst werden – überhaupt nicht im Zeitreihen-Datensatz, anhand dessen wir alle Peak-Spike-Analysen durchführen, enthalten sind (vgl. Tabelle 5.2). Dies verdeutlicht ein weiteres Mal, dass die statistische Unsicherheit, die im Personendatensatz enthalten ist, bei der Analyse der RTR-Zeitreihen ignoriert wird.

Im Folgenden untersuchen wir, inwiefern es die Interpretation der aggregierten RTR-Zeitreihen betrifft, wenn wir die statistische Unsicherheit berücksichtigen. Als erstes Beispiel dient uns ein Ausschnitt der Bewertung von Mappus und Schmid durch die Zuschauer, die vor dem Duell noch keine Wahlabsicht geäußert hatten. Dargestellt wird der Verlauf von der 15. Minute bis zur 35. Minute, da in diese 20 Minuten die meisten Peaks fallen (Bachl, 2013a, S. 153). Die Unentschiedenen bilden mit n = 39 die kleinste Gruppe in unserer Stichprobe. Da die Größe des Standardfehlers negativ mit der Fallzahl

zusammenhängt, sollten hier die schwerwiegendsten Konsequenzen zu sehen sein. Viele Studien berichten sogar die Ergebnisse von Peak-Spike-Analysen mit noch geringeren Fallzahlen (vgl. den Forschungsstand in Kapitel 3.4.1). Um eine Einschätzung davon zu erhalten, wie sich noch kleinere Fallzahlen auf die Präzision der Ergebnisse auswirken, betrachten wir zudem Zufallsstichproben aus unserer Stichprobe mit n = 13 und n = 26. Als erstes Kriterium zur Identifikation der Peaks wählen wir einen festen Grenzwert bei Abweichung von 16.7 Skalenpunkten von der Skalenmitte. Dies entspricht der Abweichung von einem Skalenpunkt der siebenstufigen Skala, wie sie auch in anderen Studien als Kriterium gewählt wurde (Maurer & Reinemann, 2003; Reinemann & Maurer, 2007b). Abbildung 5.9 zeigt die RTR-Reihe der Unentschiedenen und ihre 95%-Konfidenzintervalle für die drei Stichprobenumfänge.

Der Vergleich der drei Facetten, in denen die Verläufe für die unterschiedlichen Fallzahlen abgetragen sind, weist in zweierlei Hinsicht auf die Bedeutung des Stichprobenumfangs hin. Zum einen werden die Konfidenzintervalle mit sinkender Fallzahl weiter. Zum anderen unterscheiden sich besonders die Zeitreihe auf Basis der gesamten Stichprobe und die Zeitreihe der kleinsten Teilstichprobe mit 13 Probanden an einigen Stellen recht deutlich voneinander. Wäre ein (zufällig ausgewähltes) Drittel der Unentschiedenen am Tag unserer Erhebung nicht erschienen, hätten sich die Befunde auf Basis der hier dargestellten Analyse nicht wesentlich verändert. Hätte es aber einen Ausfall von zwei Dritteln der Teilstichprobe gegeben, so würden sich die Ergebnisse an einigen Stellen verändern. Die Erkenntnis, dass die Analyse einer Teilstichprobe von n = 13 keine überaus stabilen Ergebnisse zeigt, mag eigentlich offensichtlich sein. Fakt ist jedoch, dass in der Literatur einige Analysen berichtet werden, die sich auf Peak-Spike-Analysen mit Fallzahlen dieser Größenordnung stützen. Doch selbst die Betrachtung der Zeitreihe aller Unentschiedenen und ihrer Konfidenzintervalle offenbart eine Unsicherheit, die bei der Analyse von Mittelwerten kleinerer Gruppen eigentlich nicht überrascht, im gebräuchlichen Vorgehen der Peak-Spike-Analyse aber ignoriert wird. Wir können klar sehen, dass sich die Konfidenzintervalle an vielen Stellen mit den festgesetzten Grenzwerten für die Peak-Identifikation überschneiden. Dies ist sowohl für Passagen der Fall, deren Punktschätzer die Grenze überschreiten – die Passage würde fälschlicherweise als bedeutsam eingeordnet – als auch für Passagen, in denen die Punktschätzer innerhalb der Grenzen liegen – die Passage würde fälschlicherweise als nicht bedeutsam aussortiert.

Abbildung 5.10 verdeutlicht dieses Problem noch einmal für die Mittelwert-Zeitreihe des gesamten Publikums. Neben der gesamten Stichprobe von n = 176 Probanden werden auch zufällig ausgewählte Teilstichproben mit Umfängen von einem Drittel (n = 59) bzw. zwei Drittel (n = 118) der gesamten Stichpro-



Anmerkungen
Mittlere Bewertung von Mappus und Schmid von der 15. bis zur 35. Minute des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch die Unentschiedenen. Mittelwert-Zeitreihe (blau), 95%-Konfidenzintervall um den Mittelwert (graue Fläche), Peak-Grenzen bei ±16.7 Skalenpunkten (horizontale Linien). Die Facetten zeigen unterschiedliche Stichprobenumfänge. Die Farbe im Hintergrund zeigt an, wer gerade das Wort hatte.

Abbildung 5.9: Aggregierte RTR-Zeitreihe mit Konfidenzintervall für die Unentschiedenen

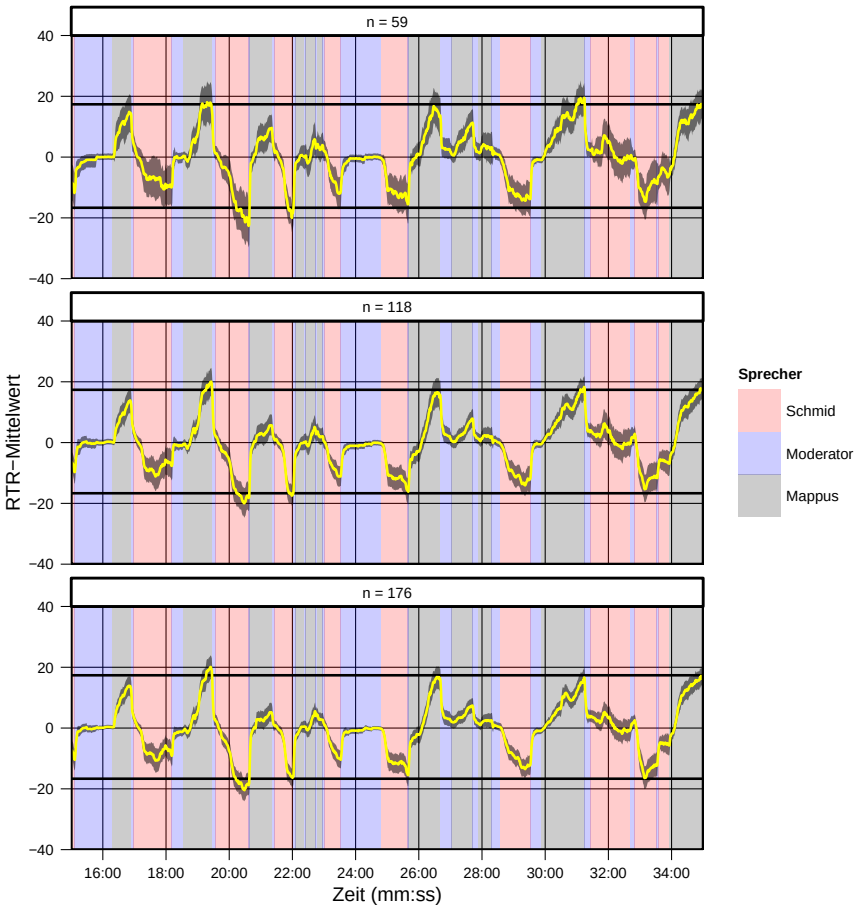
be dargestellt, um einen Eindruck von der Unsicherheit zu erhalten, die sich aus kleineren, jedoch in der Literatur üblichen Fallzahlen ergibt. In allen Darstellungen wird die gleiche Verteilung der Lagerzugehörigkeit angenommen und die Zeitreihe als gewichteter Mittelwert berechnet. Da die Schätzung des Standardfehlers bzw. des Konfidenzintervalls eines gewichteten Mittelwerts mit asymptotischen Verfahren fehleranfällig ist, wird die Verwendung eines Bootstrap-Verfahrens empfohlen (Gatz & Smith, 1995a, 1995b). Daher werden die hier präsentierten Konfidenzintervalle mithilfe des Bootstrap (Efron & Tibshirani, 1986, 1991) geschätzt.[44]

Auch die Analyse der RTR-Zeitreihe für das gesamte Publikum ist mit einer recht großen Unsicherheit behaftet. Die Ergebnisse erstaunen angesichts der bereits in den Abbildungen 5.3 und 5.8 dargestellten Varianz der individuellen Zeitreihen des gesamten Publikums nicht. Zwar sind die Fallzahlen im Vergleich zur Betrachtung einer Subgruppe größer, jedoch wirkt die ebenfalls größere Streuung einem Schrumpfen der Konfidenzintervalle entgegen. In der Folge ist auch hier eine beträchtliche Anzahl von Überschneidungen der Signifikanzgrenzen mit den Konfidenzintervallen zu beobachten. Eine Erhöhung der Fallbasis der aggregierten RTR-Zeitreihe durch die Zusammenfassung größerer Teilgruppen ist demnach keine geeignete Maßnahme, um die Unsicherheit einer Peak-Spike-Analyse wesentlich zu reduzieren. Wenn wir das erreichen wollen, müssen wir die Zahl der Probanden in allen relevanten Teilgruppen der Stichprobe steigern.

5.2.3 Erweiterung: Bootstrap-Peak-Spike-Analyse

Für die Identifikation von bedeutsamen Peaks nach relativen Grenzwerten ist die Bestimmung der statistischen Unsicherheit etwas komplexer. Wenn wir nach dem Vorschlag von Biocca et al. (1994) diejenigen Peaks finden wollen, die außerhalb von ± 1.96 Standardabweichungen um den Mittelwert der Zeitreihe liegen, ist die Information interessant, mit welcher Sicherheit dieser Grenzwert erreicht wird. Dafür reicht es jedoch nicht aus, einfach die Konfidenzintervalle der RTR-Zeitreihe mit den auf Grundlage ihrer Verteilung ermittelten Grenzwerten zu vergleichen. Wenn wir – wie es die Grundlage dieses Vergleichs ist – annehmen, dass der wahre Wert der Zeitreihe an einer anderen Stelle im Konfidenzintervall liegt, so würden sich auch der Mittelwert und die Standardabweichung der Zeitreihe verändern – und damit die Kriterien, nach denen der zugrunde liegende Grenzwert bestimmt wird. Sobald wir also zur Kenntnis

44 Das Verfahren der Peak-Spike-Analyse mit Bootstrap-Konfidenzintervallen wird im nächsten Abschnitt ausführlicher dargestellt.



Anmerkungen
Mittlere Bewertung von Mappus und Schmid von der 15. bis zur 35. Minute des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch das gesamte Publikum. Mittelwert-Zeitreihe (gelb), 95%-Konfidenzintervall auf Basis von 1000 Bootstrap-Stichproben (graue Fläche), Peak-Grenzen bei ±16.7 Skalenpunkten (horizontale Linien). Die Facetten zeigen unterschiedliche Stichprobenumfänge. Die Farbe im Hintergrund zeigt an, wer gerade das Wort hatte.

Abbildung 5.10: Aggregierte RTR-Zeitreihe mit Konfidenzintervall für das gesamte Publikum

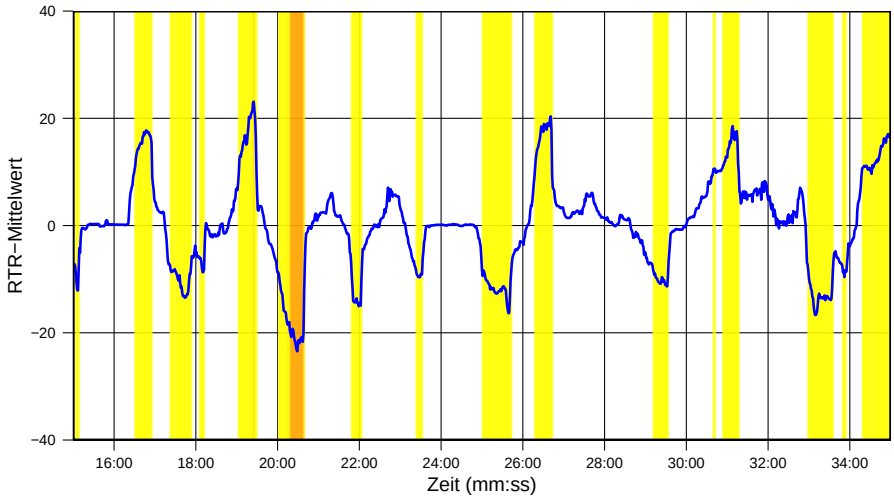
nehmen, dass mit einer aggregierten RTR-Zeitreihe eine statistische Unsicherheit verbunden ist, müssen wir auch folgern, dass diese Unsicherheit ebenso für auf der empirischen Verteilung dieser Zeitreihe basierende Grenzwerte gilt.

Um mit diesem Problem umzugehen, haben wir ein Bootstrap-Verfahren entwickelt, das es ermöglicht, die Unsicherheit zu bestimmen, mit der eine Messung der aggregierten RTR-Zeitreihe über oder unter einem relativ zu dieser Zeitreihe ermittelten Grenzwert liegt. Der Vorteil des Bootstrap ist, dass auch für Parameter, für die asymptotisch keine Konfidenzintervalle bestimmt werden können (z.B. für den Median: Efron & Tibshirani, 1986, S. 55), eine Schätzung auf Basis der wiederholten Ziehung von Bootstrap-Stichproben aus der gesamten Stichprobe möglich ist (Efron & Tibshirani, 1986, 1991). Bootstrap-Verfahren sind vergleichsweise rechenintensiv, lassen sich aber einfach programmieren und mit modernen Computern in angemessener Rechenzeit durchführen. Ziel des Verfahrens ist es, den Parameter „Auswahl als bedeutsamer Peak" für jede Sekunde des TV-Duells zu schätzen. Der Parameter ist dichotom mit den Ausprägungen 0 (kein bedeutsamer Peak) und 1 (bedeutsamer Peak). Geschätzt werden soll ein Konfidenzintervall für die Information, dass zu einer Sekunde die Ausprägung 1 vorliegt. Im Folgenden beschreiben wir die grundsätzliche Logik der Bootstrap-Peak-Analyse am Beispiel der RTR-Bewertung durch die Unentschiedenen:[45]

1. Aus der Stichprobe aller Unentschiedenen (n = 39) werden 1000 Bootstrap-Stichproben der gleichen Größe mit Zurücklegen gezogen.
2. Für jede der 1000 Bootstrap-Stichproben wird eine Mittelwert-Zeitreihe aus den individuellen RTR-Messungen der jeweils gezogenen Probanden berechnet.
3. Für jede der 1000 Mittelwert-Zeitreihen werden der Mittelwert und die Standardabweichung berechnet, um die Peak-Grenzen bei $M \pm 1.96SD$ zu bestimmen.
4. Für jede Sekunde der 1000 Mittelwert-Zeitreihen wird bestimmt, ob der Wert außerhalb der Peak-Grenzen der jeweiligen Zeitreihe liegt und damit als bedeutsamer Peak identifiziert werden kann.
5. Für jede Sekunde wird gezählt, in wie vielen der 1000 Bootstrap-Stichproben sie als bedeutsamer Peak identifiziert wurde.
6. Über die Perzentile der Häufigkeitsverteilung kann nun ein Konfidenzintervall konstruiert werden. Das 95%-Konfidenzintervall einer Sekunde,

45 Das technische Vorgehen ist in Anhang B dokumentiert.

die häufiger als 975 Mal identifiziert wurde, liegt vollständig außerhalb der Peak-Grenze. Dies sind die Passagen, bei denen mit angemessener Sicherheit davon ausgegangen werden kann, dass sie sich signifikant von den Bewertungen der übrigen Passagen abheben. Das 95%-Konfidenzintervall einer Sekunde, die in 1000 Bootstrap-Stichproben seltener als 25 Mal als bedeutsam identifiziert wurde, liegt vollständig unterhalb der Peak-Grenze. Die Konfidenzintervalle aller anderen Sekunden überschneiden sich mit der Peak-Grenze.



Anmerkungen
Mittlere Bewertung von Mappus und Schmid von der 15. bis zur 35. Minute des TV-Duells auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus) durch die Unentschiedenen (n = 39). Mittelwert-Zeitreihe (blau). Passagen, die in mindestens 975 der 1000 Bootstrap-Stichproben als bedeutsam identifiziert wurden, sind orange markiert. Passagen, die in mindestens 25 der 1000 Bootstrap-Stichproben als bedeutsam identifiziert wurden, sind gelb markiert.

Abbildung 5.11: Aggregierte RTR-Zeitreihe mit bedeutsamen Peaks nach einer Bootstrap-Peak-Analyse für die Unentschiedenen

Abbildung 5.11 visualisiert die Ergebnisse der beschriebenen Bootstrap-Peak-Analyse. Da nicht nur die Mittelwert-Zeitreihen, sondern auch die Grenzwerte über die Bootstrap-Stichproben hinweg variabel sind und damit in

der grafischen Repräsentation ein Konfidenzintervall besitzen müssten, steht die intuitive Präsentationsform der Abbildungen 5.9 und 5.10 nicht mehr zu Verfügung. In Abbildung 5.11 sind daher die Passagen, in denen das 95%-Konfidenzintervall vollständig über dem Grenzwert liegt, als orange Flächen markiert. Alle Passagen, bei denen sich das 95%-Konfidenzintervall mit dem Grenzwert überschneidet, können an der gelben Fläche erkannt werden.

Die große Zahl gelber Markierungen ruft wiederum die geringe Präzision der Peak-Spike-Analyse in Erinnerung. Da ein Konfidenzintervall auf Basis von nur 39 Probanden sehr weit ist, können wir für viele Aussagen der Kandidaten nicht mit angemessener Sicherheit ausschließen, dass sie unter den bedeutsamen Momenten der Debatte zu finden waren. Der hier dargestellte Duellausschnitt führt zwar zu einer leichten Überschätzung dieses Phänomens, da wir zur Illustration die Phase der Debatte gewählt haben, in der nach einer einfachen Peak-Analyse die meisten Peaks in der Zeitreihe der Unentschiedenen zu finden sind. Doch auch im gesamten Duellverlauf werden so deutlich mehr Stellen identifiziert, die bedeutsam sein könnten. Im Umkehrschluss ist es auch klar, dass nur an wenigen Stellen die Bewertung durch die Unentschiedenen so eindeutig war, dass diese Passagen mit großer Sicherheit als bedeutsam gelten können. In der gesamten Debatte waren dies nur fünf teils sehr kurze Peaks. Im Vergleich dazu machen wir in unserer einfachen Analyse 16 bedeutsame Peaks dieser Zeitreihe aus (Bachl, 2013a, S. 153).

Zwischenfazit: Konsequenzen der Zeitreihen-Aggregation für die Generalisierbarkeit der Befunde Der letzte Abschnitt hat verdeutlicht, dass die aggregierten RTR-Zeitreihen eine relativ geringe Präzision aufweisen. Verantwortlich hierfür sind zum einen die große Varianz zwischen den Wertungen der zu einem Aggregat zusammengefassten Rezipienten, zum anderen die relativ geringen Fallzahlen der TV-Duell-Studien, vor allem bei der Analyse von Teilstichproben. Wir können damit nicht sehr genau angeben, wie bestimmte Gruppen in der Grundgesamtheit die Kandidaten zu einem Zeitpunkt bewertet haben, und wir können nur relativ ungenau einschätzen, welche Stellen der RTR-Zeitreihen auf bemerkenswerte Stimulusinhalte hindeuten.

Die geringe Präzision in Hinblick auf die Bewertung in der Grundgesamtheit fügt den ohnehin bestehenden Problemen der nicht-repräsentativen Stichproben für die Indikatorqualität der Befunde weitere Einschränkungen hinzu. Die oft sehr weiten Konfidenzintervalle verdeutlichen, dass wir die aggregierten RTR-Zeitreihen nur sehr vorsichtig als einen Indikator für die Bewertung der Kandidaten außerhalb der Stichprobe interpretieren sollten. Vor dem Hintergrund der in solchen Rezeptionsstudien realisierbaren Stichproben sind die hier

berichteten Konfidenzintervalle auch nicht als eine Information über die Lage der wahren Bewertungen in der Grundgesamtheit aller Debattenrezipienten zu verstehen. Sie sollen vielmehr verdeutlichen, dass wir selbst in der Übertragung der deskriptiven Befunde auf die Grundgesamtheit, für die unsere Stichprobe tatsächlich steht, keine sehr genauen Aussagen machen können. Diese Probleme lassen sich schwerlich beheben – sicherlich nicht durch die Wahl besserer Analyseverfahren und aufgrund der praktischen Hürden einer Rezeptionsstudie mit großer apparativer Voraussetzung auch kaum durch eine wesentliche Verbesserung der Stichprobenqualität. Gerade darum sollte jede Beschreibung der aggregierten Kandidatenbewertungen im Debattenverlauf, wenn sie auch ein Indikator für die Bewertung der Kandidaten in einem größeren Kontext sein will, verantwortungsvoll mit der hier präsentierten Unsicherheit umgehen und die Ergebnisse nur vorsichtig interpretieren.

Die relativ geringe Präzision der Zeitreihen hat auch für die Peak-Spike-Analyse als induktives Analyseverfahren Konsequenzen. Die weiten Konfidenzintervalle haben zur Folge, dass unsere Information darüber, ob eine RTR-Reihe den zur Identifikation eines Peaks gesetzten Grenzwert überschreitet, recht ungenau ist. Nun mag man einwenden, dass dies in einer explorativen Analyse, die der Generierung neuer Hypothesen dient, weniger konsequenzenreich sei als ein fälschliches Falsifizieren oder Stützen einer Hypothese. Dieser Einwand gilt jedoch nur bedingt. Denn in fast allen hier gesichteten TV-Duell-Studien mit Peak-Analysen – einschließlich unserer eigenen – ist das Überschreiten eines festgesetzten Grenzwerts durch die RTR-Kurven die Bedingung dafür, dass eine Passage des Duells überhaupt zur interpretativen Exploration von besonders wirksamen Merkmalen des Debatteninhalts herangezogen wird. Dass die Vernachlässigung der übrigen Inhalte eine Schwäche der Peak-Spike-Analyse ist, haben wir bereits als Konsequenz ihrer induktiven Analyselogik diskutiert. Nun kommt aber noch hinzu, dass die Entscheidung, ob eine Passage interpretiert oder vernachlässigt wird, auf der Grundlage eines sehr unpräzisen Schätzers vorgenommen wird. Hätte sich die Zusammensetzung unserer Stichprobe aus derselben Grundgesamtheit nur leicht anders zusammengesetzt, so wäre wahrscheinlich die eine oder andere Stelle des Duells (nicht) zur weiteren Analyse aufgenommen worden.

Zukünftige Peak-Spike-Analysen sollten dieser Unsicherheit über ihr Auswahlkriterium Rechnung tragen. Prinzipiell stehen dafür drei Möglichkeiten zur Wahl. *Erstens* könnten nur diejenigen Peaks ausgewählt werden, deren gesamtes Konfidenzintervall außerhalb der festgelegten Grenzen liegt. Dadurch würden deutlich weniger Peaks erkannt, was die Grundlage der Interpretation weiter einschränken und die Kritik an der Vernachlässigung großer Teile des Duells verschärfen würde. Allerdings wäre die Sicherheit, dass es sich wirk-

lich um die (nach den selbst gewählten Kriterien) bemerkenswerten Passagen handelt, recht groß. *Zweitens* könnten sämtliche Peaks berücksichtigt werden, deren Konfidenzintervall sich mit den Grenzwerten überschneidet. So würden recht viele Peaks ausgewählt, was die Interpretation aufwändiger macht und zudem die Gefahr erhöht, auch bedeutungslose Passagen aufzugreifen. Dafür würden so alle Inhalte berücksichtigt, die mit einer gewissen Sicherheit zu den bemerkenswerten Stellen gehören könnten. *Drittens* könnten die präsentierten Befunde zum Anlass genommen werden, von einer strikten Orientierung an quantitativen Grenzen generell Abstand zu nehmen. Stattdessen könnten nach einem qualitativ-systematischen Verfahren z.B. so viele Passagen in absteigender Reihenfolge ihrer Bewertung zur Interpretation herangezogen werden, bis die auf Basis der ersten Stellen entwickelte Erklärung sich ausreichend erhärtet oder widerlegt wird (Rust, 1985). Welche dieser Varianten gewählt wird, hängt von den Spezifika der jeweiligen Analyse ab. In der einen oder anderen Form sollten die hier präsentierten Befunde zur (mangelnden) Präzision der RTR-Zeitreihen auf jeden Fall berücksichtigt werden, da die reflektierte Auswahl der Inhalte die Basis für die Qualität der folgenden Interpretationen, Ableitungen und generierten Hypothesen ist.

5.2.4 Empfehlungen

In den vorangegangen Teilkapiteln haben wir gezeigt, wie bei Peak-Spike-Analysen der über eine Personenaggregation gebildeten RTR-Zeitreihen vorgegangen wird. Dabei haben wir festgestellt, dass das Verfahren Einschränkungen unterliegt, die größtenteils auf den unangemessenen Umgang mit den Folgen der gewählten Aggregationslogik zurückzuführen sind. Die Darstellung von RTR-Messungen als über alle oder Teilgruppen der Probanden aggregierte Verlaufskurven wird wegen ihrer unkomplizierten Durchführung und einfachen Kommunizierbarkeit immer ein wichtiges Verfahren der deskriptiven und explorativen Analyse von RTR-Daten bleiben. Auch die Peak-Spike-Analyse zur induktiven Identifikation besonders wirksamer Passagen des Stimulus ausgehend von den aggregierten Publikumsreaktionen ist für explorativ angelegte Arbeiten ein bewährtes Verfahren. Im Folgenden fassen wir knapp die sechs wichtigsten Empfehlungen zusammen, die wir aus unserer Diskussion ableiten können:

1. Durch die Aggregation der individuellen RTR-Messungen zu RTR-Zeitreihen geht der Bezug zur Individualebene verloren. Dies sollte insbesondere bei der Interpretation der RTR-Verlaufskurven und der identifizierten Peaks berücksichtigt werden. Nur eine Interpretation, die sich auch inhaltlich auf

die Aggregate bezieht, ist dem Aggregatniveau der RTR-Zeitreihen angemessen. Interpretationen, die auf die individuellen Bewertungen abzielen und sie z.B. durch die individuellen Informationsverarbeitungsprozesse erklären wollen, finden sich in den aggregierten Daten nicht wieder.

2. Um einen Eindruck von den hinter den aggregierten RTR-Zeitreihen liegenden individuellen RTR-Messungen zu erhalten, sollten diese zumindest immer auch einer visuellen Analyse unterzogen werden. Dies gilt insbesondere für die RTR-Messungen zu den Ausschnitten der Stimuli, die als bedeutsame Peaks einer aggregierten Zeitreihe identifiziert werden. Da diese Ausschnitte die Interpretation der Befunde dominieren, sollte auch berücksichtigt werden, wie sie von den individuellen Probanden bewertet werden.

3. Wenn davon auszugehen ist, dass die Voreinstellungen der Probanden die individuellen RTR-Bewertungen stark prägen, sollte dies auch bei der Analyse der RTR-Zeitreihen durch die Bildung von Teilaggregaten berücksichtigt werden. Es ist in den Studien zur Kandidatenbewertung in TV-Duellen gute Praxis, die Verläufe getrennt nach politischen Lagern zu untersuchen. Von der (zusätzlichen) Analyse einer Zeitreihe für das gesamte Publikum ist vor dem Hintergrund der enormen Varianz zwischen den individuellen Messungen abzuraten. Dies gilt vor allem für die öffentliche Kommunikation der Ergebnisse, in der trotz des Hinweises auf die eingeschränkte Aussagekraft dieser Darstellung das Bild des „größten Erfolgs beim gesamten Publikum“ hängen bleibt.[46] Wenn mit einer Peak-Spike-Analyse die beim gesamten Publikum erfolgreichsten Aussagen identifiziert werden sollen, sollte nach dem Vorbild von Reinemann und Maurer (2005) nach Stellen des Duells gesucht werden, an denen alle nach politischem Lager gebildeten Teilgruppen eine überdurchschnittliche Zustimmung zeigen.

4. In allen Analysen der aggregierten RTR-Zeitreihen muss die Unsicherheit, die um die Mittelwert-Zeitreihen herum besteht, berücksichtigt werden. Die beträchtliche interindividuelle Varianz und die (sehr) kleinen Fallzahlen, auf denen die Zeitreihen basieren, sorgen dafür, dass die Zeitreihen nur relativ unpräzise Indikatoren sind. Dies ist für die deskriptiven, explorativen Analysen noch nicht *per se* ein Problem, da (auch wegen der mangelhaften Repräsentativität der Stichproben) ein direkter Inferenzschluss auf die Grundgesamtheit aller Rezipienten eines Stimulus ohnehin nicht möglich

46 Diese Erfahrung haben wir selbst bei der öffentlichen Kommunikation der Befunde aus der Rezeptionsstudie zum baden-württembergischen TV-Duell gemacht.

ist. Da die RTR-Zeitreihen trotzdem meist auch als ein Indikator für den Bewertungsverlauf in einer Grundgesamtheit aufgefasst werden, müssen diese Interpretationen mit einer der Unsicherheit angemessenen Vorsicht erfolgen. Im Idealfall sollten dazu die Konfidenzintervalle um die Zeitreihen auch in den Ergebnispräsentationen dargestellt werden, um nicht den Eindruck einer hochpräzisen technischen Messung zu vermitteln. Besonders kritisch zu sehen ist vor diesem Hintergrund die Präsentation von RTR-Zeitreihen live während der Ausstrahlung eines TV-Duells. Mehrere Studien, in denen eingeblendete RTR-Zeitreihen experimentell variiert wurden, haben nachgewiesen, dass solche Präsentationen die Bewertung der Kandidaten in einer Debatte beeinflussen (Davis, Bowers & Memon, 2011; Weaver III, Huck & Brosius, 2009; Wolf, 2010). Da es kaum möglich scheint, dem Publikum die mit den RTR-Kurven verbundene Unsicherheit in angemessener Form zu erläutern, raten wir dazu, von solchen Einblendungen Abstand zu nehmen.

5. Besonders relevant ist die Berücksichtigung der die aggregierten RTR-Zeitreihen umgebenden Unsicherheit bei der Identifikation der bedeutsamen Passagen des Stimulus mit Peak-Spike-Analysen. Da bei der Untersuchung längerer Stimuli häufig nur die Inhalte des Stimulus auf Wirkungen auslösende Merkmale hin interpretiert werden, die in einer solchen Analyse als relevanter Peak identifiziert werden, beeinflusst die Präzision der Identifikation direkt das Ergebnis der Analysen. Je nachdem, ob die Vermeidung von Fehlern erster oder zweiter Art bei der Peak-Identifikation als wichtiger eingeschätzt wird, sollten entweder nur Peaks ausgewählt werden, die bei einer vorgegebenen Irrtumswahrscheinlichkeit über den gewählten Grenzwerten liegen, oder alle Peaks ausgewählt werden, für die mit einer vorgegeben Irrtumswahrscheinlichkeit nicht ausgeschlossen werden kann, dass sie über dem Grenzwert liegen. Besonders für sehr kleine (Teil-) Stichproben, deren RTR-Zeitreihen mit einer entsprechend großen Unsicherheit behaftet sind, sollte auch der Verzicht auf einen quantitativen Grenzwert in Betracht gezogen werden. Wie die Präzision der Peak-Identifikation bei festen und relativen Grenzwerten evaluiert werden kann, haben wir oben ausführlich gezeigt.

6. Schließlich sollten natürlich die Einschränkungen der Peak-Spike-Analyse, die in früheren Arbeiten bereits herausgestellt wurden, auch weiterhin beachtet werden. Hier sei an die Einschränkung des induktiven Vorgehens, das zu einer Vernachlässigung der nicht als Peak identifizierten Inhalte führt, und an die notwendige Systematik bei der Interpretation der den Peaks vorangehenden Stimulusinhalte (positives Vorbild: Reinemann & Maurer, 2005) erinnert.

5.3 Verfahren zur deduktiven Analyse

Auch die meisten deduktiv angelegten Analysen zur Erklärung der unmittelbaren Kandidatenbewertungen in TV-Duellen bauen auf über die Rezipienten aggregierten RTR-Zeitreihen auf (Kapitel 5.3.1). Seltener werden Verfahren eingesetzt, die auf einer Zusammenfassung der RTR-Messungen über mehrere Messzeitpunkte basieren (Kapitel 5.3.2). Abschließend fassen wir die Ergebnisse unserer Diskussionen zusammen und nennen mögliche Lösungen für die aufgeworfenen Probleme (Kapitel 5.3.3). Einen dieser Lösungsansätze, die Analyse der RTR-Messungen mit Mehrebenenmodellen, stellen wir in Kapitel 6 ausführlich vor.

5.3.1 Verfahren auf Basis der Aggregation über Personen

Wie die induktiv angelegten Arbeiten, die RTR-Zeitreihen mit Peak-Spike-Analysen untersuchen, ziehen auch die meisten deduktiven Analysen über Personen aggregierte RTR-Messungen als abhängige Variable heran. Dabei kann zwischen einem zeitreihenanalytischen Ansatz und einem aussagenanalytischen Ansatz unterschieden werden. Im erstgenannten Ansatz wird die RTR-Zeitreihe direkt als abhängige Variable genutzt. Im zweitgenannten Ansatz wird eine weitere Aggregation über die Zeit vorgenommen, in der mehrere aufeinanderfolgende aggregierte RTR-Messungen zu Aussagen zusammengefasst werden. Anschließend diskutieren wir zusammenfassend die Einschränkungen, die sich auch bei diesen Vorgehensweisen aus der Aggregation der individuellen RTR-Messungen zu Publikumsaggregaten ergeben.

Zeitreihenanalytisches Vorgehen

Den ambitioniertesten statistischen Ansatz zur Verknüpfung von Merkmalen des Debatteninhalts und deren unmittelbaren Bewertung durch das Publikum legt Nagel (2012) vor (vgl. auch Nagel et al., 2012).[47] Vielfältige inhaltsanalytisch sekundengenau erfasste Merkmale der Debatte auf verbaler, visueller und vokaler (Stimmhöhe, Sprechlautstärke, Sprechgeschwindigkeit) Ebene werden mit den ebenfalls auf Sekundenbasis vorliegenden aggregierten RTR-Zeitreihen kombiniert. Damit liegt für die Analyse ein klassischer Zeitreihen-Datensatz vor. Er besteht aus Zeitreihen zum Debatteninhalt in jeder Sekunde als unabhängige Variablen und für jede Analyse aus einer Zeitreihe zur aggregierten

47 Das hier vorgestellte Verfahren ähnelt den zeitreihenanalytischen Ansätzen, mit denen die rezeptionsbegleitend erfassten Emotionen während der Musikrezeption erklärt werden (z.B. Schubert, 1999, 2004).

Kandidatenbewertung durch alle Zuschauer bzw. Teilgruppen von Zuschauern als abhängige Variable. Das statistische Verfahren, das die Autorin zur Verknüpfung dieser Zeitreihen entwickelt, soll drei Bedingungen erfüllen: Es soll eine Latenzzeit zwischen Vorkommen und Wirkung der Debattenmerkmale berücksichtigen, multivariat sein und dem Zeitreihencharakter der Daten gerecht werden.

Die Latenzzeit wird empirisch als das Zeitintervall bestimmt, nach dem der Effekt der wichtigsten unabhängigen Variablen auf die Publikumszeitreihen am größten ist. Dazu werden die Korrelationen der abhängigen RTR-Zeitreihe mit den um unterschiedliche Zeitintervalle verschobenen Zeitreihen der wichtigsten inhaltlichen Merkmale untersucht. Da die Korrelationen bei einer Verzögerung um vier Sekunden die größten Beträge aufweisen, wird diese Zeitspanne in den folgenden Analysen als Latenzzeit angenommen. Eine Diskussion bestehender zeitreihenanalytischer Verfahren (ARIMA- und GLS-Regression) kommt zu dem Schluss, dass diese zwar in einiger Hinsicht den geforderten Kriterien entsprächen, jedoch zu große Anteile der Varianz in der RTR-Zeitreihe entfernen würden, um schlüssige Ergebnisse zu erzielen. Als alternatives statistisches Verfahren wird die „Regression mit gleitenden Summen“ (Nagel, 2012, S. 167) entwickelt. Dabei werden alle Zeitreihen der Debatteninhalte zuerst in Dummy-Zeitreihen für jedes vorkommende Merkmal zerlegt. Die Werte der drei aufeinanderfolgenden Sekunden jeder Zeitreihe zu Debattenmerkmalen werden miteinander verrechnet, um zu modellieren, dass die Wirkung von Merkmalen nicht von einer Sekunde auf die nächste auftritt und sofort wieder verschwindet, sondern eine gewisse Zeitspanne vorhält. In verschiedenen Modellen werden schließlich die RTR-Zeitreihen auf eine große Zahl von so aufbereiteten unabhängigen Zeitreihen der Debatteninhalte regressiert. Dies soll vergleichende Befunde zur Wirksamkeit der unterschiedlichen Merkmale des Debatteninhalts ermöglichen.

Das Vorgehen wird transparent dokumentiert und ausführlich begründet. Besonders die Modellierung der unabhängigen Variablen kann gut nachvollzogen werden. Der Ansatz hat jedoch aus unserer Sicht drei wesentliche Einschränkungen. *Erstens* erfordert das gewählte Vorgehen die Definition einer exakten Transferfunktion, die beschreibt, wann und in welcher Weise die Effekte der Debatteninhalte in der Zeitreihe der Publikumsurteile sichtbar werden. Aus der Kombination von Latenzzeit und gleitenden Summen ergibt sich die folgende Transferfunktion: Ein Merkmal, das in Sekunde 0 auftritt, wirkt in Sekunde 4 mit einem Gewicht von 0.5, in Sekunde 5 mit einem Gewicht von 1 und in Sekunde 6 mit einem Gewicht von 0.5. In der Logik des gewählten Regressi-

onsmodells[48] ist diese Transferfunktion statisch, das heißt, dass sie 1) für die Wirkung aller Merkmale gilt, 2) für die Wirkung auf die Bewertung durch alle untersuchten Teilgruppen des Publikums gilt, und 3) über den gesamten Verlauf der 90-minütigen Debatte gilt.

Aber sind diese Annahme plausibel? So weist die Autorin selbst darauf hin, dass die schnellere Verarbeitung visueller Merkmale im Vergleich zu verbalen Botschaften ein wesentlicher Grund dafür ist, dem visuellen Kanal ein größeres Wirkpotenzial zuzuschreiben (Nagel, 2012, S. 53). Das Problem der für alle Merkmale statischen Transferfunktion wäre grundsätzlich zu lösen, indem für die unterschiedlichen Merkmale auf Basis theoretischer Überlegungen oder empirischer Tests eigene Transferfunktionen definiert werden. In Anbetracht der Vielzahl von Merkmalen, die in den Modellen berücksichtigt werden, ist die Wahl einer einzigen Transferfunktion aus pragmatischen Gründen nachvollziehbar. Würde aber beispielsweise – wie es auf Basis der präsentierten theoretischen Überlegungen durchaus möglich wäre – die Wirkung der visuellen Merkmale sofort nach einer Sekunde auftreten und nach einer weiteren Sekunde wieder verschwinden, so könnte dieser Effekt mit der gewählten Transferfunktion nicht erfasst werden. Auf solche Probleme macht auch die qualitative Validierung der Latenzzeiten (Nagel, 2012, S. 156) aufmerksam. Es zeigt sich, dass die Latenzzeit bei markanten Passagen zwischen 3 und 6 Sekunden schwankt. Zwar überlappen sich die so identifizierten Latenzzeiten mit der Transferfunktion, ihrer exakten Form entsprechen sie aber nicht.

Schließlich stellt eine statische, auf Sekundenbasis definierte Transferfunktion besonders hohe Anforderungen an die sekundengenaue Erfassung der Debattenmerkmale in der standardisierten Inhaltsanalyse. Denn nur wenn es gelingt, das Auftreten eines Merkmals wirklich auf die Sekunde genau zu identifizieren und diese Identifikationsgenauigkeit über alle Merkmale und die gesamte Debatte hinweg konstant ist, so kann auch seine Wirkung auf die Zuschauerurteile mit dieser Präzision modelliert werden. Über die Messgenauigkeit der vorliegenden Inhaltsanalyse kann hier nicht geurteilt werden. Nach unseren eigenen Erfahrungen mit einem ähnlich aufgebauten, wenn auch deutlich weniger umfangreichen Codebuch, das ebenfalls eine Erfassung der Inhalte auf Sekundenbasis vorsieht (Bachl, Käfferlein & Spieker, 2013a, 2013b),

48 Dies gilt jedoch nur für das hier gewählte Regressionsmodell, nicht für Regressionsmodelle im Allgemeinen. Verfahren der modernen Zeitreihenanalyse erlauben durchaus die Definition dynamischer Transferfunktionen. Diese Modelle erfordern jedoch die mehrfache Aufnahme der um unterschiedliche Lags verschobenen unabhängigen Zeitreihen (Kirchgässner & Wolters, 2007). In Anbetracht der ohnehin schon großen Zahl der Regressoren ist es nachvollziehbar, dass die Autorin diese Modelle nicht berücksichtigt.

würden wir uns eine derartige Präzision trotz einer insgesamt akzeptablen Reliabilität nicht zutrauen.

Zweitens folgt aus der Berücksichtigung sehr vieler Kategorien des Debatteninhalts, die zudem für jede Ausprägung zu Zeitreihen (gewichteter) Dummy-Variablen transformiert werden, dass die Modelle zur Erklärung der aggregierten RTR-Zeitreihen eine sehr große Anzahl an Prädiktoren umfassen. Die genaue Zahl der geschätzten Koeffizienten schwankt von Modell zu Modell und kann wegen einer Beschränkung der Darstellung auf die Koeffizienten, die ein Signifikanzniveau von mindestens $p < .05$ erreichen, nicht genau nachvollzogen werden. Bereits das einfache „Blockmodell" (Nagel, 2012, S. 171), das nur die verbalen und visuellen Merkmale sowie Interaktionen innerhalb der verbalen Merkmale, jedoch keine Interaktionen zwischen verbalen und visuellen Merkmalen umfasst, enthält bereits 52 zu schätzende Parameter.[49] Für das zentrale Ziel der Arbeit, den relativen Erklärungsbeitrag der einzelnen Kommunikationsmodalitäten unter Kontrolle des Erklärungsbeitrags aller anderen Modalitäten zu bestimmen, ist die Formulierung dieser umfangreichen Modelle größtenteils unproblematisch. Die Merkmale der einzelnen Blöcke werden dazu nacheinander in das Modell aufgenommen und die Veränderung des korrigierten R^2 bei Aufnahme eines Blocks als dessen relativer Erklärungsbeitrag interpretiert (Nagel, 2012, Kap. 12.4).

Wenn aber die Koeffizienten der einzelnen Merkmale von Interesse sind, nimmt die Nützlichkeit dieser sehr umfangreichen Modelle deutlich ab. Ob es bei der Formulierung der Modelle mit einer derart großen Zahl an Regressoren bereits um eine Überanpassung des Modells an die Daten handelt, kann von außen betrachtet nicht beurteilt werden.[50] Sicher feststellen lässt sich, dass sich die präsentierten Regressionstabellen nur sehr schwer in Hinblick auf die tatsächliche Bewertung bestimmter, mit mehreren Merkmalen versehenen Debatteninhalte interpretieren lassen. Zunächst einmal liegt dies daran, dass das Ergebnis einer Regression mit mehr als 50 Parametern ohnehin nur schwer in seiner Gesamtheit zu erfassen ist. Im vorliegenden Fall kommt hinzu, dass sich die Koeffizienten einzelner Inhaltsmerkmale und deren Signifikanztests aufgrund der für sie gewählten Dummy-ähnlichen Skalierung immer nur im Verhältnis zur Konstanten der Regression interpretieren lassen. Die Konstante beschreibt dabei den Mittelwert der Sekunden, in der alle inhaltlichen Kate-

49 Dies geht aus einer Auflistung der Zahl der Regressoren in den Blöcken hervor: „Inhalt: 17; Argumentation: 22 inklusive Interaktionsterme; Rhetorik: 5; nonverbales Verhalten: 7" (Nagel, 2012, S. 201, Fußnote 202). Dazu kommt die Schätzung der Regressionskonstante.

50 Folgen des sogenannten „overfitting" können eine ineffiziente Schätzung der Parameter (auch als Folge der häufig auftretenden Multikollinearität) und eine mangelnde Übertragbarkeit des Modells über den untersuchten Datensatz hinaus sein (Babyak, 2004).

gorien in ihrer Referenzausprägung auftreten (Nagel, 2012, S. 174-175). Aus Tabelle 1 geht hervor, dass dies die Sekunden sind, in denen kein Thema, kein Bezugsgegenstand, keine Evidenzen, keine Appelle, keine Passagenfiguren, kein Gemeinplatz, kein Bezugsobjekt, keine Tendenz, kein Handlungsbezug, kein Blickkontakt des Sprechers oder des Zuhörers, kein Lächeln und keine Gestik des Sprechers bzw. keine optische Kommentierung durch den Zuhörer zu erkennen waren (Nagel, 2012, S. 121). Bei vielen dieser Referenzausprägungen drängt sich die Frage auf, was ein Unterschied zu dieser inhaltsleeren Kategorie inhaltlich bedeuten kann. Verschärft wird dieses Problem noch durch die nicht überraschende Tatsache, dass bestimmte Merkmale häufiger in Kombination miteinander vorkamen (Nagel, 2012, Kap. 10). Um die durchschnittliche Wirkung dieser Kombinationen zu ermitteln, müssten in einem additiven Regressionsmodell die Koeffizienten unter Berücksichtigung der Verteilung der übrigen Merkmale miteinander verrechnet werden, was alleine auf Basis der Regressionstabellen nicht möglich ist.

Auch ein Vergleich der Einflussstärke von Merkmalen mit gleichgerichteten Koeffizienten ist in dieser Darstellung unmöglich. Diese Einschränkung macht sich z.B. bei der Interpretation des Befunds bemerkbar, dass in den Modellen zur Erklärung der Bewertung beider Kandidaten durch alle Zuschauer die Koeffizienten für alle Sachthemen entweder signifikant positiv oder nicht signifikant sind. Die Beträge der Regressionskoeffizienten lassen sich lediglich deskriptiv vergleichen. Es kann aber nicht festgestellt werden, ob ein Thema einen signifikant größeren Einfluss als ein anderes Thema hatte, oder ob sich die Bewertung eines Themas mit einem signifikanten Koeffizienten signifikant von der Bewertung eines Themas mit einem nicht signifikanten Koeffizienten unterscheidet (Gelman & Stern, 2006).

Noch komplexer wird die Interpretation der Regressionsmodelle durch die Aufnahme von Interaktionen zwischen verschiedenen Merkmalen, mit denen der Effekt der relationalen Strategien beschrieben wird. So wird z.B. eine Selbstpräsentation von Schröder durch die Interaktion der Merkmale „positive Tendenz“ und „Bezug zum Regierungslager“ untersucht. Die Koeffizienten solcher multiplikativer Interaktionseffekte werden jedoch anders geschätzt und getestet als die Koeffizienten einfacher Regressoren im linear-additiven Modell, und ihre Aufnahme verändert die Interpretation der Koeffizienten aller Regressoren, die an der Interaktion beteiligt sind (Brambor, Clark & Golder, 2006; Hayes, 2013). Insofern erfasst z.B. die Interpretation, dass „nur die acclaim-Strategie einen signifikant positiven Effekt ($b = 0,08$) auf die kurzfristigen Urteile über Schröder hatte“ (Nagel, 2012, S. 180), die Bedeutung des Koeffizienten nicht vollständig. Da das Bezugsobjekt „Regierungslager“ einen negativen Koeffizienten von $b = -0.07$ aufweist, reicht die Kombination

mit einer positiven Tendenz gerade einmal aus, um den negativen Effekt von Aussagen, die sich ohne eine positive Tendenz auf das Regierungslager bezogen, wieder auszugleichen. An späteren Stellen der Analyse werden gar dreifache Interaktionseffekte berechnet, was die Komplexität der Ergebnisinterpretation als eine Kombination von einfachen linearen sowie konditionalen Effekten weiter steigert.

Die Kritik an den beiden genannten Punkten soll nicht als Generalkritik an der von Nagel (2012) verfolgten analytischen Strategie verstanden werden. Wenn es das Ziel einer Analyse ist, die sekündlichen Schwankungen der unmittelbaren Kandidatenbewertungen möglichst vollständig zu erklären und dafür eine möglichst vollständige inhaltsanalytische Beschreibung des TV-Duells heranzuziehen, dann sind die getroffen Entscheidungen als pragmatische Vereinfachung nachvollziehbar (Verwendung einer statischen Transferfunktion) bzw. die einzige Möglichkeit (Formulierung von sehr umfangreichen Regressionsmodellen). Die Diskussion der Einschränkungen dieser zeitreihenanalytischen Vorgehensweise soll vielmehr verdeutlichen, warum wir im Folgenden eine andere Analysestrategie für Tests der Wirkungen *ausgewählter* Inhaltsmerkmale auf die *generelle Tendenz der Veränderung* der Kandidatenbewertung wählen. Für das Erreichen dieses Ziels erscheint es uns nicht angebracht, diese Einschränkungen in Kauf zu nehmen.

Mit der Verwendung der für das gesamte Publikum sowie Teilgruppen nach Kandidatenlager und Involvement Mittelwert-aggregierten RTR-Zeitreihen ist die *dritte* Einschränkung allerdings eine wesentliche. Damit gelten dieselben Probleme, die wir bereits ausführlich für die Peak-Spike-Analyse aggregierter RTR-Zeitreihen dargestellt haben, auch für den zeitreihenanalytischen Ansatz: Zum einen wird die vermutlich recht große statistische Unsicherheit um die aggregierten Zeitreihen ignoriert, die durch die Varianz innerhalb der zusammengefassten Gruppen und die teils sehr geringen Fallzahlen entsteht. Zum zweiten beziehen sich alle Analysen immer nur auf die Bewertung der Kandidaten durch die gebildeten Aggregate. Interpretationen auf Individualniveau sind so nicht möglich. Zusätzlich sind die inferenzstatistischen Tests, die in diesen deduktiv orientierten Analysen durchgeführt werden, in ihrer Aussagekraft beschränkt. Da das zeitreihenanalytische Vorgehen diese Probleme, die sich aus der Analyse über Personen aggregierter RTR-Daten ergeben, mit der Aussagenanalyse teilt, werden sie im Anschluss an den folgenden Abschnitt gemeinsam dargestellt.

Aussagenanalytische Ansätze

Als aussagenanalytischer Ansatz lässt sich das Vorgehen von J. Maier (2009) und Strömbäck et al. (2009) einordnen. Der Datensatz, der in diesen Analysen ausgewertet wird, gleicht einem typischen Datensatz aus einer Inhaltsanalyse. Fälle sind die Aussagen der Kandidaten, zu denen der Sprecher und verschiedene inhaltliche Merkmale erfasst werden. Über einen Zeitcode, der Anfang und Ende der Aussagen markiert, werden aus den Zeitreihen-Datensätzen der aggregierten RTR-Verläufe Informationen zur mittleren Bewertung dieser Aussage durch das gesamte Publikum bzw. Teilgruppen zugespielt. Bei J. Maier (2009) geschieht dies über die Bildung des Mittelwerts[51] über die Werte der RTR-Zeitreihen während der Aussage und den folgenden zwei Sekunden als Latenzzeit. Bei Strömbäck et al. (2009) werden die Werte der RTR-Zeitreihen zu Beginn und am Ende der Aussage in den Datensatz übernommen. Aus beiden Werten wird dann die Differenz berechnet, sodass negative Werte eine Verschlechterung der Bewertung während der Aussage, positive Werte eine Verbesserung der Bewertung während der Aussage bedeuten. Die mittlere Bewertung bzw. die Differenz wird in den statistischen Analysen als abhängige Variable, die inhaltsanalytisch erfassten Merkmale der Aussage als unabhängige Variable verwendet. Als statistische Verfahren dienen T-Tests, Varianzanalysen und Regressionsmodelle. Damit ist die Aussagenanalyse ein vergleichsweise unkompliziertes Vorgehen zur Verknüpfung von Inhaltsanalyse- und RTR-Daten. Ihre größte Einschränkung liegt wiederum in der Verwendung über die Personen aggregierter RTR-Messungen. Zudem gehen durch die zusätzliche Zusammenfassung der RTR-Zeitreihen über die Zeitgrenzen der Aussagen weitere Informationen über den zeitlichen Verlauf der Kandidatenbewertung verloren.

Konsequenzen der Aggregation der individuellen RTR-Messung über Personen für zeitreihen- und aussagenanalytische Ansätze

Gemeinsam haben beide Ansätze, dass sie die individuellen RTR-Messungen über (Gruppen von) Rezipienten zusammenfassen. Die abhängigen Variablen der Analysen sind entweder die oben ausführlich dargestellten aggregierten RTR-Zeitreihen, oder sie basieren auf diesen Zeitreihen. Daher gelten alle oben

51 Das Vorgehen bei der Datenfusion wird nicht im Detail erläutert. Nur diese Information steht zur Verfügung: „Die RTR-Bewertung einer Aussage wird hier definiert als der spontane Eindruck, den die Probanden von der Aussage selbst sowie den sich daran anschließenden zwei Sekunden hatten." (J. Maier, 2009, S. 184, Fußnote 8). Da im Folgenden mehrmals von einer durchschnittlichen Bewertung der Kandidaten die Rede ist, gehen wir davon aus, dass die Zusammenfassung über die Aussage hinweg über den Mittelwert erfolgt.

diskutierten Probleme, die sich aus der Aggregation über Personen ergeben, auch für diese Analyseverfahren. Da sie die Varianz zwischen den Personen in einer Gruppe und innerhalb einzelner Personen vernachlässigen, können sie nur Aussagen auf Aggregatniveau treffen. Annahmen über individuelle Prozesse der Duellwahrnehmung können so nicht überprüft werden. Dies ist vor allem für die Studien problematisch, die ihre Argumentation auf psychologische Theorien zur individuellen Informationsverarbeitung stützen (Nagel, 2012; Nagel et al., 2012; Strömbäck et al., 2009). Auch können Interaktionen zwischen Personenmerkmalen wie den Voreinstellungen und den Merkmalen des Debatteninhalts nur indirekt untersucht werden, indem die Ergebnisse für anhand von Personenmerkmalen gebildeten Aggregaten verglichen werden. Ein statistischer Hypothesentest über solche Interaktionen (oder auch für Haupteffekte der Personenmerkmale) kann nicht durchgeführt werden, da die individuellen Personen und ihre Eigenschaften nicht in den Datensätzen erfasst sind.

Doch auch die in den Analysen durchgeführten inferenzstatistischen Tests zum Einfluss der Merkmale des Debatteninhalts sind nicht unproblematisch. Dies ist für die Verfahren eine wesentliche Einschränkung, da sie in ihrer Analyselogik deduktiv angelegt sind und zeigen wollen, wie sich bestimmte Merkmale der Kandidatenaussagen auf ihre unmittelbare Bewertung auswirken. Ursache für die beschränkte Aussagekraft der inferenzstatistischen Tests ist das Ignorieren der bereits im Kontext der Peak-Spike-Analyse detailliert aufgezeigten Unsicherheit um die Mittelwert-Punktschätzer, aus denen die RTR-Zeitreihen bestehen. Weder die Fallzahl noch die Varianz zwischen den Personen ist in den Zeitreihen-Datensätzen enthalten und kann für die inferenzstatistischen Tests herangezogen werden.

Beim Vergleich der RTR-Mittelwerte von Sekunden bzw. Aussagen, die ein bestimmtes Merkmal aufweisen bzw. nicht aufweisen, handelt es sich um die Auswertung einer Messwiederholung, da die Bewertungen aller Sekunden bzw. Aussagen bei denselben Personen gemessen wurden. Ein angemessener Test dieses Unterschieds beruht daher nicht auf den absoluten Mittelwerten, sondern auf den individuellen Differenzen der Messungen zu allen Zeitpunkten innerhalb der Personen. Wir können das Problem, dass sich beim Test der Messwiederholungsunterschiede in diesen Verfahren auf Basis der aggregierten RTR-Messungen ergibt, leicht nachvollziehen, indem wir die einfachste Variante der Messwiederholung, in der nur zwei Messungen bei denselben Personen miteinander verglichen werden, betrachten. Der Unterschied zwischen diesen Messungen kann mit einem t-Test für verbundene Stichproben auf seine statistische Bedeutsamkeit getestet werden. Der empirische t-Wert, mit

dem ein Konfidenzintervall als Indikator für die Unsicherheit um die Differenz konstruiert werden kann, ist (Bortz & Schuster, 2010, S. 125):

$$t = \frac{\overline{d}}{SD_d}\sqrt{n} \tag{5.9}$$

mit

$$\overline{d} = \sum_i d_i / n = M_1 - M_2 \tag{5.10}$$

und

$$SD_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}} \tag{5.11}$$

wobei $\overline{d}$ der Mittelwert aller individuellen Differenzen $d_i = y_{i1} - y_{i2}$ innerhalb der einzelnen Probanden i, SD_d die Standardabweichung aller individuellen Differenzen d_i innerhalb der einzelnen Probanden i, M_1 und M_2 die Mittelwerte der individuellen Messungen y_{i1} bzw. y_{i2} und n die Zahl aller Probanden sind.

Mit den Informationen eines über Personen aggregierten Datensatzes können wir nach Formel 5.10 zwar den mittleren Unterschied $\overline{d}$ zwischen zwei RTR-Messungen bestimmen, nicht aber die in Formel 5.11 angegebene Streuung SD_d der individuellen Differenzen und die Fallzahl n der Vergleichspaare. Dafür wären die individuellen RTR-Messungen der Probanden notwendig. Und genau so, wie wir die Differenz zwischen zwei Messwiederholungen nicht ohne die Streuung der individuellen Differenzen und die Zahl der Vergleichspaare auf Signifikanz testen können, ist ein Test für mehr als zwei Messzeitpunkte ebenfalls nicht möglich.

Die Streuung, die in den Analysen über Personen aggregierter RTR-Daten berücksichtigt wird, ist lediglich die Streuung zwischen den aggregierten Bewertungen der Sekunden bzw. Aussagen. Implizit wird damit angenommen, dass die durch die individuellen Messungen verursachte Unsicherheit vernachlässigbar klein ist. Dies würde bedeuten, dass die Streuung der individuellen Differenzen sehr klein und/oder die Fallzahl der Personen sehr groß ist. Beides

ist in TV-Duell-Studien mit rezeptionsbegleitenden Messungen jedoch in der Regel nicht gegeben. Die Signifikanztests der Zeitreihen- und Aussagenanalysen vernachlässigen damit eine wichtige Quelle der statistischen Unsicherheit und müssen kritisch betrachtet werden. Sie geben nur darüber Auskunft, ob sich die Punktschätzer der Aggregate signifikant voneinander unterscheiden. Damit beziehen sich diese Inferenzen *nicht* auf die Population anderer Rezipienten desselben TV-Duells. Stattdessen wird eine Aussage darüber getroffen, ob sich die Bewertungen durch die hier gebildeten Aggregate verändern würden, wenn sich die erfassten Merkmale des Debatteninhalts anders zusammengesetzt hätten. Dieser Inferenzschluss mag für andere Forschungsziele – etwa die Untersuchung von Vermutungen, wie die Bewertungen der Kandidaten durch die untersuchten Rezipientengruppen ausgefallen wären, wenn die Kandidaten andere Inhalte geäußert hätten – durchaus interessant sein. Zur Beantwortung der Frage, ob die deskriptiv festgestellte Wirkung eines Merkmals des Debatteninhalts sich auf andere Rezipienten außerhalb der Stichprobe verallgemeinern lässt, tragen die Signifikanztests über Unterschiede zwischen Messzeitpunkten in den aggregierten Zeitreihen nichts bei. Damit ist es nicht möglich, mit diesen Tests etwas über die allgemeinen Wirkungen der untersuchten Inhaltsmerkmale auf andere Rezipienten, und in der Fortführung schon gar nicht über die Wirkung ähnlicher Inhalte in anderen Debatten auf andere Rezipienten auszusagen (vgl. dazu auch die typischen Forschungsfragen in TV-Duell-Studien, Kapitel 2).

Hinzu kommt, dass die Signifikanztests dieser Analysen eine sehr große Power besitzen, da ihre Fallzahl die Zahl der Sekunden bzw. die Zahl der Aussagen ist. Dies macht sich vor allem bei den Zeitreihenanalysen bemerkbar. Beispielsweise berichtet Nagel (2012) für ihre Regressionsmodelle Fallzahlen von bis zu n = 1799 Sekunden, die maximale Fallzahl, die dem RTR-Aggregat zugrunde liegt, ist n = 72. So entsteht die paradoxe Situation, dass in der Analyse auch verhältnismäßig kleine Unterschiede zwischen den Punktschätzern der Zeitreihe als signifikant eingestuft werden, während die Unterschiede zwischen diesen Punktschätzern wegen der Varianz der individuellen RTR-Messungen und der kleinen Fallzahlen mit einer recht große Unsicherheit behaftet sind. Hieraus ergibt sich vor allem in der Kommunikation der Ergebnisse ein Dilemma: Statistische Befunde mit einer geringen Irrtumswahrscheinlichkeit ($p < .05$) gelten in der Regel als gut abgesichert. Wenn aber eine wesentliche Quelle der statistischen Unsicherheit bereits vor der Analyse durch die Aggregation der Daten ausgeblendet wird, besteht die Gefahr, dass trotz des nominalen Unterschreitens der konventionellen Signifikanzgrenze ein Fehler erster Art besteht. In der Kommunikation der Ergebnisse muss dann klar herausgestellt werden, dass sich der Signifikanztest lediglich auf die Unterschiede zwischen

den Aggregat-Punktschätzern bezieht. Dies gilt insbesondere, wenn auf Basis der Ergebnisse ein Inferenzschluss auf Ebene der Personen gezogen werden soll, da die aus den individuellen Messungen entstehende Unsicherheit überhaupt nicht Bestandteil des statistischen Tests ist.

Wir müssen also zu dem Schluss kommen, dass die zur Analyse der Wirkung bestimmter Merkmale auf die unmittelbare Bewertung der Kandidaten eingesetzten Analyseverfahren, die auf über Personen aggregierten RTR-Messungen beruhen, eine wesentliche Schwäche aufweisen, wenn es darum geht, ihre Befunde statistisch abzusichern. Da diese Analyseverfahren in deduktiven Studien zum Test von Hypothesen über die Wirkung bestimmter Inhaltsmerkmale auf die unmittelbare Bewertung der Kandidaten eingesetzt werden, stellt ein nur beschränkt gültiger Signifikanztest ein elementares Problem dar. Dazu kommen die fehlenden Fähigkeiten, die Ergebnisse auf Individualniveau zu interpretieren und Eigenschaften der Rezipienten explizit (d.h. nicht nur durch den Vergleich verschiedener Aggregate) in die Tests einzubeziehen. Das Analyseverfahren, dass wir in Kapitel 6 vorstellen, soll geeignet sein, diese Einschränkungen zu beheben. Der wesentliche Ansatzpunkt dafür wird sein, auf die Bildung von Aggregaten der RTR-Messungen *vor* der Analyse zu verzichten.

5.3.2 Verfahren auf Basis der Aggregation über Messzeitpunkte

Neben dem verbreiteten Vorgehen, die RTR-Messungen von Personengruppen zu aggregieren, besteht auch die Möglichkeit, die einzelnen Messungen, die bei einer Person vorgenommen wurden, für diese Person zu einer oder mehreren Variablen zusammenzufassen. Die am häufigsten anzutreffende Vorgehensweise ist hier, einen Mittelwert über alle RTR-Messungen einer Person zu bilden, um so zu erfassen, wie diese Person die Kandidaten in der Debatte insgesamt bewertet hat (Bachl, 2013b; J. Maier, 2013; Maurer & Reinemann, 2009). Meist dienen die so gebildeten Variablen dazu, die unmittelbaren Kandidatenbeurteilungen als unabhängige Variablen in ein Modell zur Erklärung post-rezeptiv gemessener Einstellungen zu integrieren.

Diese Zusammenfassung sämtlicher RTR-Messungen kann aber auch als abhängige Variable verwendet werden, um herauszufinden, welche Personenmerkmale die durchschnittliche Bewertung des gesamten Duells beeinflusst haben. Da es eine solche Zusammenfassung sämtlicher Messzeitpunkte unmöglich macht, den Effekt *bestimmter* Debatteninhalte auf die Kandidatenbewertung zu untersuchen, wollen wir diese extremste Form der Aggregation über die Zeit hier nicht weiter vertiefen. Zur Analyse der Wirkungen bestimmter Inhalte können über die Zeit zusammengefasste RTR-Messungen jedoch herangezogen

werden, wenn die Zusammenfassung auf Basis der Inhalte stattfindet, die zu diesen Zeitpunkten besprochen wurden.

Einfache Beispiele für eine solche inhaltsbasierte Zusammenfassung über die Zeit sind unsere Analysen der Kandidatenbewertungen des Duells Mappus gegen Schmid in Abhängigkeit von den besprochenen Themen. Um herauszufinden, wie die Kandidaten in den zehn thematisch abgegrenzten Abschnitten der Debatte von den Zuschauern bewertet wurden, haben wir die RTR-Messungen während jedes Abschnitts für jeden Probanden zusammengefasst (Bachl, 2013a, S. 154-158). In einer detaillierteren Auswertung sind wir ebenso für die RTR-Messungen während der drei bildungspolitischen Blöcke vorgegangen (Bachl & Vögele, 2013). Nach dieser Zusammenfassung enthält der Personendatensatz zehn bzw. drei neue Variablen, die über die durchschnittliche Bewertung der Kandidaten in diesen Themenfeldern durch jeden individuellen Probanden Auskunft geben. Die Zusammenfassung ist in diesen Beispielen vergleichsweise einfach, da die Themenblöcke sequentiell nacheinander angeordnet sind. Die Inhaltsanalyse, mit der bestimmt wird, welche RTR-Messungen zusammenzufassen sind, beschränkt sich hier auf das Protokollieren der Start- und Endzeiten jedes Blocks sowie die Identifikation der Passagen, in denen die Moderatoren gesprochen haben und die daher nicht in die Mittelwerte der Kandidatenbewertung einfließen sollen.

Konzeptionell ähnlich, in der Umsetzung jedoch komplexer, ist das Vorgehen von Spieker (2011), der die RTR-Messungen während Aussagen, die eine bestimmte Relation enthielten, zusammenfasst. Dazu wird in einer standardisierten Inhaltsanalyse die Debatte zuerst in Aussagen zerteilt. Für jede Aussage werden (neben anderen, hier nicht relevanten Merkmalen) Sprecher, Relation sowie Start- und Endzeit erfasst. Auf dieser Basis werden dann alle RTR-Messungen eines Probanden, die in eine Aussage mit den Relationen Selbstpräsentation, Angriff oder Verteidigung von Merkel oder Steinmeier fallen, zu jeweils einem Mittelwert zusammengefasst. Der Personendatensatz enthält damit sechs neue Variablen, die über die durchschnittliche Bewertung der Kandidaten während ihrer Verwendung der drei Relationen informieren.

Für die statistische Auswertung der so erstellten Datensätze stehen die üblichen Verfahren zur Verfügung. Unterschiede in der durchschnittlichen Bewertung eines Themenblocks oder einer Relation in Abhängigkeit von den Voreinstellungen der Probanden können mit einfachen t-Tests für unabhängige Stichproben, Varianzanalysen, Korrelationsverfahren oder Regressionsanalysen untersucht werden. Für den Vergleich der in Messwiederholung vorliegenden Bewertungen unterschiedlicher Themen oder Relationen sind t-Tests für verbundene Stichproben oder Varianzanalysen mit Innersubjektfaktoren geeignet. Soll analysiert werden, ob es einen Interaktionseffekt zwischen einem Personen-

merkmal und einem inhaltlichen Merkmal gibt, kann z.B. eine Varianzanalyse mit dem Inhaltsmerkmal als Innersubjektfaktor und dem Personenmerkmal als Zwischensubjektfaktor verwendet werden.

Der Vorteil der Aggregation von RTR-Messungen über die Zeit ist die Möglichkeit, Personenmerkmale und Interaktionen zwischen Personen- und Inhaltsmerkmalen zur Erklärung heranzuziehen. Das Vorgehen hat jedoch auch zwei entscheidende Einschränkungen: *Erstens* kann nur eine begrenzte Zahl von inhaltlichen Merkmalen untersucht werden, da durch jedes weitere Merkmal bzw. jede weitere Merkmalskombination weitere Messwiederholungsvariablen hinzukommen. Wollten wir beispielsweise die inhaltlichen Merkmale der beiden vorgestellten Studien kombinieren, so ergäben sich für die Kombination von Themen, Sprecher und Relation $10 \times 2 \times 3 = 60$ Variablen. Auch die Information, welche RTR-Wertung ein Proband vor dem Beginn einer Aussage angibt, kann nicht sinnvoll in einem solchen Datensatz erfasst werden. Dies ist für Analysen, in denen die RTR-Werte nach Aussagen zusammengefasst werden (z.B. Spieker, 2011), die an unterschiedlichen Stellen der Debatte liegen, eine relevante Einschränkung. Es ist davon auszugehen, dass die Bewertung der jeweils vorangegangen Debatteninhalte die Bewertung des aktuellen Debatteninhalts wesentlich beeinflusst.[52]

Zweitens geht mit der Aggregation über die Zeit die intraindividuelle Varianz in den RTR-Messungen der einzelnen Probanden verloren. Damit geht zunächst einmal ein inhaltlicher Informationsverlust einher. Wir können nicht mehr unterscheiden, auf die Bewertungen welcher konkreten Inhalte des Stimulus Unterschiede zwischen den Probanden zurückgehen. Dieser Informationsverlust scheint auf den ersten Blick akzeptabel zu sein, wollen wir doch in solchen Analysen herausfinden, wie z.B. Angriffe generell – also im Mittel über alle Vorkommen – bewertet werden. Dies ist jedoch genauso ein Trugschluss wie die Annahme, dass bei der Analyse von über Personen aggregierten RTR-Daten wegen ihres Messwiederholungscharakters die Informationen des Personendatensatzes vernachlässigt werden dürften. Bereits Biocca et al. (1994, S. 41) warnen vor dem „potential for serious distortions“ bei der Analyse von über die Zeit zusammengefassten RTR-Messungen. Tabelle 5.3 verdeutlicht die Problematik anhand eines einfachen Beispiels. Dargestellt sind die mittleren RTR-Bewertungen von fünf Angriffen durch drei Probanden, sowie die mittlere Bewertung über alle Angriffe mit ihrer Standardabweichung und ihrem 95%-Konfidenzintervall.

52 Allerdings wird der Ausgangswert einer Aussage auch in den hier vorgestellten Analysen auf Aussagenebene (J. Maier, 2009; Strömbäck et al., 2009) nicht als Kontrollvariable berücksichtigt, obwohl dies in der dort gegebenen Datenstruktur möglich wäre.

Tabelle 5.3: Probleme der Zusammenfassung von RTR-Messungen über die Zeit

ID	A_1	A_2	A_3	A_4	A_5	M_A	SD_A	95%-KI$_{M_A}$
1	0	0	0	0	0	0	0	$[0;0]$
2	-10	-10	-10	-10	$+40$	0	22.4	$[-19.6;19.6]$
3	-10	$+40$	-10	-10	$+40$	10	27.4	$[-14.0;34.0]$

Anmerkungen
ID: Personen-ID; A_i: RTR-Messungen der Angriffe; M_A: mittlere Bewertung aller Angriffe; SD_A: Standardabweichung der Bewertung aller Angriffe, 95%-KI$_{M_A}$: 95%-Konfidenzintervall um M_A).

Unschwer ist zu erkennen, welchen Informationsverlust die Zusammenfassung der RTR-Bewertungen der einzelnen Angriffe zu einer einzigen Variable verursacht. Die Probanden 1 und 2 bewerten jeden Angriff unterschiedlich, haben jedoch denselben Gesamtmittelwert. Dagegen unterscheiden sich die Probanden 2 und 3 nur in der Bewertung des Angriffs 2 und stimmen bei den übrigen Angriffen überein, ihre Gesamtbewertung der Angriffe unterscheidet sich jedoch. Nur bei Proband 1 ist der Gesamtmittelwert wirklich aussagekräftig für die Bewertung der Angriffe. Die Standardabweichungen der Probanden 2 und 3 zeigen deutlich, dass der jeweilige Mittelwert nicht sehr typisch für die Bewertung der Aussagen ist. Das Konfidenzintervall in der letzten Spalte zeigt, warum auch ein statistischer Test, der nur auf die Variable des Gesamtmittelwerts eingeht, begrenzte Aussagekraft hat. In einer solchen Auswertung würden nur die Punktschätzer der Mittelwerte berücksichtigt. Nach diesen bewertet Proband 3 Angriffe positiver als die anderen beiden Probanden. Die Streuung zwischen den einzelnen Angriffen würde aber bei einer solchen Auswertung ebenso wenig berücksichtigt wie die Information, dass die Mittelwerte lediglich auf fünf bewerteten Angriffen beruhen. Die Konfidenzintervalle der Mittelwerte für alle drei Personen überlappen sich deutlich, wir können nicht sagen, dass eine der Personen die Angriffe insgesamt deutlich besser oder schlechter bewertet als eine andere. Damit sind auch die Signifikanztests der Analysen, die lediglich einen Gesamtmittelwert aus allen RTR-Messungen zu einem Merkmal des Debatteninhalts berücksichtigen, nur beschränkt aussagekräftig, da sie eine wesentliche Quelle von Variation vernachlässigen.

Analyseverfahren, die auf über die Zeit zusammengefassten RTR-Messungen basieren, haben also konzeptionell ähnliche Probleme wie die Verfahren auf Basis über Personen aggregierter RTR-Messungen. Der Fokus der Probleme verschiebt sich entsprechend des Vorgehens bei der Aggregation von den

Personen auf die Inhalte. Es können nur eine begrenzte Anzahl von inhaltlichen Merkmalen sinnvoll untersucht werden, und es gehen Informationen über die Unterschiede zwischen den Bewertungen einzelner Stimulusinhalte verloren. Damit nimmt auch die Aussagekraft der inferenzstatistischen Tests ab, da sie die Varianz zwischen den einzelnen RTR-Messungen als Quelle von Unsicherheit vernachlässigen.

5.3.3 Zwischenfazit

Deduktive Analysen unterscheiden sich von den induktiven, explorativen Analysen dadurch, dass sie Annahmen oder explizit ausformulierte Hypothesen über die Wirkung bestimmter Merkmale der Inhalte der Stimuli prüfen wollen. Sie führen dazu standardisierte Inhaltsanalysen zur Erfassung der relevanten Merkmale durch und verknüpfen diese Informationen in verschiedener Weise mit den durch RTR-Messungen erfassten Reaktionen der Rezipienten auf das Vorkommen dieser Merkmale. Unser Ergebnis, dass die Signifikanztests in den bestehenden Verfahren nur beschränkt aussagekräftig sind, da sie – egal, ob sie die RTR-Messungen über Personen oder über Messzeitpunkte zusammenfassen – jeweils eine wesentliche Quelle von Unsicherheit vernachlässigen, macht sie weitgehend ungeeignet, um das Ziel eines inferenzstatistischen Hypothesentests zu erreichen. Von der Verwendung dieser Verfahren in ihrer bestehenden Form raten wir daher ab.

Die Verfahren, die auf über Personen aggregierte RTR-Messungen zurückgreifen, müssen so erweitert werden, dass die personenbezogene Varianz in den Tests berücksichtigt wird und Personenmerkmale als erklärende Variablen verwendet werden können. Umgekehrt müssen die Verfahren, in denen die RTR-Messungen über mehrere Messzeitpunkte zusammengefasst werden, so modifiziert werden, dass die Varianz zwischen den Messungen zu verschiedenen Zeitpunkten in die Signifikanztests einfließt. Für diese notwendigen Erweiterungen stehen verschiedene analytische Modelle zur Verfügung, die sich in jüngerer Vergangenheit auch in der Kommunikationswissenschaft steigender Verbreitung erfreuen. Welche Modellklasse den größten Nutzen verspricht, hängt vom genauen Ziel der Analyse ab. Das elaborierteste Vorgehen in der Literatur zur Analyse von rezeptionsbegleitend gemessener Kandidatenbewertungen in TV-Duellen präsentiert Nagel (2012). Es ist dafür entwickelt, möglichst große Anteile der Varianz in den Kandidatenbewertungen durch möglichst viele Merkmale des Debatteninhalts zu erklären. Größter Mangel des Verfahrens ist der Rückgriff auf die aggregierten RTR-Zeitreihen, wodurch die Modellierung von Interaktionen mit Rezipientenmerkmalen sowie Interpretationen auf Individualniveau unmöglich werden. Zudem vernachlässigen

die Signifikanztests die durch die personenbezogene Varianz verursachte Unsicherheit und beziehen sich so nur auf die Punktschätzer der Zeitreihen. Das letztgenannte Problem ließe sich recht einfach mit dem für die Absicherung der Peak-Spike-Analyse entworfenen Bootstrap-Verfahren lösen. Dadurch könnten realistische Konfidenzintervalle und Signifikanztests unter Berücksichtigung der personenbezogenen Unsicherheit ermittelt werden. Allerdings würde sich so an der grundsätzlichen Modellspezifikation nichts ändern, wodurch die beiden erstgenannten Probleme ungelöst blieben.

Vielversprechender erscheint daher der Einsatz von Time-Series Cross-Sectional Modellen, wie sie z.B. von Wang und Kollegen zur Modellierung von physiologischen Messungen eingesetzt wurden (Wang, Morey & Srivastava, 2012; Wang, Solloway et al., 2012). Grundsätzlich kann in dieser Modellklasse die zeitreihenanalytische Spezifikation einschließlich des entwickelten Verfahrens zur gewichteten Modellierung der Zeitreihen der Debattenmerkmale beibehalten werden. Allerdings werden statt der aggregierten RTR-Zeitreihe sämtliche individuellen RTR-Zeitreihen als abhängige Variablen verwendet. Mit diesen Zeitreihen können dann auch weitere Personenmerkmale in das Modell eingebracht werden. Da alle individuellen Messungen im Modell berücksichtigt sind, können mit geeigneten Schätzverfahren konsistente Standardfehler und Signifikanztests ermittelt werden. Offen ist allerdings, ob sich die starke Heterogenität in den zeitlichen Assoziationen der individuellen Zeitreihen, wie sie sich bei der Prüfung der Reliabilität der individuellen Messungen zeigt (vgl. Abbildung 4.3, S. 106), mit einem halbwegs sparsamen Modell angemessen abbilden lässt.

In der vorliegenden Arbeit wollen wir jedoch einen anderen Weg beschreiten, der zu einem Verfahren führen soll, das geeignet ist, Annahmen über die Wirkung ausgewählter Merkmale des Debatteninhalts zu überprüfen. Damit können die im Folgenden vorgestellten Modelle auch als Weiterentwicklung der Aussagenanalyse und der auf über die Messzeitpunkte zusammengefassten RTR-Messung basierenden Verfahren betrachtet werden.

6 Mehrebenenmodelle der unmittelbaren Kandidatenbewertung

In diesem Kapitel zeigen wir, wie die RTR-Messungen der unmittelbaren Kandidatenbewertungen in TV-Duell-Studien so modelliert werden können, dass sowohl Personen- als auch Inhaltscharakteristika angemessen berücksichtigt werden. Dazu setzen wir Verfahren der Mehrebenenanalyse ein. Allgemein gesprochen erlaubt die Mehrebenenanalyse, die RTR-Messungen gleichzeitig als Bewertung durch individuelle Rezipienten *und* als Bewertung während bestimmter Debatteninhalte zu betrachten. Damit entfällt die problematische Entscheidung der bekannten Verfahren, eine Aggregation entweder über Rezipienten oder über Messzeitpunkte vorzunehmen, um eine Verknüpfung mit dem jeweils anderen Merkmalsträger der RTR-Messungen herzustellen. Im finalen Modell, dem kreuzklassifizierten Wachstumskurvenmodell, ist es möglich, die dynamische Entwicklung der Kandidatenbewertung in Abhängigkeit von Rezipienten- und Inhaltscharakteristika zu analysieren. Da auf eine Aggregation der Daten vor der Analyse verzichtet werden kann, wird die Unsicherheit, die durch unterschiedliche Bewertungen der individuellen Rezipienten entsteht, ebenso berücksichtigt wie die durch die unterschiedliche Bewertung der Inhalte verursachte Unsicherheit. Die Inferenzschlüsse, die auf Basis von RTR-Messungen in TV-Duell-Studien getroffen werden sollen (vgl. Kapitel 2), verdeutlichen den Bedarf nach einem analytischen Verfahren, das beide Quellen der Unsicherheit berücksichtigt. Bei der Beantwortung von allgemeinen Forschungsfragen, die sich auf die Bewertung von Kandidaten in anderen TV-Duellen durch andere Zuschauer beziehen, sind die Inhalte der untersuchten Debatte wie auch die untersuchten Versuchsteilnehmer lediglich als Stichproben aus größeren Grundgesamtheiten von (Debatten)- Inhalten und von Rezipienten zu sehen. Genau diese Logik des Ziehens von Stichproben auf zwei Ebenen – von Stimuli und Rezipienten – wird in der Mehrebenenmodellierung aufgegriffen. Sie ist damit, im Gegensatz zu den in Kapitel 5.3 vorgestellten Ansätzen der deduktiven Analyse, konzeptionell und theoretisch der Modellierung von RTR-Messungen in TV-Duell-Studien angemessen.

Das Mehrebenenmodell, in der englischsprachigen Literatur als „multilevel model“ (z.B. Gelman & Hill, 2006, S. 2), „mixed effect model“ (z.B. Pinheiro & Bates, 2000, S. 3) oder „hierarchical linear model“ (z.B. Raudenbush & Bryk,

2002, S. 5) bezeichnet, hat sich in den letzten Jahrzehnten etabliert. Dies zeigt sich nicht zuletzt in einer steigenden Zahl von einfachen wie fortgeschrittenen Lehrbüchern zu diesen Verfahren (z.B. Gelman & Hill, 2006; Goldstein, 2011; Hox, 2010; Langer, 2009; Pinheiro & Bates, 2000; Snijders & Bosker, 2011) sowie Handbüchern zu spezifischen Fragen (z.B. Hox & Roberts, 2011; de Leeuw & Meijer, 2008). Die Aufnahme entsprechender Kapitel in allgemeine Lehrbücher zur Datenanalyse lässt darauf schließen, dass die Mehrebenenanalyse zunehmend zum Kanon der (fortgeschrittenen) Methodenausbildung gehört (z.B. Field, 2013; Field, Miles & Field, 2012; Tabachnick & Fidell, 2007). Auch in der Kommunikationswissenschaft wurde bereits vor einiger Zeit in einem von Slater, Snyder und Hayes (2006) herausgegebenen Sonderheft von *Human Communication Research* auf die Relevanz des Mehrebenenansatzes hingewiesen.

Angesichts dieses großen Angebots an einführender wie auch speziellerer Literatur verzichten wir an dieser Stelle auf eine grundsätzliche Erläuterung der Mehrebenenmodellierung. Grundlagen der hier relevanten Modellklassen der Wachstumskurvenmodelle und kreuzklassifizierten Modelle erklären wir eingangs der entsprechenden Teilkapitel. Dort zeigen wir auch anhand von Anwendungsbeispielen, wie sich die allgemeine Logik der Mehrebenenmodelle auf die Analyse der mit RTR-Messungen erfassten unmittelbaren Kandidatenbewertungen in TV-Duellen übertragen lässt.

Zur Schätzung der Modelle verwenden wir das *R* Paket *lme4*.[53] Grundsätzlich steht mittlerweile eine große Zahl spezialisierter Softwarelösungen für die Mehrebenenanalyse zur Verfügung. Auch die meisten allgemeinen Statistikprogramme enthalten diesbezüglich mehr oder weniger umfangreiche Funktionen (vgl. für Überblicke z.B. Goldstein, 2011; Snijders & Bosker, 2011). Da sich die Softwareentwicklung für Mehrebenenmodelle noch in vollem Gang befindet, kann die Wahl der Software durchaus Folgen für mögliche Modellspezifikationen und Teile der Ergebnisse haben. Die Software-Lösungen gehen auf verschiedene Weise an die numerische Evaluation der Modelle heran, und sie erlauben eine mehr oder weniger flexible Spezifikation der Modellbestandteile. Wir haben uns für das Paket *lme4* entschieden, da es als freie Software kostenlos verfügbar und – im Sinne wissenschaftlicher Transparenz noch bedeutender – offen dokumentiert ist.[54] Zusätzlich spricht für *lme4*, dass das Paket bei der Modellschätzung mit umfangreichen kreuzklassifizierten Datensätzen schnell und zuverlässig arbeitet. So berichtet Luo (2013, S. 52), dass *lme4* bei der Schätzung

53 Während des Verfassens dieser Arbeit wurde die Version 1.0 des Pakets veröffentlicht (Bates, Maechler, Bolker & Walker, 2013). Die meisten Berechnungen wurden jedoch noch mit der Vorgängerversion 0.999999-0 (Bates, Maechler & Bolker, 2012) durchgeführt.

54 Statistisch interessierte Leserinnen und Leser können die eingesetzten numerischen Verfahren bei Bates (2013a, 2013b) nachlesen.

kreuzklassifizierter Modelle neunmal schneller arbeitet als die entsprechende Funktion in *SAS*. Die Geschwindigkeit ist für unsere Arbeit durchaus relevant, da die Schätzung der komplexesten hier vorgestellten Modelle auch mit *lme4* über 30 Minuten dauert.

Für das Verständnis der Ausführungen zu den Anwendungsbeispielen muss an dieser Stelle kurz erklärt werden, welche Modellannahmen wir hinsichtlich der Analyse des Debatteninhalts treffen. Wir haben uns dafür entschieden, die Merkmale des Debatteninhalts nicht als Eigenschaften der einzelnen Sekunden der Debatte zu betrachten, sondern sie übergeordneten Analyseeinheiten zuzuordnen. Aus dieser Perspektive betrachtet stellt unser Vorgehen eine Erweiterung des aussagenanalytischen Vorgehens (J. Maier, 2009; Strömbäck et al., 2009) bzw. der inhaltlichen Zusammenfassung der RTR-Messungen (Bachl, 2013a; Bachl & Vögele, 2013; Spieker, 2011) dar. Für diese Entscheidung sprechen zwei Gründe.

Erstens ist es sehr schwierig, den Prozess, wie sich die Wahrnehmung eines Merkmals in der unmittelbaren Kandidatenbewertung niederschlägt, auf Sekundenbasis zu modellieren. Nagel (2012) hat sich für ihre Zeitreihenmodelle ausführlich mit dieser Problematik beschäftigt und für das Aggregatniveau eine nach pragmatischen Gesichtspunkten taugliche Lösung gefunden. Allerdings muss dafür eine statische Transferfunktion definiert werden, der u.a. die implizite Annahme zugrunde liegt, dass die Effekte sich zu jedem Zeitpunkt der Debatte in gleicher Form in der aggregierten RTR-Zeitreihe wiederfinden (vgl. Kapitel 5.3.1). War diese Annahme für die Aggregatanalyse noch eine pragmatisch nachvollziehbare Vereinfachung, so ist ihre Übertragbarkeit auf die Analyse der individuellen RTR-Messungen zweifelhaft. Würden wir eine solche Transferfunktion auf Sekundenbasis für unsere Analyse nutzen, käme die weitere implizite Annahme hinzu, dass der Wirkungsprozess nicht nur bei jedem Vorkommen eines Merkmals, sondern auch bei jedem individuellen Rezipienten in derselben Form auftritt. Jedoch hat bereits die visuelle Betrachtung einiger individueller Verläufe verdeutlicht, dass solche homogenen Reaktionen der Probanden auf Individualebene nicht vorliegen (vgl. Kapitel 5.2). Auch die im Mittel recht geringen Korrelationen der individuellen Zeitreihen sprechen gegen diese Annahme (vgl. Kapitel 4.2.2).

Zweitens bezweifeln wir, dass unsere Inhaltsanalyse die notwendige Präzision in der sekundengenauen Identifikation der betreffenden Merkmale aufweist, um als Basis einer statisch definierten Wirkungsfunktion zu dienen. Zwar basieren auch unsere Daten auf einer sekundengenauen Inhaltsanalyse, und die Reliabilitätstests sprechen für eine akzeptable Zuverlässigkeit (Bachl, Käfferlein & Spieker, 2013a, 2013b). Da aber keine perfekte Präzision garantiert werden kann, würde der Transferfunktion, die ebenfalls nur eine vereinfachte

Annäherung an den wahren Wirkungsprozess darstellen kann, eine weitere Fehlerquelle hinzugefügt. Dem tragen wir Rechnung, indem wir die Inhalte als Eigenschaften längerer Passagen der Debatte auffassen.

Wir untersuchen damit in unseren Modellen, ob die Kandidaten während längerer Passagen in Abhängigkeit von den inhaltlichen Merkmalen, die in der Passage insgesamt vorkommen, unterschiedlich bewertet werden. Damit vermeiden wir die Illusion einer Präzision, Wirkungsprozesse über den gesamten Debattenverlauf und alle individuellen Rezipienten hinweg generalisiert auf Sekundenbasis darstellen zu können – und damit eine Scheingenauigkeit, die wir ohnehin nicht erreichen.

In den folgenden Analysen fassen wir Passagen der Debatte auf drei unterschiedliche Arten zusammen:

Erstens: Unter einem *Turn* verstehen wir die Zeitdauer, in der ein Kandidat das Wort hat, von dem Zeitpunkt, an dem ein Kandidat das Wort ergreift, bis zu dem Zeitpunkt, an dem die Sprecherrolle wechselt (Tapper & Quandt, 2006, 2010). In den Turns können wir die Bewertungen aller Ausführungen der Kandidaten erfassen.

Zweitens: Unter einer *Antwort* verstehen wir die ersten 30 Sekunden, die ein Kandidat spricht, nachdem ein Moderator eine Frage gestellt hat. Diese Definition hat den Vorteil, dass die RTR-Messungen der meisten Rezipienten einen Ausgangspunkt nahe des neutralen Nullpunkts der Skala haben. Dadurch können wir in diesen Passagen besonders gut die dynamische Veränderung der Bewertungen analysieren. Diese Auswahl der Analyseeinheit folgt auch aus dem *latched mode* der Messung mit RTR-Dials (vgl. Kapitel 3.1). Während zu späteren Zeitpunkten der Kandidatenaussagen nicht einwandfrei entschieden werden kann, ob eine Messung als Reaktion auf den gerade rezipierten Debatteninhalt oder als Folge einer bereits vor einiger Zeit erfolgten Bewertung aufzufassen ist, können wir die auf die Fragen der Moderatoren folgenden Messungen mit großer Wahrscheinlichkeit auf die neue Antwort zurückführen.

Die Beschränkung auf die ersten 30 Sekunden erfolgt aus pragmatischen Erwägungen. Es ist für Vergleiche der Veränderungen über die Antworten hinweg hilfreich, wenn diese in etwa gleich lang sind. Aus diesem Grund verzichten wir auch auf Antworten, die kürzer als 15 Sekunden sind. Zudem erreichen wir durch die Beschränkung, dass die einzelnen Antworten vergleichsweise konsistente Botschaften enthalten. Aus dieser Definition folgt, dass wir einige Teile des Duells in dieser Analyse nicht berücksichtigen. Vernachlässigt werden die Bewertungen der Kandidaten nach diesen ersten 30 Sekunden. Ebenso werden die Passagen nicht betrachtet, in denen sich die Kandidaten gegenseitig ins Wort fallen und die Sprecherwechsel damit ungeordnet verlaufen. An diesen Stellen ändern die Zuschauer ihre Bewertungen mittels RTR-Dial

häufig sehr schnell von einem Extrem in das andere. Bezogen auf alle Turns der Kandidaten wirkt sich diese Beschränkung nur unwesentlich aus. Von den insgesamt 74 Turns werden in den Antworten 68 berücksichtigt (Schmid: 34 von 36; Mappus: 34 von 38).[55] Stärker macht sich die Auswahl der ersten 30 Sekunden bemerkbar. Von den insgesamt 2926 Sekunden, in denen die Kandidaten im TV-Duell das Wort hatten, entfallen 1841 Sekunden (62%) auf die analysierten Antworten (Schmid: 63%; Mappus: 61%). Die Konsequenzen dieser Beschränkungen diskutieren wir im abschließenden Kapitel 7.

Drittens: Unter einem *Relationswechsel* verstehen wir die ersten zehn Sekunden, die auf den Wechsel der Relation in der Aussage eines Kandidaten folgen. Diese Einheit ist nicht formal, sondern inhaltlich abgegrenzt. Während eines Turns oder einer Antwort kann ein Kandidat (auch mehrmals) die Relation seiner Aussage ändern. So könnte ein Kandidat sich innerhalb eines Turns zuerst gegen Kritik an seiner Politik verteidigen, dann eigene politische Pläne für die Zukunft vorstellen und schließlich die Vorschläge der Gegenseite angreifen. Dieser Turn enthielte dann drei Relationswechsel, für die jeweils die Zeit im Duell erfasst wird. Auch in der Einheit der Relationswechsel betrachten wir nur nach ihrem Inhalt ausgewählte spezifische Ausschnitte der Debatte und ihrer Bewertung. Da wir uns zur Identifikation des Starts einer Einheit auf die Präzision der Inhaltsanalyse verlassen müssen, ist diese Einteilung sicherlich mit der größten Unsicherheit behaftet. Dem können wir lediglich teilweise entgegenwirken, indem wir nicht nur die Veränderung in der Sekunde des Wechsels, sondern auch die darauf folgenden Bewertungen modellieren.

Das Kapitel ist wie folgt aufgebaut: Im ersten Teilkapitel 6.1 stellen wir Wachstumskurvenmodelle als eine Form von Mehrebenenmodellen vor, mit der die Veränderung von in Messwiederholung bei denselben Untersuchungseinheiten erhobenen Variablen analysiert werden kann. In TV-Duell-Studien können so die Bewertungen eines Kandidaten während *einer* Antwort untersucht werden. Im Anwendungsbeispiel demonstrieren wir dies für die Bewertung von Mappus während einer ausgewählten Antwort. In Teilkapitel 6.2 führen wir das kreuzklassifizierte Modell ein, das es uns ermöglicht, gleichzeitig Merkmale der Rezipienten und des Stimulus zur Erklärung der RTR-Messungen heranzuziehen. Illustrativ analysieren wir mit dieser Modellklasse die unmittelbaren Bewertungen der Kandidaten in Abhängigkeit von den Voreinstellungen der Rezipienten und den Themen der Turns. Schließlich kombinieren wir in Teilkapitel 6.3 die beiden Modellklassen. Die Möglichkeiten des kreuzklassifizierten Wachstumskurvenmodells erklären wir an zwei Anwendungsbeispielen: Zuerst

55 Die Gesamtzahl der Turns weicht leicht von der Datenbasis in Kapitel 6.2 ab, da in diesen Analysen aus inhaltlichen Erwägungen einige Turns ausgeschlossen werden.

modellieren wir die unmittelbare Bewertung der Kandidaten als Funktion von Voreinstellungen der Rezipienten und Relationen der Antworten. Dann untersuchen wir die Reaktionen der Rezipienten auf einen Wechsel der Relation in Abhängigkeit von denselben Konstrukten.

6.1 Das Wachstumskurvenmodell

In diesem Teilkapitel erläutern wir den Nutzen von Wachstumskurvenmodellen für die Untersuchung der unmittelbaren Kandidatenbewertungen. Dazu stellen wir zuerst die grundlegende Funktionsweise der Modelle vor. Danach untersuchen wir, wie die Rezipienten Mappus während einer seiner Antworten bewerten. Wachstumskurvenmodelle ermöglichen die Analyse der dynamischen Veränderung der unmittelbaren Kandidatenbewertungen während einer Antwort. Damit kann ein grundlegendes Forschungsinteresse der TV-Duell-Forschung befriedigt werden: Wie verändert sich die Bewertung eines Kandidaten infolge einer Antwort in der Debatte, und wie können interindividuelle Unterschiede zwischen den Veränderungen durch Eigenschaften der Rezipienten – z.B. ihre Voreinstellungen – erklärt werden?

6.1.1 Grundlagen der Modellklasse

Die Modellierung intraindividueller Veränderungsprozesse mit Wachstumskurvenmodellen ist eine weit verbreitete Anwendung des allgemeinen mehrebenenanalytischen Ansatzes. Bereits Bryk und Raudenbush (1987) beschreiben, dass sich die mehrfache Messung eines Konstrukts bei denselben Untersuchungseinheiten als ein Mehrebenenmodell auffassen lässt. Entsprechend reichhaltig ist die methodologische Literatur zu dieser Modellklasse (z.B. Van Der Leeden, 1998; Peugh & Enders, 2005; Snijders, 1996), und auch fast alle der eingangs genannten einschlägigen Lehrbücher widmen ihr ein Kapitel. Schemer, Matthes und Wirth (2009) stellen dar, wie sich Wachstumskurvenmodelle zur Analyse von Medienwirkungen mit Paneldaten nutzen lassen. Zu den prominenten Anwendungen in der Kommunikationswissenschaft zählen Analysen zu Slaters (2007, S. 281) Ansatz der „Reinforcing Spirals“ (z.B. Schemer, 2012; Slater & Hayes, 2010; Slater, Henry, Swaim & Anderson, 2003).

Die Übertragung der Logik der Mehrebenenanalyse auf die Analyse von wiederholten Messungen ist einfach. Analog zur Idee der hierarchischen Datenstruktur, dass Schüler (L1) in Klassen (L2) oder Wähler (L1) in Wahlkreisen (L2) gruppiert sind, können wir sagen, dass die wiederholten Messungen (L1) in der Untersuchungseinheit (L2) gruppiert sind, bei der sie gemessen werden.

Wenn wir die mit RTR-Messungen erfassten Bewertungen der Rezipienten zu einer Antwort eines Kandidaten im TV-Duell untersuchen, sind damit die RTR-Messungen (L1) in den Rezipienten (L2) gruppiert. Abbildung 6.1 stellt diese Struktur in einem Klassifikationsdiagramm vereinfacht für die Bewertung einer fünfsekündigen Antwort durch drei Rezipienten dar.

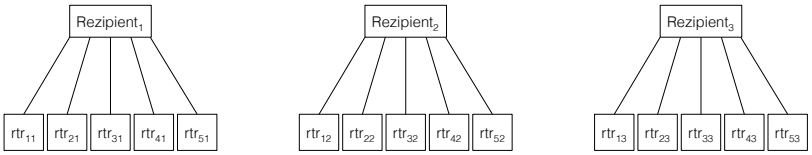


Abbildung 6.1: Klassifikationsdiagramm der Mehrebenenstruktur für eine fünfsekündige Antwort und drei Rezipienten

Die Datenstruktur berücksichtigt also, dass die RTR-Messungen eines Probanden miteinander zusammenhängen, und dass sie sich zumindest zu einem Teil mit gemessenen Charakteristika des Probanden erklären lassen. Hat ein Rezipient beispielsweise eine positive Voreinstellung zu Stefan Mappus, so liegt die Annahme nahe, dass er auch eine Antwort von Mappus während des Duells positiv beurteilt. Eigenschaften der Rezipientenebene (L2) können dazu beitragen, die individuellen RTR-Messungen (L1) zu erklären. Das einfachste Mehrebenenmodell dieser Daten, das lediglich die Zuordnung der RTR-Messungen zu einem Probanden berücksichtigt, jedoch die Veränderung über die Zeit und den Einfluss einzelner Rezipientenmerkmale vorläufig noch vernachlässigt, ist formal als die Abfolge der beiden Regressionsgleichungen definiert:[56]

$$\text{L1:} \qquad rtr_{ti} = \pi_{0i} + e_{ti} \tag{6.1}$$

$$\text{L2:} \qquad \pi_{0i} = \beta_{00} + u_{0i} \tag{6.2}$$

Auf Ebene der RTR-Messungen (L1) wird jede individuelle Messung durch Gleichung 6.1 beschrieben. Dabei ist rtr_{ti} der Wert der RTR-Messung, der zum

56 Die Notation von Mehrebenenmodellen erfolgt in der Literatur nicht einheitlich. Wir nutzen im Folgenden weitgehend die Notation von Hox (2010, Kap. 2 & 5).

Zeitpunkt t von Rezipient i abgegeben wird. Um den gewohnten Koeffizienten β für die Gleichungen auf Ebene der Personen zur Verfügung zu haben, werden die Koeffizienten auf Ebene der Messwiederholungen mit π bezeichnet. Im hier beschriebenen einfachen Fall ist π_{0i} der Koeffizient für die Regressionskonstante (im Folgenden: Intercept). Da wir die Veränderung über die Zeit hier noch außen vor lassen und nur ein Intercept geschätzt wird, entspricht π_{0i} der mittleren RTR-Bewertung der Antwort durch den Rezipienten i. Das Subskript i hinter dem Koeffizienten verdeutlicht, dass wir annehmen, dass dessen Werte zwischen den Rezipienten variieren. Sie werden daher als *Random Effects* bezeichnet. e_{ti} ist der L1-Fehlerterm, der die Abweichung der RTR-Messung zum Zeitpunkt t von der mittleren RTR-Bewertung π_{0i} durch diesen Rezipienten angibt.

Auf Ebene der Rezipienten (L2) ergibt sich nach Gleichung 6.2 die mittlere Bewertung der Antwort π_{0i} durch den Rezipienten i als Gesamtmittelwert der RTR-Bewertungen aller Rezipienten β_{00} und die rezipientenspezifische Abweichung u_{0i} von diesem Gesamtmittelwert. Der Koeffizient β variiert nicht zwischen den Rezipienten – er wird daher als *Fixed Effect* bezeichnet. Er ist der Populationsschätzer der mittleren Bewertung der Antwort.

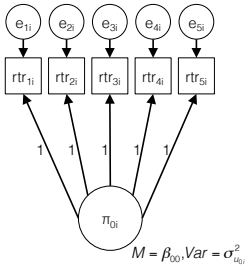
Als Gleichung für die Erklärung einer individuellen RTR-Messung rtr_{ti} durch das Mehrebenenmodell erhalten wir durch Einsetzen von Gleichung 6.2 in Gleichung 6.1:

$$rtr_{ti} = \beta_{00} + u_{0i} + e_{ti} \tag{6.3}$$

Da das Modell nur einen Intercept aufweist, wird es als *Intercept-Only-Modell* bezeichnet. Und da auch keine Einflüsse spezifischer Rezipienteneigenschaften modelliert werden, erhält es zusätzlich die Bezeichnung *unkonditionales* oder *leeres* Intercept-Only-Modell. Warum eine RTR-Messung rtr_{ti} vom Gesamtmittelwert aller RTR-Messungen β_{00} abweicht, lässt sich nach diesem Modell durch die Abweichung u_{0i} der mittleren Bewertung des Rezipienten i vom Gesamtmittelwert und die Abweichung e_{ti} der einzelnen RTR-Messung von der mittleren Bewertung des Rezipienten i erklären. Die rezipientenspezifische Varianzkomponente $\sigma^2_{u_{0i}}$ kann durch Eigenschaften der Rezipienten, wie z.B. ihrer Lagerzugehörigkeit oder ihren Voreinstellungen, erklärt werden. Die L1-Residualvarianz $\sigma^2_{e_{ti}}$ kann als Messfehler verstanden werden, den wir mit dieser Modellspezifikation nicht erklären können.

Die Bezeichnung der L1-Residualvarianz als Messfehler stellt eine Verbindung der Mehrebenenmodelle von Messwiederholungen zu ihrer Spezifikation als Strukturgleichungsmodelle her. Curran (2003) legt ausführlich dar, dass

jedes Mehrebenenmodell auch als Strukturgleichungsmodell spezifiziert werden kann. Die konzeptionelle Äquivalenz der beiden Vorgehen werden wir im Folgenden nutzen, um die unseren Analysen zugrunde liegende Logik zu erläutern.[57] Ein wesentlicher Nutzen von Strukturgleichungsmodellen besteht darin, ein latentes Konstrukt durch ein Messmodell aus mehreren manifesten Variablen zu beschreiben (z.B. Geiser, 2010, S. 41). In weiteren Analysen können die latenten Variablen anstelle der manifesten Variablen als unabhängige und/oder abhängige Variablen genutzt werden. In unserer Analyse wollen wir die unmittelbaren Bewertungen der Kandidaten während einer Antwort als latente Variable(n) spezifizieren. Die beobachteten RTR-Messungen sind ihre manifesten Indikatoren. In der klassischen Darstellungsweise der Strukturgleichungsmodelle entspricht das in Gleichung 6.3 spezifizierte leere Intercept-Only-Modell dem Messmodell in Abbildung 6.2.



Eigene Darstellung in Anlehnung an Curran (2003, S. 544)

Abbildung 6.2: Einfaches Messmodell der latenten Bewertung eines Kandidaten während einer fünfsekündigen Antwort

Die Konstante des Intercept-Only-Modells entspricht dem einfachen Mittelwert der RTR-Messungen. Aus diesem Grund sind alle Faktorladungen des so spezifizierten Messmodells auf 1 gesetzt. Jede RTR-Messung des Rezipienten i besitzt einen Fehlerterm e_{ti}, der die Abweichung der Messung vom Mittelwert π_{0i} ausdrückt. Der Rezipienten-Mittelwert π_{0i} entspricht der im Messmodell als latente Variable geschätzten Bewertung des Kandidaten durch den Rezi-

57 Um der Verständlichkeit Willen nehmen wir einige formal-statistisch nicht vollkommen korrekte Verkürzungen in der Darstellung der Übereinstimmung von Mehrebenen- und Strukturgleichungsmodellen vor. Der formale Nachweis der statistischen Äquivalenz findet sich bei Curran (2003).

pienten i. Sie hat einen Mittelwert von β_{00} und eine Varianz von $\sigma^2_{u_{0i}}$. Es ist einfach zu erkennen, dass alle in den Gleichungen 6.1 bis 6.3 definierten Koeffizienten in diesem Messmodell enthalten sind. Demzufolge können wir das unkonditionale Mehrebenenmodell als Messmodell für die latente Bewertung der Antwort durch die Rezipienten auffassen.

Entsprechend einfach ist auch die Erweiterung des Modells um Rezipienteneigenschaften als L2-Prädiktoren zur Erklärung der unmittelbaren Bewertung. Wenn wir z.B. eine Dummy-Variable *lg_map* zur Kennzeichnung, dass ein Rezipient dem Lager von Mappus angehört, berücksichtigen, erweitert sich die L2-Gleichung des Intercept-Only-Modells zu:

$$\pi_{0i} = \beta_{00} + \beta_{01} \times lg\_map_i + u_{0i} \tag{6.4}$$

Der Koeffizient β_{01} gibt in dieser Gleichung an, in welchem Umfang die mittlere Bewertung durch die Anhänger von Mappus von der Bewertung durch die übrigen Rezipienten abweicht. Denken wir wiederum in der Logik der Strukturgleichungsmodelle, so können wir uns in Abbildung 6.2 einen Pfad von der Variable *lg_map* auf die latente Variable π_{0i} mit dem Pfadkoeffizienten β_{01} vorstellen. Der Vergleich der Varianzen $\sigma^2_{u_{0i}}$ zwischen den Modellen mit und ohne den L2-Prädiktor ermittelt ein R^2-ähnliches Maß. Es gibt Auskunft darüber, welcher Anteil der personenspezifischen Varianz in der latenten Variable π_{0i} durch die Information, ob ein Rezipient dem Lager Mappus angehört, erklärt wird (z.B. Hayes, 2006; Hox, 2010).

Bisher haben wir im Intercept-Only-Modell die Veränderung der RTR-Verläufe über die Zeit vernachlässigt. Inhaltlich entspricht dieses Modell der einfachen Mittelwertbildung über alle RTR-Messungen eines Rezipienten während der Antwort. Anders als bei der Aggregation vor einer Analyse bleiben die Abweichungen hier als Varianz $\sigma^2_{e_{ti}}$ im Modell erhalten. Dadurch wird zumindest berücksichtigt, dass die mittlere RTR-Bewertung keine perfekte Beschreibung aller RTR-Bewertungen in diesem Zeitraum ist. Das Intercept-Only-Modell dient uns im Folgenden als Referenzgröße zur Einschätzung der Modellverbesserung gegenüber der Annahme eines einfachen Modells ohne Berücksichtigung der zeitlichen Dynamik. Die Veränderung über die Zeit wird in den Wachstumskurvenmodellen berücksichtigt, indem ein oder mehrere L1-Parameter basierend auf dem Messzeitpunkt der RTR-Messung in die Modellformulierung eingebracht werden. Das gebräuchlichste Wachstumskurvenmodell besteht aus zwei Parametern: einem Intercept, der angibt, wo der geschätzte RTR-Wert zum Zeitpunkt 0 (also zum Beginn der Antwort) liegt, und einem linearen Slope, der angibt, wie sich die RTR-Messungen von

diesem Punkt ab verändern. Als unkonditionales Modell ohne L2-Prädiktoren ist dieses Modell formal durch die folgenden Regressionsgleichungen definiert:

$$\text{L1:} \qquad rtr_{ti} = \pi_{0i} + \pi_{1i} \times zeit_{ti} + e_{ti} \tag{6.5}$$

$$\text{L2:} \qquad \begin{aligned} \pi_{0i} &= \beta_{00} + u_{0i} \\ \pi_{1i} &= \beta_{10} + u_{1i} \end{aligned} \tag{6.6}$$

$$\text{Gesamt:} \qquad rtr_{ti} = \beta_{00} + \beta_{10} \times zeit_{ti} + u_{0i} + u_{1i} \times zeit_{ti} + e_{ti} \tag{6.7}$$

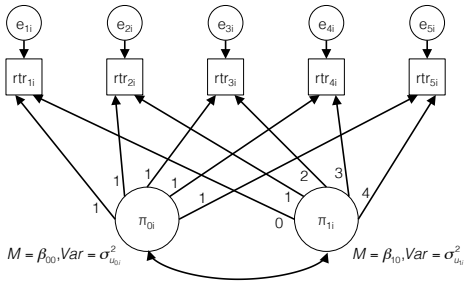
Auf Ebene der individuellen Messungen wird der zusätzliche L1-Prädiktor $zeit_{ti}$ eingeführt (Gleichung 6.5). Er gibt an, zu welchem Zeitpunkt t der Antwort die Messung beim Rezipienten i vorgenommen wird. Die Ausprägungen des Prädiktors reichen für die 30-sekündige Antwort von 0 bis 29. Der Slope-Koeffizient π_{1i} quantifiziert, um welchen Betrag sich die RTR-Messung in jeder Sekunde verändert.[58] Auch die Bedeutung des Intercept-Schätzers π_{0i} ändert sich durch diese L1-Spezifikation. Er gibt an, welcher Wert für den RTR-Verlauf des Rezipienten i zum Beginn der Antwort (also wenn $zeit_{ti} = 0$) vorhergesagt wird.

Da sowohl der L1-Intercept als auch der L1-Slope innerhalb der Rezipienten variieren, müssen auf Ebene der Rezipienten beide Parameter durch eine L2-Gleichung beschrieben werden (Gleichung 6.6). β_{00} ist dabei der Populationsschätzer für die Ausgangsbewertung zum Zeitpunkt 0, β_{10} der Populationsschätzer für die Veränderung der RTR-Bewertungen. Die Varianzen der L2-Fehlerterme $\sigma^2_{u_{0i}}$ und $\sigma^2_{u_{1i}}$ geben an, wie stark die geschätzten Ausgangswerte bzw. Veränderungen der einzelnen Rezipienten um die Populationsschätzer variieren. Sie stellen die Varianzkomponenten dar, die durch Rezipienteneigenschaften als L2-Prädiktoren erklärt werden können.

Die Analogie von Mehrebenen- und Strukturgleichungsmodellen wurde im Kontext der Wachstumskurvenmodelle bereits ausführlich diskutiert (z.B. Chou, Bentler & Pentz, 1998; Curran, 2003; Hox & Stoel, 2005; Stoel & Galindo Garre, 2011). So, wie wir oben bereits das einfache Intercept-Only-Modell als Messmodell dargestellt haben, kann auch das Wachstumskurvenmodell als ein Strukturgleichungsmodell aufgefasst werden. Es wird als latentes Wachstumskurvenmodell bezeichnet, da die Koeffizienten, die Ausgangspunkt und

58 Im Allgemeinen werden alle Regressionskoeffizienten außer der Konstanten (Intercept) als Slope bezeichnet. Im Folgenden soll sich die Bezeichnung Slope jedoch nur auf die Koeffizienten zur zeitlichen Veränderung beziehen.

Veränderung über die Zeit charakterisieren, als latente Variablen geschätzt werden. Abbildung 6.3 zeigt, wie sich das in Gleichungen 6.5 bis 6.7 beschriebene Mehrebenenmodell als latentes Wachstumskurvenmodell darstellen lässt.



Eigene Darstellung in Anlehnung an Hox und Stoel (2005, S. 3)

Abbildung 6.3: Wachstumskurvenmodell der latenten RTR-Bewertung einer fünfsekündigen Antwort

In der Begrifflichkeit der Strukturgleichungsmodelle werden hier zwei latente Variablen geschätzt, um die Bewertung der Antwort über die Zeit beschreiben zu können. Beide latente Variablen haben einen Mittelwert β_{00} bzw. β_{10}, mit dem Inferenzen über den Verlauf der Bewertungen in der Grundgesamtheit gezogen werden können, und Varianzen $\sigma^2_{u_{0i}}$ bzw. $\sigma^2_{u_{1i}}$, die durch Merkmale der Rezipienten erklärt werden können. Die Spezifikation als Modell mit einem latenten Intercept und einem latenten linearen Slope zeigt sich in der schematischen Darstellung als Fixierung der Faktorladungen. Alle Pfade vom latenten Intercept auf die manifesten RTR-Messungen sind auf 1 gesetzt, die vom latenten Slope steigen linear von 0 bis $t_{\max} - 1$ (im illustrativen Beispiel: 4; für die Analyse im nächsten Teilkapitel: 29) an.

Wenn wir Eigenschaften der Rezipienten, z.B. die Zugehörigkeit zum Lager Mappus, als L2-Prädiktoren in das Modell aufnehmen, werden die L2-Gleichungen wie folgt erweitert:

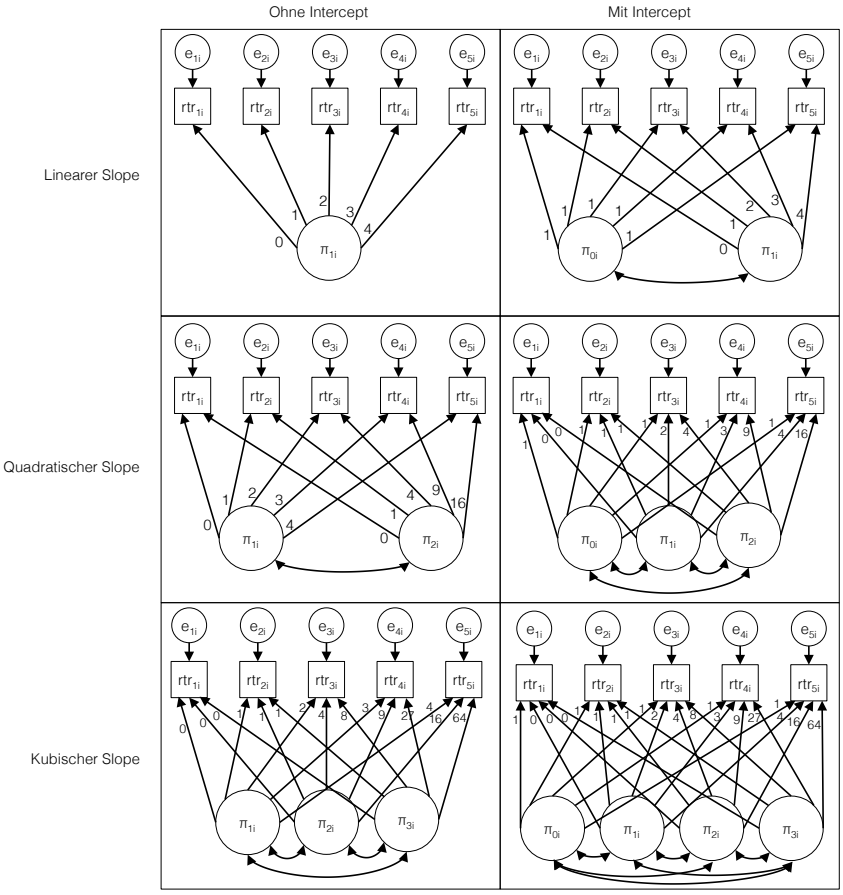
$$\pi_{0i} = \beta_{00} + \beta_{01} \times lg\_map_i + u_{0i} \tag{6.8}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} \times lg\_map_i + u_{1i} \tag{6.9}$$

Dabei quantifiziert β_{00} den vorhergesagten Ausgangswert des RTR-Verlaufs für Rezipienten, die nicht dem Lager Mappus angehören, β_{01} die Abweichung der Anhänger von Mappus von diesem Ausgangswert, β_{10} die Veränderung der Nicht-Mappus-Anhänger in einer Sekunde und β_{11} die Abweichung der Anhänger von Mappus von dieser Veränderung. Im Strukturgleichungsschema in Abbildung 6.3 können wir uns wiederum eine manifeste Variable „Lager: Mappus“ vorstellen, deren Pfade mit den Koeffizienten β_{01} und β_{11} auf die latenten Variablen π_{0i} und π_{1i} zeigen. Die Varianzaufklärung durch den L2-Prädiktor wird durch einen Vergleich der Varianzen $\sigma^2_{u_{0i}}$ und $\sigma^2_{u_{1i}}$ zwischen dem unkonditionalen und dem Modell mit L2-Prädiktor ermittelt. Da, wie im Schema des Strukturgleichungsmodells gut zu erkennen, nun zwei latente Variablen erklärt werden, erhalten wir in der Folge auch zwei R^2-ähnliche Kennzahlen: eine für die Varianzaufklärung im Intercept und eine für die Varianzaufklärung im Slope.

In der folgenden Analyse der Rezipientenreaktionen auf die Antwort von Mappus werden wir verschiedene L1-Parametrisierungen, oder, in der Begrifflichkeit der Strukturgleichungsmodelle, verschiedene Messmodelle für die individuellen RTR-Verläufe auf ihre Eignung prüfen. Neben dem Intercept-Only-Modell (vgl. Gleichung 6.3 und Abbildung 6.2) werden Wachstumskurvenmodelle verglichen, die sich in ihrer Intercept- und Slope-bezogenen Spezifikation unterscheiden. Zum einen testen wir Modelle, in denen der Intercept nicht frei geschätzt wird, sondern alle vorhergesagten RTR-Verläufe am neutralen Nullpunkt der RTR-Skala beginnen müssen. Da wir die Reaktionen in der Folge einer Frage der Moderatoren untersuchen, sollten die RTR-Bewertungen nahe dieses Ausgangswerts starten. Zum anderen testen wir Modelle, deren Slope-Parameter eine lineare bzw. nicht-lineare Veränderung der RTR-Bewertungen während der Antwort spezifizieren. Die nicht-linearen Veränderungen werden durch polynomiale Transformationen der Zeitvariable modelliert (z.B. Bortz & Schuster, 2010, S. 199-200). Tabelle 6.1 gibt einen Überblick über die L1-Gleichungen der zu testenden Modelle. Abbildung 6.4 überträgt diese Gleichungen in die Logik der latenten Wachstumskurvenmodelle. Ein weniger abstrakter Eindruck der Spezifikationen wird für die hier zu untersuchende Antwort im nächsten Teilkapitel in Abbildung 6.6 vermittelt.

Alle hier skizzierten Annäherungen an die vielfältigen Verläufe der individuellen RTR-Messungen sind sehr stark vereinfacht und nehmen dadurch auch statistische Fehler bewusst in Kauf. Das Intercept-Only-Modell vernachlässigt Veränderungen der Bewertungen während der Antwort – sowohl die Veränderung der mittleren Bewertungen als auch die Veränderung der Varianz um die mittlere Bewertung herum. Die Modelle, die keinen frei geschätzten Intercept enthalten und damit die geschätzten Verläufe durch den Nullpunkt der



Anmerkung
Natürlich sind die Modelle mit höheren Polynomen für eine fünffache Messwiederholung deutlich überparametrisiert. Die Darstellung ist dementsprechend illustrativ zu verstehen.

Abbildung 6.4: Überblick über die L1-Spezifikationen, dargestellt als latente Wachstumskurvenmodelle für eine fünfsekündige Antwort

Tabelle 6.1: Überblick über die L1-Spezifikationen der Wachstumskurvenmodelle

	Intercept	
Slope	nein	ja
linear	$\pi_{1i} \times zeit_{ti} + e_{ti}$	$\pi_{0i} + \pi_{1i} \times zeit_{ti} + e_{ti}$
quadratisch	$\pi_{1i} \times zeit_{ti} + \pi_{2i} \times zeit_{ti}^2 + e_{ti}$	$\pi_{0i} + \pi_{1i} \times zeit_{ti} + \pi_{2i} \times zeit_{ti}^2 + e_{ti}$
kubisch	$\pi_{1i} \times zeit_{ti} + \pi_{2i} \times zeit_{ti}^2 + \pi_{3i} \times zeit_{ti}^3 + e_{ti}$	$\pi_{0i} + \pi_{1i} \times zeit_{ti} + \pi_{2i} \times zeit_{ti}^2 + \pi_{3i} \times zeit_{ti}^3 + e_{ti}$

Anmerkung
Dargestellt ist nur die rechte Seite der L1-Gleichung. Da in allen Gleichungen die individuelle RTR-Messungen des Rezipienten i zum Zeitpunkt t beschrieben werden, ist auf der linken Seite jeweils rtr_{ti} zu ergänzen.

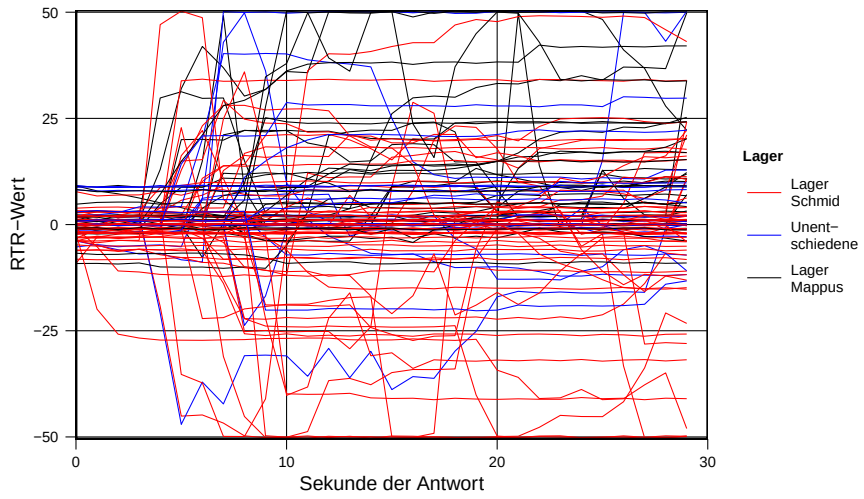
RTR-Skala zwingen, vernachlässigen, dass auch um die geschätzten Ausgangswerte herum Varianz besteht. Die Spezifikationen mit linearen Slope-Termen vernachlässigen schließlich, dass sich die beobachteten Verläufe nicht durch kontinuierliche Veränderungen pro Sekunde beschreiben lassen und wegen der Grenzen der RTR-Skala nicht stetig linear weitergeführt werden. Es wird also auch zu evaluieren sein, welche dieser Einschränkungen zu welchen Folgen für die Modellschätzung führen.

6.1.2 Bewertung von Mappus während einer Antwort

In den folgenden Abschnitten demonstrieren wir die Modellierung der Bewertung von Stefan Mappus während einer Antwort. Zuerst zeigen wir die Modellgüte der beschriebenen Parametrisierungen des L1-Wachstumskurvenmodells. Danach erklären wir die Bewertungen in Abhängigkeit von einigen Rezipientencharakteristika. Als Beispiel ziehen wir einen Angriff von Mappus auf die finanz- und steuerpolitischen Pläne der Oppositionsparteien heran. Abbildung 6.5 gibt einen Überblick über die individuellen Bewertungen während der Antwort.[59] Berücksichtigt werden in dieser Analyse nur Bewertungen, deren Wert zu Beginn der Antwort nahe des neutralen Skalenmittelpunkts

59 Der Kontext der Antwort und die aggregierten Bewertungen durch die Anhänger von Mappus und Schmid sowie die Unentschiedenen können dem wörtlichen Transkript der Debatte entnommen werden (Bachl, Käfferlein, Krafft, Schmalz & Vögele, 2013).

liegen. RTR-Verläufe, die außerhalb des Korridors $[-9; 9]$ beginnen, werden ausgeschlossen. Damit liegen 136 gültige RTR-Verläufe vor.



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus) durch n = 136 Rezipienten.
Wortlaut der Antwort: Zunächst einmal wüsste ich nicht, wo man bei der EDV-Ausstattung des Landes eine halbe Milliarde Euro einspart. Und was Herr Schmid gerade nicht erwähnt hat, ist, dass er sehr wohl eine Erhöhung der Grunderwerbssteuer vorschlägt. – Unterbrechung Schmid: Nein, das habe ich nicht. – Doch, das haben Sie mit den Grünen gemeinsam vorgestellt, auf 4,5 Prozentpunkte. Erst einmal stimmt's schon deshalb nicht, weil davon 55 Prozent der Mehreinnahmen an die Kommunen gehen. Die fließen also gar nicht in den Landeshaushalt. Und zum Zweiten will ich halt gerade nicht, dass man zuerst große Versprechungen macht und dann nachher bei den Menschen wieder Steuererhöhungen holt.

Abbildung 6.5: Beobachtete Bewertungen von Mappus während der Antwort

Wie bereits im vorangegangenen Kapitel demonstriert, besteht eine erhebliche Varianz zwischen den Verläufen der individuellen Bewertungen. Zwar ist unter den negativen Bewertungen ein Muster nach den Lagern zu erkennen – Ablehnung kam nur von Anhängern der Oppositionsparteien sowie einigen Unentschiedenen. Zustimmung gab es für die Antwort jedoch in allen drei Lagern. Schließlich zeigt sich auch wieder, dass ein großer Teil der Rezipienten in diesen ersten 30 Sekunden der Aussage noch gar keine Bewertung abgibt.

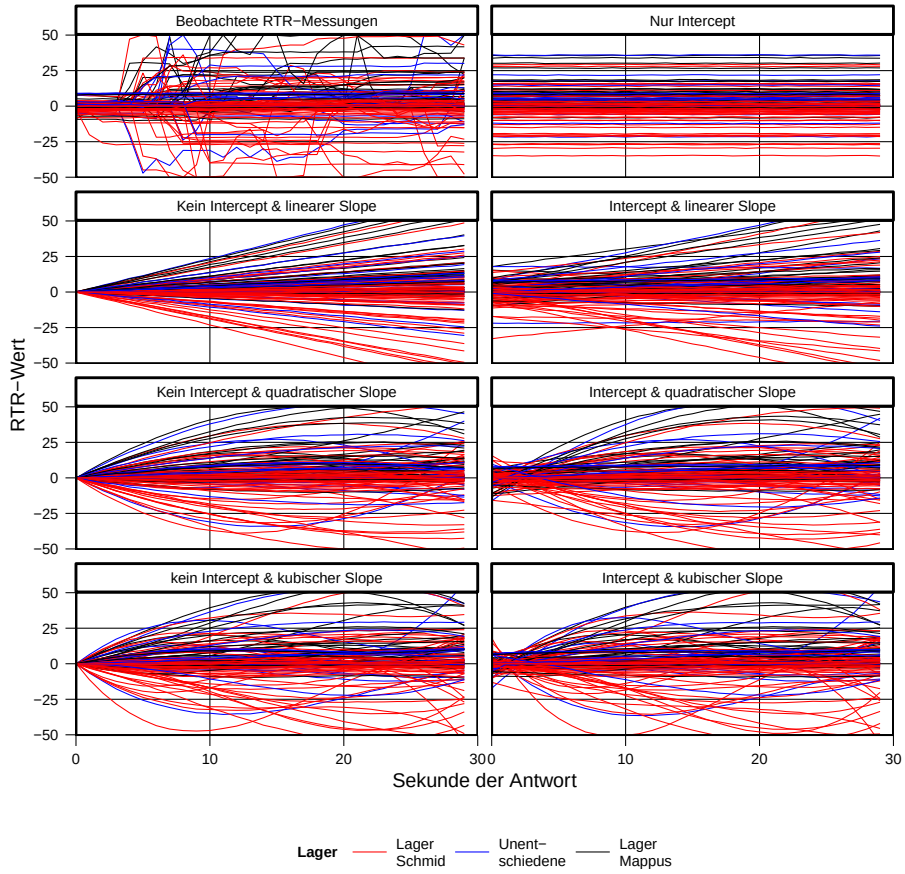
L1-Modellspezifikation

Die individuellen Bewertungen der Antwort modellieren wir nun mit unterschiedlich parametrisierten Wachstumskurvenmodellen. Abbildung 6.6 gibt einen Überblick über die durch diese Modelle vorhergesagten individuellen Verläufe. Links oben sind zum Vergleich noch einmal die beobachteten RTR-Messungen dargestellt. Rechts oben finden sich die vorhergesagten Werte durch das Intercept-Only-Modell. Hier wird die dynamische Veränderung ausgehend vom neutralen Skalenmittelpunkt nicht modelliert. Für jede Bewertung eines Probanden wird stattdessen derselbe RTR-Wert vorhergesagt. So wird klar, dass dieses Modell konzeptionell der Berechnung eines einfachen Mittelwerts über die Bewertungen in den 30 Sekunden durch einen Probanden entspricht.

Die übrigen Facetten zeigen die vorhergesagten Werte durch die im letzten Teilkapitel formal definierten L1-Wachstumskurven mit unterschiedlichen Parametrisierungen. In der linken Spalte sind die Modelle abgebildet, die auf die Schätzung eines Intercepts verzichten und damit erzwingen, dass alle vorhergesagten RTR-Verläufe am Nullpunkt der Skala beginnen. Die in der rechten Spalte abgebildeten Modelle erlauben einen frei geschätzten Intercept, was sich in einer interindividuellen Varianz in den geschätzten Verläufen bereits zu Beginn der Antwort niederschlägt.

In den Zeilen finden sich von oben nach unten Slope-Parametrisierungen mit zunehmender Komplexität: der lineare Slope als einfache Regressionsgerade sowie eine quadratische und eine kubische Parametrisierung als polynomiale Regressionen zweiter und dritter Ordnung. Vergleichen wir die vorhergesagten mit den beobachteten Werten, so bilden die nicht-linearen Modelle einige Charakteristika der beobachteten Daten etwas besser ab. Durch den quadratischen (und kubischen) Term können sie berücksichtigen, dass die Bewertung im Zeitverlauf nicht immer weiter steigt, sondern an einem gewissen Punkt die Grenzen der RTR-Skala erreicht sind. Auch gelingt es ihnen teilweise, Verläufe zu beschreiben, die keine monotonen Veränderungen aufweisen, sondern beispielsweise im Zeitverlauf zuerst ansteigen, dann aber wieder absinken.

Eine weitere Konsequenz der polynomialen Modellierung der Wachstumskurven offenbart der Vergleich der Modelle mit und ohne Intercept. Die Aufnahme eines quadratischen (und kubischen) Terms verringert die Varianz der Intercepts zwischen den Rezipienten im Vergleich zum linearen Wachstumskurvenmodell. Die nicht-linearen Modelle mit und ohne Intercept sind sich deutlich ähnlicher als die entsprechenden linearen Modelle. Im linearen Modell mit Intercept weichen die vorhergesagten Werte zum Start der Antwort stärker voneinander ab als die beobachteten Werte. Dies ist der Einschränkung geschuldet, dass im einfachen, linearen Modell mit Intercept nur zwei



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus) durch n = 136 Rezipienten. Links oben: Beobachtete Bewertung; Rechts oben: Intercept-Only-Modell; Folgend: Wachstumskurvenmodelle mit verschiedener Parametrisierung von Intercept und Slope.

Abbildung 6.6: Vorhergesagte Bewertungen von Mappus während der Antwort

Parameter optimiert werden können, um eine gute Passung zum beobachteten RTR-Verlauf zu erreichen: der Schnittpunkt mit der y-Achse (Intercept) und die Steigung von diesem Punkt (Slope). Wenn ein Proband seine Bewertung bereits zu einem frühen Zeitpunkt der Antwort verändert und sie dann über einen längeren Zeitraum beibehält, so wird die Schätzung genauer, wenn die Regressionsgerade an einem der späteren Bewertung näheren Punkt startet, statt nahe des Skalenmittelpunkts.

Hierdurch ändert sich die Interpretation des Intercept-Parameters: Er kann als ein latenter Ausgangspunkt der unmittelbaren Kandidatenbewertung aufgefasst werden. In ihm spiegeln sich starke Voreinstellungen gegenüber dem Kandidaten und seinen in der Antwort vertretenen Positionen wider. Wenn z.B. ein Rezipient eine gute Meinung von Mappus hatte und insbesondere seine Konzepte in der Finanz- und Steuerpolitik positiv bewertete, dann ist es wahrscheinlich, dass er, sobald Sprecher und Thema der Antwort zu erkennen waren, sehr schnell eine positive Bewertung abgab. Dazu musste er gar nicht abwarten, was Mappus genau zu diesem Thema zu sagen hatte – alleine der Sprecher und das Thema reichen aus, um die bereits bestehende Einstellung abzurufen. Diese Interpretation lässt sich gut mit verschiedenen Modellen zur Informationsverarbeitung bei der Bildung von (politischen) Urteilen verbinden, beispielsweise mit den Routen der Informationsverarbeitung im heuristisch-systematischen Modell (Chaiken, 1980) oder im Elaboration-Likelihood-Modell (Petty & Cacioppo, 1986). Speziell im Kontext der politischen Kommunikation ist auch die Unterscheidung zwischen „On-Line Versus Memory-Based Process Models of Political Evaluation“ (Lavine, 2002, S. 225) zu nennen. Effekte auf den Intercept würden demnach dafür sprechen, dass die Debatteninhalte nicht systematisch evaluiert werden, bevor eine Bewertung abgegeben wird, sondern als Cues dienen, die bestehende Einstellungen der Rezipienten abrufen. Wenn wir dieses Muster in den Bewertungen während vieler Antworten in den kreuzklassifizierten Wachstumskurvenmodellen systematischer untersuchen (Kapitel 6.3), zeigt sich der Vorteil einer differenzierten Interpretation von Intercept- und Slope-bezogenen Koeffizienten noch deutlicher.

Schließlich müssen wir noch beachten, wie gut die Entwicklung der interindividuellen Varianz im Zeitverlauf durch die verschiedenen Spezifikationen abgebildet werden kann. Bereits aus einer einfachen Betrachtung der beobachteten RTR-Verläufe wird deutlich, dass die Varianz zwischen den individuellen RTR-Verläufen im Laufe der Antwort größer wird (technisch: Heteroskedastizität, vgl. ausführlicher Kapitel 6.4). Im Intercept-Only-Modell, das die Zeit nicht als Prädiktor enthält, wird von einer im Zeitverlauf konstanten Varianz der Messungen ausgegangen. Die Modelle ohne Intercept nehmen spezifikationsbedingt fälschlicherweise an, dass zum Beginn einer Antwort keinerlei Varianz

zwischen den Rezipienten besteht. Um diese Fehlspezifikation auszugleichen, müssen die Verläufe in der Folge stärker ansteigen, was am Ende der Antworten zu einer Überschätzung der Varianz führt. In dieser Hinsicht sind nur die Modelle mit Intercept und einem oder mehreren Slope-Koeffizienten korrekt spezifiziert. Sie berücksichtigen, dass bereits zu Beginn der Antwort Varianz existiert, die im Verlauf der Antwort weiter wächst. Die Folgen der mehr oder weniger adäquaten Repräsentation der über die Zeit variablen Varianz zeigen sich später in den Konfidenzintervallen der von den Modellen vorhergesagten L2-Verläufe (vgl. z.B. Abbildung 6.8, S. 200). Die Vorhersagen des Intercept-Only-Modells sind zu Beginn der Antwort zu unpräzise und am Ende der Antwort zu präzise, für die Modelle ohne Intercept gilt das Umgekehrte. Nur die Modelle mit Intercept- und Slope-Parametern können zu jedem Zeitpunkt der Antwort eine Schätzung mit einer der Varianz angemessenen Konfidenz abgeben.

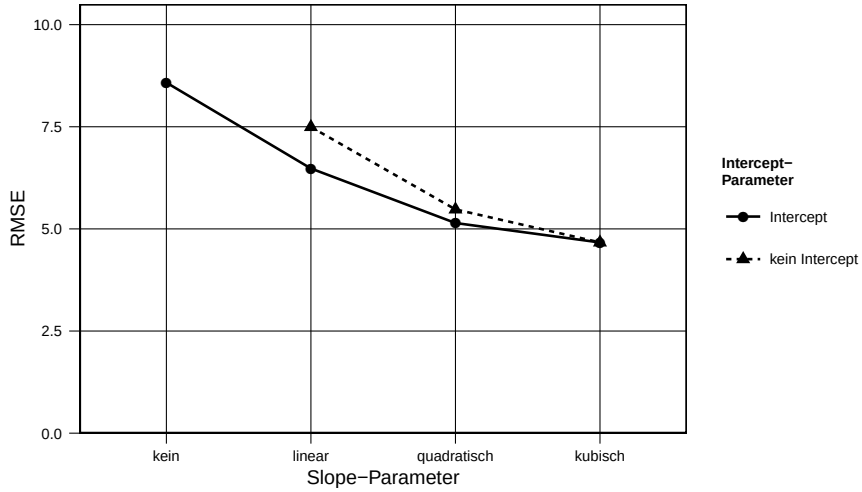
Die visuellen Vergleiche zeigen, dass die Modellierung als Wachstumskurve immer nur eine mehr oder weniger gute Annäherung an die empirische Realität sein kann. Um zu entscheiden, welche L1-Spezifikation für weitere Analysen am besten geeignet ist, müssen aus datenanalytischer Perspektive zwei Kriterien gegeneinander abgewogen werden (Hox, 2010, S. 104-105): Einerseits sind nach dem Prinzip der Sparsamkeit („parsimony“) Modelle zu bevorzugen, die zur Beantwortung der Forschungsfrage mit möglichst wenigen Parametern auskommen. Aus datenanalytischer Perspektive ist eine Orientierung an diesem Prinzip geboten, um eine Überanpassung („overfitting“) des Modells an die vorliegenden Daten zu vermeiden. Eine Überanpassung würde dazu führen, dass sich die anhand der vorliegenden Daten generierten Modelle nicht auf andere Datensätze übertragen lassen. Aus inhaltlicher Perspektive ist ein sparsames Modell von Vorteil, da es sich einfacher interpretieren lässt. Bereits in einer einfachen polynomialen Regression fällt es schwer, die Koeffizienten direkt zu interpretieren. In einem Mehrebenenmodell, in dem die Parameter der polynomialen Terme auf L1-Ebene innerhalb der L2-Einheiten variieren, kommen neben den Fixed-Effect-Koeffizienten der Polynome ebenso viele Random-Effect-Varianzen hinzu. Wenn in einem zweiten Analyseschritt L2-Merkmale zur Erklärung der interindividuellen Unterschiede in das Modell eingebracht werden, muss für jede Random-Effects-Varianz die Varianzreduktion geprüft und interpretiert werden. In den hier präsentierten kubischen Modellen entspricht das der Interpretation von jeweils drei bzw. vier (inkl. Intercept) Fixed-Effect-Koeffizienten und Varianzkomponenten.

Andererseits sollen Modelle so spezifiziert sein, dass sie möglichst große Anteile der Varianz in der abhängigen Variable berücksichtigen und erklären können. In der einfachen Regression entspricht dies der Auswahl von Mo-

dellen, die ein möglichst hohes R^2 erreichen. Für die hier vorgeschlagenen Wachstumskurvenmodelle ist diese Entscheidung etwas komplexer, da auf L1 die beobachteten Messungen durch die individuellen Wachstumskurven abgebildet und erst auf L2 die Parameter der individuellen Wachstumskurven durch Merkmale der L2-Einheiten erklärt werden. Die L1-Residuen (also die Abweichungen zwischen den individuellen beobachteten RTR-Verläufen und den durch die individuellen Wachstumskurven vorhergesagten Werte) können damit in den folgenden Analyseschritten nicht mehr erklärt werden. Nach dieser Überlegung sind Modellspezifikationen zu bevorzugen, die die L1-Residuen minimieren und somit möglichst große Teile der in den Messungen enthaltenen Informationen einer Erklärung durch L2-Prädiktoren zugänglich machen. Oder einfacher ausgerückt: Die durch die Modelle vorhergesagten RTR-Verläufe sollten den empirisch beobachteten RTR-Verläufen möglichst ähnlich sein.

Um entscheiden zu können, welche Spezifikationen der Wachstumskurven ein vorteilhaftes Verhältnis von Sparsamkeit und Datenpassung bieten, sind in Abbildung 6.7 die Vorhersagefehler der Modelle in Abhängigkeit von ihren Intercept- und Slope-Parametrisierungen dargestellt. Der Vorhersagefehler wird durch den Root Mean Squared Error (RMSE), die Wurzel der mittleren quadrierten Abweichung der vorhergesagten von den beobachteten RTR-Messungen, quantifiziert.

In der Darstellung der Vorhersagefehler lassen sich drei wesentliche Punkte erkennen. *Erstens*: Die Modellpassung der Wachstumskurvenmodelle ist in jeder Parametrisierung besser als die Passung des Intercept-Only-Modells, das die zeitliche Dynamik vernachlässigt und nur den Mittelwert aller RTR-Messungen eines Probanden erfasst. Selbst der Vorhersagefehler des zweiten Modells mit nur einem Parameter, einem linearen Slope durch den Nullpunkt der Y-Achse, passt wesentlich besser zu den beobachteten Verläufen. *Zweitens*: Mit der Anzahl der Parameter steigt auch die Modellpassung, der relative Nutzen zusätzlicher Parameter nimmt jedoch in den komplexeren Modellen ab. Während die quadratischen Modelle eine wesentliche Verbesserung gegenüber den linearen Modellen erreichen, fällt der Vorteil der kubischen gegenüber den quadratischen Modellen nur vergleichsweise gering aus. *Drittens*: Nur in der linearen Spezifikation gibt es einen wesentlichen Unterschied zwischen den Modellen mit und ohne Intercept. Sobald eine nicht-lineare Spezifikation des Wachstumsparameters gewählt wird, bringt die zusätzliche Aufnahme eines Intercepts kaum noch eine Verringerung des Vorhersagefehlers. Dies lässt sich auch im Vergleich der vorhergesagten Verläufe in Abbildung 6.6 ablesen. Die Formen der durch nicht-lineare Wachstumskurven geschätzten Verläufe mit und ohne Intercept sind einander sehr ähnlich.



Anmerkung
RMSE: Wurzel der mittleren quadrierten Abweichung der vorhergesagten von den beobachteten RTR-Messungen.

Abbildung 6.7: Vorhersagefehler der Wachstumskurvenmodelle

Nach einer Abwägung von Datenpassung und Sparsamkeit, aber auch der inhaltlichen Interpretierbarkeit der Modelle halten wir zwei Spezifikationen für gut geeignet:

- Das Modell *mit Intercept und einem linearen Slope* weist im Vergleich mit den weniger komplexen Modellen einen akzeptablen Vorhersagefehler auf. Weitere Analysen legen nahe, dass der Vorhersagefehler in etwa entsprechend der im Zeitverlauf zunehmenden Varianz größer wird und damit zu jedem Zeitpunkt der Antwort eine angemessene Schätzung erlaubt. Zudem zeichnet sich das Modell durch eine gute Interpretierbarkeit aus, bei der den beiden Parametern distinkte Bedeutungen zugeschrieben werden können. Der latente Intercept wird von schnellen und starken Reaktionen verändert, die als Indikatoren für eine heuristische, von Cues geprägte Bewertung der Kandidaten gelten können. Der latente lineare Slope quantifiziert dagegen stärker die Bewertung als Ergebnis einer kognitiven Verarbeitung der Inhalte der Antwort. Somit bietet dieses

Modell eine gute Rückbindung zu theoretischen Konzepten der Informationsverarbeitung.

- Das Modell *ohne Intercept und einem quadratischen Slope* passt besser zu den Daten als eine lineare Modellierung. Die zusätzliche Aufnahme eines Intercepts in dieses Modell bringt nur eine kleine Verbesserung, ebenso die Aufnahme eines zusätzlichen kubischen Terms. Beide ausgewählten Modelle besitzen damit nur zwei Parameter, was die Gefahr einer Überanpassung verringert. Die Modellierung des nicht-linearen Wachstums durch Polynome hat jedoch den Nachteil, dass die einzelnen Parameter nicht nach dem Beispiel der erstgenannten Spezifikation direkt interpretiert werden können.

Außerdem führen wir die folgenden Analysen mit zwei weiteren Spezifikationen durch:

- Anhand eines Vergleichs mit dem *Intercept-Only-Modell* prüfen wir, welche Vorteile die Modellierung von Wachstumskurven gegenüber der einfachen Zusammenfassung als mittlere Bewertung während der gesamten Antwort hat.
- Trotz der relativ großen Vorhersagefehler ziehen wir das Modell heran, das *nur einen linearen Slope* enthält. Damit prüfen wir, ob diese einfachste Form des Wachstumskurvenmodells inhaltlich zu ähnlichen Ergebnissen führt. Wenn diese Modellierung mit komplexeren Spezifikationen konsistente Befunde liefert, so hätte sie für Anwendungen, die vor allem eine einfache Kommunizierbarkeit der Ergebnisse erfordern, einen gewissen Reiz.

L2-Modellspezifikation

Im nächsten Analyseschritt erklären wir die interindividuellen Unterschiede zwischen den durch die Wachstumskurven abgebildeten RTR-Verläufen mit Eigenschaften der Rezipienten. Dazu präsentieren wir zwei Analysen. Als erstes berücksichtigen wir lediglich den Faktor „Lagerzugehörigkeit“ als unabhängige Variable. Dieses Modell entspricht einer Übertragung der in der aggregationsbasierten Verfahren verwendeten Aufteilung der Probanden nach ihrer Lagerzugehörigkeit. In den bekannten Verfahren ist dieses einfache Modell notwendig, da die Aggregate nur auf Grundlage kategorial skalierter Merkmale der Rezipienten gebildet werden können (vgl. Kapitel 5). In einer

weitergehenden Analyse nutzen wir die Möglichkeit der Wachstumskurvenmodelle, auch mehrere und quasi-metrisch skalierte L2-Prädiktoren zur Erklärung der RTR-Bewertungen heranzuziehen. So können wir demonstrieren, dass der hier präsentierte Analyseansatz nicht nur im Vorteil ist, da er den Informationsverlust der Aggregation vermeidet, sondern auch, da er eine flexiblere Modellierung des Einflusses mehrerer und informationsreicherer Personenvariablen erlaubt.

Bei der Formulierung der L2-Spezifikationen gehen wir immer nach demselben Prinzip vor. Zuerst schätzen wir unkonditionale Modelle, die außer den Fixed Effects der Wachstumskurvenparameter keine L2-Parameter enthalten. Dann bringen wir in weiteren Schritten Rezipientenmerkmale in das Modell ein. Durch den Vergleich mit dem unkonditionalen Modell bzw. den Modellen mit weniger L2-Prädiktoren prüfen wir, ob die neuen Prädiktoren das Modell verbessern. Alle Analysen werden für die oben beschriebenen vier Modellierungen der individuellen RTR-Messungen durchgeführt. Die Schätzung mehrerer L2-Modellspezifikationen für unterschiedliche L1-Wachstumskurven führt zwangsläufig dazu, dass im Folgenden eine große Zahl von Modellen präsentiert werden muss. Um die Darstellung bestmöglich zu systematisieren, führen wir eine Notation zur Modellbezeichnung ein. Die Parametrisierung der Wachstumskurven auf L1 werden wie folgt bezeichnet:

- Modell M0: Intercept-Only-Modell ohne Wachstumsparameter
- Modell M1: Wachstumskurvenmodell ohne Intercept mit linearem Slope
- Modell M2: Wachstumskurvenmodell mit Intercept und linearem Slope
- Modell M3: Wachstumskurvenmodell ohne Intercept mit quadratischem Slope

Den Ausbau der L2-Spezifikation innerhalb dieser Modelle machen wir durch eine zweite Zahl kenntlich, die in der Bezeichnung der Modelle nach einem Punkt (.) erfolgt. Dabei steht das Modell M.0 für das leere Modell ohne Personenmerkmale, beginnend mit M.1 aufsteigend werden die komplexer werdenden Modellierungen mit zusätzlichen Personenmerkmalen bezeichnet. So ist z.B. Modell M2.0 das leere Modell, in dem die individuellen RTR-Messungen durch einen Intercept und einen linearen Slope beschrieben werden. In Modell M2.1 wird das Personenmerkmal Lagerzugehörigkeit berücksichtigt. Ein Vergleich der Modelle M2.0 und M2.1 zeigt uns, ob die Lagerzugehörigkeit zur Erklärung der individuellen Wachstumskurvenparameter beiträgt.

Effekte der Lagerzugehörigkeit Wir beginnen mit einer einfachen Modellierung, in der wir die unkonditionalen Modelle um den L2-Prädiktor Lagerzugehörigkeit erweitern. An diesen einfachen Modellen erläutern wir ausführlicher die einzelnen Schritte der Modellformulierung, Modellprüfung und Interpretation. Dabei soll auch die Hypothese getestet werden, dass die Lagerzugehörigkeit einen Einfluss auf die unmittelbare Bewertung von Mappus während der Antwort hat. Wir erwarten auf Basis des Forschungsstands (vgl. Kapitel 3.4), aber auch auf Grundlage der visuellen Inspektion der individuellen RTR-Messungen (vgl. Abbildung 6.5) und der L1-Wachstumskurven (vgl. Abbildung 6.6):

> **H1**: Die unmittelbare Bewertung von Mappus während seiner Antwort unterscheidet sich nach der Lagerzugehörigkeit der Rezipienten.

Als einzige Personenvariable wird der dreistufige Faktor Lagerzugehörigkeit eingebracht. Von den 136 Rezipienten, die für die Analyse dieser Antwort gültige RTR-Messungen aufweisen, gehören 37 (27%) zum Lager Mappus, 71 (52%) zum Lager Schmid und 28 (21%) zu den Unentschiedenen. Im Regressionsmodell ist der Faktor durch zwei Dummy-Variablen „Lager: Mappus“ und „Lager: Schmid“ repräsentiert, die Unentschiedenen bilden die Referenzausprägung.

Tabelle 6.2 zeigt die Vergleiche zwischen den leeren Modellen (M.0) und den Modellen mit Lagerzugehörigkeit als L2-Prädiktor (M.1).[60] In den ersten beiden Zeilen wird das Informationskriterium *AICc* berichtet. Das *AIC* gibt Auskunft darüber, wie sich die Modellpassung (ausgedrückt durch die Deviance) unter der Berücksichtigung der Steigerung der Modellkomplexität durch die Aufnahme zusätzlicher Parameter verändert. Die Variante *AICc* korrigiert

60 Es ist zu beachten, dass für den Vergleich von Modellen mit unterschiedlichen Fixed Effects anhand der Deviance bzw. Deviance-basierter Maße wie dem *AICc* die Ergebnisse einer Maximum-Likelihood-Schätzung (ML) verwendet werden müssen (z.B. Hox, 2010, S. 51). Da wir für die Schätzung der Fixed Effects die Restricted-Maximum-Likelihood-Schätzung (REML) bevorzugen (z.B. Hox, 2010, S. 40-42), werden die Modelle zum Vergleich dieser Maße mit der ML-Methode neu geschätzt.

zusätzlich für die Größe der Stichprobe.[61] Das Modell, das einen niedrigeren *AICc*-Wert aufweist, ist zu bevorzugen. In allen L1-Spezifikationen liegt der *AICc*-Wert des Modells mit der Lagerzugehörigkeit als L2-Prädiktor unter dem Wert des leeren Modells. Daraus können wir schließen, dass die Aufnahme des Prädiktors die Modellpassung stärker verbessert, als dass sie die Komplexität des Modells erhöht.

Tabelle 6.2: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Lagerzugehörigkeit

	Modell 0	Modell 1	Modell 2	Modell 3
$AICc_{M.0}$	29810	28758	28001	26784
$AICc_{M.1}$	29795	28743	27989	26773
$Deviance_{M.0}$	29804	28752	27988	26771
$Deviance_{M.1}$	29784	28732	27967	26752
ΔDeviance (χ^2)	19	19	21	20
Freiheitsgrade	2	2	4	4
Signifikanz	<.001	<.001	<.001	.001

Anmerkungen
Informationskriterium *AICc* und Kenngrößen des LR-Tests zwischen leerem Modell (M.0) und Modell mit Lagerzugehörigkeit als L2-Prädiktor (M.1). Modell 0: Intercept-Only-Modell; Modelle 1-3: Wachstumskurvenmodelle mit linearem Slope, Intercept und linearem Slope, quadratischem Slope.
Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Einer vergleichbaren Logik folgt der „deviance test“ oder „likelihood ratio test“ (im Folgenden: LR-Test) (Snijders & Bosker, 2011, S. 97). Die Deviance ist ein Maß für die fehlende Passung des Modells zu den Daten. Ein Modell mit einer im Vergleich geringeren Deviance passt besser zu den Daten. Der Vorteil des Deviance-Tests ist, dass die Differenz der Deviance gegen eine χ^2-Verteilung mit der Differenz der Parameterzahl der Modelle als Freiheitsgrade

61 Burnham und Anderson (2004, S. 270) empfehlen, immer *AICc* zu verwenden, da es in kleinen Stichproben verlässlicher ist und sich mit steigender Stichprobengröße asymptotisch dem *AIC* nähert. Die Korrektur für den Stichprobenumfang ist allerdings in Mehrebenenmodellen nicht trivial, da wir entscheiden müssen, die Einheiten welcher Ebene zur Berechnung des Korrekturfaktors herangezogen werden sollen (Hox, 2010, S. 50). In klassischen Mehrebenenmodellen, in denen die individuellen L1-Einheiten auf höheren Ebenen zu Gruppen zusammengefasst werden, ist die L1-Fallzahl meist die sinnvollste Wahl. In Wachstumskurvenmodellen sind jedoch die eigentlichen Stichprobeneinheiten auf L2 angesiedelt. Daher ist es in den hier dargestellten Modellen naheliegend, den Korrekturfaktor anhand der Zahl der L2-Einheiten, also der Fallzahl der Probanden, zu ermitteln. Die hier präsentierten *AICc*-Werte werden mit dem *R* Paket *AICcmodavg* (Mazerolle, 2013) für die angemessene L2-Fallzahl berechnet.

getestet werden kann. Über diesen Test können wir bestimmen, ob eine Modellerweiterung – hier um den Faktor Lagerzugehörigkeit und seine Interaktionen mit den Zeit-Variablen – dazu beiträgt, das Modell signifikant zu verbessern. Der Deviance-Test ist damit kein Signifikanz-Test eines einzelnen Koeffizienten, sondern prüft die gesamte Modellverbesserung.[62] Damit kann Hypothese 1 zum Einfluss der Lagerzugehörigkeit auf die unmittelbare Bewertung getestet werden. Tabelle 6.2 ist zu entnehmen, dass die Hypothese von den Daten gestützt wird. In allen vier L1-Spezifikationen führt die Berücksichtigung der Lagerzugehörigkeit zu einer signifikanten Abnahme der Deviance.

Nachdem nun geklärt ist, dass die Aufnahme des L2-Prädiktors Lagerzugehörigkeit das Modell verbessert, wollen wir der Frage nachgehen, wie gut der Faktor die interindividuellen Unterschiede in der unmittelbaren Bewertung von Mappus während der Antwort erklären kann. Dazu sind im ersten Teil von Tabelle 6.3 die Varianzkomponenten der Modelle dargestellt. Im zweiten Teil der Tabelle wird die anteilige Reduktion der Varianzkomponenten in den Modellen M.1 in Vergleich zu den leeren Modellen M.0 berichtet.

Innerhalb des leeren Intercept-Only-Modells M0.1 gibt die Intraklassen-Korrelation ρ über das Verhältnis der Varianzkomponenten auf L2 (Varianz der individuellen Intercepts) und L1 (Residualvarianz) Auskunft. Sie ist ein Maß für den Anteil der Varianz, der durch die Zuordnung der RTR-Messungen zu den Probanden, bei denen sie erhoben wurden, erklärt werden kann (Hox, 2010, S. 15). Hier ergibt sich

$$\rho = \frac{\sigma^2_{\text{Intercept}_{\text{Mo.o}}}}{\sigma^2_{\text{Intercept}_{\text{Mo.o}}} + \sigma^2_{\text{L1-Residuum}_{\text{Mo.o}}}} = \frac{142.03}{142.03 + 76.13} = .65 \tag{6.10}$$

Damit können 65 Prozent der Varianz in den individuellen RTR-Messungen durch die Schätzung eines Intercepts für jeden Probanden erklärt werden. Dieser Anteil quantifiziert die durch L2-Prädiktoren maximal *erklärbare* Varianz. Leider steht für Modelle mit Random Slopes, wie wir sie in den L1-Wachstumskurvenmodellen spezifizieren, kein vergleichbares Maß zur Verfügung, da die Varianzen der Random Slopes von der Skalierung der unabhängigen Variablen abhängen. Wir können daher lediglich die Zuordnung von 65 Prozent der Streuung zur Varianzkomponente der Rezipienten im Intercept-

62 Es sei hier bereits darauf hingewiesen, dass der LR-Test für Fixed Effects eine erhöhte α-Fehlerrate aufweisen kann (z.B. Manor & Zucker, 2004; Pinheiro & Bates, 2000). Für die in diesem Teilkapitel präsentierten Tests ist dies unproblematisch, da die Effekte der Lagerzugehörigkeit deutlich jenseits des konventionellen Signifikanzniveaus von $p < .05$ liegen. In den Kapiteln 6.2 und 6.3 kommen wir ausführlicher auf diese Problematik zurück.

Tabelle 6.3: Varianzerklärung der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Lagerzugehörigkeit

	Modell 0	Modell 1	Modell 2	Modell 3
Varianzkomponenten M.0				
Intercept	142.03		49.47	
Slope		0.56	0.40	3.02
Slope²				0.00
L1-Residuum	76.13	58.09	44.74	32.05
Varianzkomponenten M.1				
Intercept	124.75		47.41	
Slope		0.49	0.36	2.78
Slope²				0.00
L1-Residuum	76.13	58.09	44.74	32.05
Varianzreduktion M.1 gegenüber M.0 (Anteil)				
Intercept	.12		.04	
Slope		.12	.08	.08
Slope²				.04
L1-Residuum	.00	.00	.00	.00

Anmerkungen
Varianzkomponenten der Modelle M.0 und M.1 sowie anteilige Reduzierung der Varianzkomponenten im Vergleich zwischen leerem Modell (M.0) und Modell mit Lagerzugehörigkeit als L2-Prädiktor (M.1). Modell 0: Intercept-Only-Modell; Modelle 1-3: Wachstumskurvenmodelle mit linearem Slope, Intercept und linearem Slope, quadratischem Slope.
Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Only-Modell als eine Untergrenze für die insgesamt durch Rezipientenmerkmale erklärbare Varianz annehmen. Der Vergleich der Modellpassungen hat ergeben, dass der Vorhersagefehler und damit auch die Residualvarianzen der Wachstumskurvenmodelle deutlich unter dem Intercept-Only-Modell liegen (vgl. Abbildung 6.7). Demnach muss der in den Wachstumskurvenmodellen durch Charakteristika der Rezipienten erklärbare Anteil der Varianz über dem des Intercept-Only-Modells liegen.

Der Vergleich der Varianzkomponenten kann dazu herangezogen werden, R^2-ähnliche Maße der Varianzaufklärung durch die in M.1 berücksichtigen L2-Prädiktoren zu berechnen. Für den einfachsten Fall des Intercept-Only-Modells ergibt sich als $R^2_{2_{\text{Intercept}}}$ auf Ebene der Rezipienten (L2) (Hox, 2010, S. 71-72):

$$R^2_{2_{\text{Intercept}}} = 1 - \frac{\sigma^2_{\text{Intercept}_{\text{Mo.1}}}}{\sigma^2_{\text{Intercept}_{\text{Mo.0}}}} = 1 - \frac{124.75}{142.03} = .12 \tag{6.11}$$

Um diese Kenngrößen richtig einzuordnen, müssen wir uns in Erinnerung rufen, dass die L2-Prädiktoren nur die Varianz in den Parametern der Random Effects erklären können. Die Lagerzugehörigkeit erklärt 12 Prozent der personenbezogenen L2-Varianzkomponente, die insgesamt 65 Prozent der gesamten Varianz im Modell ausmacht. Anders ausgedrückt handelt es sich bei dieser Varianzerklärung nicht um zusätzlich erklärte Varianz im Gesamtmodell. Die Varianz wurde bereits dadurch erklärt, dass wir die Zuordnung der RTR-Messungen zu den Rezipienten in der Mehrebenenmodellierung berücksichtigt haben. Nun können wir aussagen, dass 12 Prozent der Varianz zwischen den Personen auf ihre Lagerzugehörigkeit zurückzuführen ist.

Für das einfachste Wachstumskurvenmodell M1 mit nur einem variablen Parameter, einem linearen Slope, kann die Varianzaufklärung im Slope-Parameter durch das Verhältnis der Slope-Varianz in den Modellen M1.0 und M1.1 äquivalent bestimmt und interpretiert werden. Die Berücksichtigung des Faktors Lagerzugehörigkeit verringert die personenbezogene Varianzkomponente ebenfalls um 12 Prozent.[63] Allerdings können wir für dieses Modell nicht genau bestimmen, wie groß der Anteil der rezipientenbezogenen Varianzkomponente ist. Aus dem geringeren Vorhersagefehler gegenüber dem Intercept-Only-Modell können wir aber schließen, dass das einfache Wachstumskurvenmodell einen größeren Varianzanteil der Erklärung durch L2-Prädiktoren zugänglich macht, die Berücksichtigung der Lagerzugehörigkeit von dieser größeren Komponente einen ebenso großen Anteil erklären kann und der Faktor damit in Modell M1 auch insgesamt mehr zur Erklärung der individuellen RTR-Messungen beiträgt.

Die Interpretation der Reduzierung der personenbezogenen Varianzkomponenten in den Modellen M2 und M3 ist etwas komplexer, da hier jeweils zwei freie Parameter geschätzt werden. Es zeigt sich, dass die Lagerzugehörigkeit die Varianz in den individuellen Intercepts um 4 Prozent reduziert. Die Varianz der Slope-Parameter verringert sich um 8 Prozent. Im quadratischen Wachstumskurvenmodell nimmt die Varianz im Parameter des linearen Terms um 8

63 Es muss an dieser Stelle darauf hingewiesen werden, dass die auf zwei Dezimalstellen exakte Übereinstimmung der R^2_2 von Mo und M1 zufällig zustande kommt, ebenso, dass die Summe der R^2_2 von M2 und M3 12 Prozent ergibt. Durch die konzeptionelle Ähnlichkeit sollte die Größenordnung der Werte in etwa übereinstimmen, sie sind jedoch nicht notwendigerweise exakt gleich.

Prozent ab, die im Parameter des quadratischen Terms um 4 Prozent. Da wir die relative Größe der beiden Varianzkomponenten zueinander nicht bestimmen können, müssen die Reduzierungen jeweils für sich betrachtet interpretiert werden. Auch ein direkter Vergleich mit den R_2^2-Maßen der Modelle M0 und M1 ist nicht möglich. Es zeigt sich, dass der neue L2-Prädiktor einen geringen Teil der Varianz in den variablen Parametern der Wachstumskurven erklären kann. Inhaltlich bedeutet dies, dass die Lagerzugehörigkeit in geringem Umfang dazu beiträgt, die Unterschiede in den individuellen Bewertungen von Mappus während dieser Antwort zu erklären.

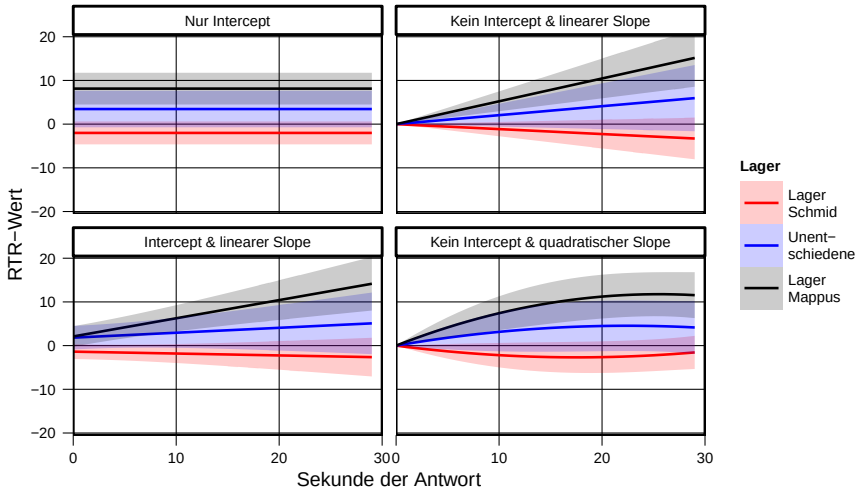
Noch nicht untersucht haben wir bisher, in welche Richtung sich die Lagerzugehörigkeit auf die unmittelbare Bewertung von Mappus auswirkt. Dies können wir anhand der in Tabelle 6.4 dargestellten Fixed Effects der Modelle M.1 oder einfacher anhand der durch die Modelle vorhergesagten RTR-Bewertungen für die drei Rezipientengruppen in Abbildung 6.8 untersuchen.

Tabelle 6.4: Effekt der Lagerzugehörigkeit auf die Bewertung von Mappus während der Antwort

	Modell 0.1 β (s.e.)	Modell 1.1 β (s.e.)	Modell 2.1 β (s.e.)	Modell 3.1 β (s.e.)
Intercept	3.45 (2.13)		1.82 (1.38)	
Lager Mappus	4.67 (2.83)		0.28 (1.83)	
Lager Schmid	-5.46 (2.52)*		-3.20 (1.63)*	
Zeit		0.21 (0.13)	0.11 (0.12)	0.40 (0.32)
Zeit x Lg. Mappus		0.32 (0.18)*	0.30 (0.15)*	0.52 (0.42)
Zeit x Lg. Schmid		-0.32 (0.16)*	-0.16 (0.14)	-0.71 (0.38)*
Zeit²				-0.01 (0.01)
Zeit² x Lg. Mappus				-0.01 (0.01)
Zeit² x Lg. Schmid				0.02 (0.01)

Anmerkungen
REML-Koeffizienten β mit ihren Standardfehlern (s.e) und Signifikanzniveau (Einseitige T-Tests mit Freiheitsgraden nach Satterthwaite-Approximation): $^*p < .05$. Modell 0: Intercept-Only-Modell; Modelle 1-3: Wachstumskurvenmodelle mit linearem Slope, Intercept und linearem Slope, quadratischem Slope.
Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Die Fixed-Effect-Koeffizienten des Intercept-Only-Modells M0.1 können direkt als die mittlere Bewertung von Mappus während der Antwort durch die drei Probandengruppen interpretiert werden. Da die Angehörigkeit zu einem der beiden Kandidatenlager mit Dummy-Variablen im Modell repräsentiert ist, entspricht der Intercept-Koeffizient der Bewertung von Mappus durch die Un-



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch die Modelle in Tabelle 6.4. Die Flächen zeigen 95%-Konfidenzintervalle.

Abbildung 6.8: Effekt der Lagerzugehörigkeit auf die Bewertung von Mappus während der Antwort

entschiedenen. Die Koeffizienten der beiden Lager geben an, wie die mittleren Bewertungen durch ihre Anhänger von der Bewertung durch die Unentschiedenen abweichen. Die Bewertung durch die eigenen Anhänger ist am positivsten, die durch die Anhänger Schmids am negativsten und die durch die Unentschiedenen liegt dazwischen. Das Ergebnis der Verrechnung der Koeffizienten können wir einfacher aus Abbildung 6.8 ablesen. Die Konfidenzintervalle der geschätzten RTR-Kurven zeigen, dass die Bewertungen durch Anhänger von Regierung und Opposition sich signifikant voneinander unterscheiden. Die Bewertungen durch die eigenen Anhänger liegen zudem signifikant über dem neutralen Mittelpunkt der Skala, während die Oppositionsanhänger ihn nur neutral, nicht aber signifikant negativ bewerten. Die mittlere Bewertung der Unentschiedenen ist zwar deskriptiv im positiven Bereich, unterscheidet sich aber ebenfalls nicht signifikant vom Nullpunkt.

Der Vergleich der Fixed Effects des einfachsten Wachstumskurvenmodells M1.1 ist ebenfalls noch anhand der Koeffizienten in Tabelle 6.4 zu erfassen.

Ausgehend von der Modellannahme, dass alle Bewertungen am Mittelpunkt der RTR-Skala beginnen, steigt die mittlere Bewertung durch die Unentschiedenen in jeder Sekunde der Antwort um $\beta_{\text{Zeit}} = 0.21$ Skalenpunkte. Für die Bewertung durch die Anhänger von Mappus ergibt sich eine Steigung von $\beta_{\text{Zeit}} + \beta_{\text{Zeit x Lg. Mappus}} = 0.53$ Skalenpunkten in der Sekunde. Die durchschnittliche Bewertung durch die Anhänger von Schmid fällt in jeder Sekunde minimal um $\beta_{\text{Zeit}} + \beta_{\text{Zeit x Lg. Schmid}} = -0.11$ Skalenpunkte. Was diese Koeffizienten für die Bewertung von Mappus im Verlauf der Antwort bedeuten, ist in Abbildung 6.8 deutlich einfacher zu erkennen. Von seinen Anhängern erfährt Mappus im Verlauf der Antwort recht deutliche Zustimmung, während die Veränderung der Bewertung der Unentschiedenen im Mittel zwar positiver wird, sich jedoch nie signifikant von der neutralen RTR-Position unterscheidet. Die geringe negative Steigung der Anhänger Schmids ist kaum nennenswert, sie deckt sich über die gesamte Antwort mit der neutralen Skalenmitte.

Um die Bewertung von Mappus nach Modell M2.1 zu interpretieren, müssen gleichzeitig die Koeffizienten des Intercepts und des linearen Slopes betrachtet werden. Dabei ist zu beachten, dass die Intercept-bezogenen Koeffizienten über die *geschätzten* mittleren Ausgangspunkte der RTR-Verläufe der Gruppen informieren, die von den beobachteten Ausgangspunkten abweichen können. Ein höherer Intercept-bezogener Koeffizient weist darauf hin, dass die Bewertung zu einem frühen Zeitpunkt der Antwort einen höheren Wert erreicht, sei es durch einen tatsächlich höheren Wert zum Beginn der Antwort oder durch eine (starke) positive Veränderung nach kurzer Zeit. Die Slope-bezogenen Koeffizienten geben über die Veränderung von diesem geschätzten Ausgangswert aus Auskunft (vgl. auch die Ausführungen zu Abbildung 6.6). Die geschätzten mittleren Bewertungen durch die Unentschiedenen ($\beta_{\text{Intercept}} = 1.82$) und die Anhänger von Mappus ($\beta_{\text{Intercept}} + \beta_{\text{Lager: Mappus}} = 2.10$) starten auf etwa gleichem Niveau. Für die Anhänger von Schmid liegt der Ausgangswert mit $\beta_{\text{Intercept}} + \beta_{\text{Lager: Schmid}} = -1.38$ niedriger. Während sich die Bewertungen der Unentschiedenen ($\beta_{\text{Zeit}} = 0.11$) und der Oppositionsanhänger ($\beta_{\text{Zeit}} + \beta_{\text{Zeit x Lg. Schmid}} = -0.05$) im Zeitverlauf kaum verändern, steigt die mittlere Bewertung durch die Regierungsanhänger ($\beta_{\text{Zeit}} + \beta_{\text{Zeit x Lg. Mappus}} = 0.41$) sichtbar an. Inhaltlich lässt sich dies so interpretierten: Die durchschnittliche Bewertung durch eigene Anhänger und Unentschiedene steigt relativ schnell um einen geringen Wert an, die Bewertung durch die Anhänger des anderen Kandidaten sinkt relativ schnell um einen gewissen Wert. Nur im eigenen Lager gewinnt Mappus im Verlauf seiner Antwort noch in bedeutenderem Umfang an Zustimmung. Bei der Interpretation des Modells mit zwei L1-Parametern und den entsprechenden Fixed-Effect-Koeffizienten wird der Vorteil der grafischen

Inspektion des Modell-Outputs deutlich. Das beschriebene Muster kann aus Abbildung 6.8 direkt abgelesen werden.

Fast vollständig auf die grafische Repräsentation des Modells verlassen müssen wir uns bei der Interpretation von Modell M3.1. Die Veränderung in allen Gruppen wird beschrieben durch einen linearen Wachstumskoeffizienten und einen quadratischen Dämpfungskoeffizienten, der dem linearen Wachstum entgegenwirkt. Die einzelnen Koeffizienten besitzen in dieser Spezifikation keine direkt interpretierbare inhaltliche Bedeutung. In Abbildung 6.8 wird aber die Logik ihres Zusammenwirkens deutlich. Der Dämpfungskoeffizient führt dazu, dass die Bewertung nicht bis zum Ende der Antwort linear weiter steigt bzw. fällt, sondern an einem gewissen Punkt ein Maximum bzw. Minimum erreicht. Inhaltlich ergibt sich hier ein mit den anderen L1-Spezifikationen konsistentes Bild. Die Bewertung durch die eigenen Anhänger entwickelt sich im Verlauf der Antwort deutlich positiv. Auch die Zustimmung der Unentschiedenen wächst, jedoch überschneidet sich das Konfidenzintervall durchgängig mit dem neutralen Skalenmittelpunkt. Die Urteile der Anhänger Schmids verlaufen deskriptiv leicht negativ, inferenzstatistisch bleibt dies jedoch bedeutungslos.

Insgesamt können wir bis zu diesem Punkt der Analyse festhalten, dass die Lagerzugehörigkeit entsprechend Hypothese 1 zwar signifikant, allerdings nur in relativ geringem Umfang zur Erklärung der Bewertung von Mappus während der Antwort beiträgt. Die Richtung der beobachteten Unterschiede nach Lagerzugehörigkeit entspricht den Erwartungen. Dies zeigt sich konsistent über alle L1-Spezifikationen.

Effekte der Voreinstellungen Die flexiblen Analysemöglichkeiten des Wachstumskurvenmodells zeigen wir, indem wir die Modelle um die Voreinstellungen der Rezipienten zu den Kandidaten und ihren Parteien erweitern. Die Voreinstellungen wurden jeweils anhand der üblichen elfstufigen Skalometer-Fragen erfasst (vgl. Kapitel 4). Eine volle Berücksichtigung von solchen quasimetrisch skalierten Variablen ist in Ansätzen, die auf Aggregationen über Personen beruhen, kaum möglich, da sehr viele Aggregate über entsprechend wenige Personen gebildet und die Befunde jeweils miteinander verglichen werden müssten. Tabelle 6.5 stellt die wichtigsten Kennwerte der zusätzlichen L2-Prädiktoren vor.

Die univariaten Maße liegen im Bereich dessen, was unsere zugunsten der Anhänger der Oppositionsparteien verzerrte Stichprobe erwarten lässt. Auffällig ist jedoch der fehlende Zusammenhang zwischen den Einstellungen zur CDU und Mappus auf der einen und der SPD und Schmid auf der anderen Seite. Die Kandidaten und Parteien beider Lager werden offenbar unabhängig

Tabelle 6.5: Überblick über die Voreinstellungen der Rezipienten

			Korrelation (r)			
	M	SD	(1)	(2)	(3)	(4)
(1) Skalometer Mappus	−1.36	2.85	1.00			
(2) Skalometer Schmid	0.60	1.68	−.00	1.00		
(3) Skalometer CDU	0.53	2.95	.64	−.04	1.00	
(4) Skalometer SPD	1.12	2.16	.01	.67	.14	1.00

Anmerkungen
Alle Variablen auf einer Skala von −5 (Halte überhaupt nichts von diesem Politiker / dieser Partei) bis +5 (Halte sehr viel von diesem Politiker / dieser Partei); n = 136 Rezipienten.

voneinander bewertet. Eine zu erwartende negative Korrelation (je positiver Schmid (die SPD) beurteilt wird, desto negativer wird Mappus (die CDU) bewertet) findet sich nicht.

Im Folgenden schätzen wir für jede L1-Spezifikation Modelle, die die Voreinstellungen zu den Kandidaten (M.2), den Parteien (M.3) sowie alle vier Variablen (M.4) als L2-Prädiktoren enthalten. In einem finalen Modell (M.5) entfernen wir alle L2-Prädiktoren, die nach den Ergebnissen von LR-Tests nicht signifikant zur Verbesserung der Modelle beitragen, um ein möglichst übersichtliches Ergebnis zu erhalten. Als Hypothesen formulieren wir:

> **H2**: Je besser Mappus (H2a) und die CDU (H2b) vor dem Duell bewertet werden, desto positiver wird Mappus während der Antwort bewertet.

> **H3**: Je besser Schmid (H3a) und die SPD (H3b) vor dem Duell bewertet werden, desto negativer wird Mappus während der Antwort bewertet.

Tabelle 6.6 vergleicht die Modelle mit den zusätzlichen L2-Prädiktoren für die Intercept-Only-L1-Spezifikation. Sowohl die Berücksichtigung der Kandidaten-Skalometer (M0.2) als auch die der Partei-Skalometer (M0.3) stellen signifikante Verbesserungen gegenüber dem leeren Modell (M0.0) dar. Die niedrigeren $AICc$-Werte und die Reduzierung der personenbezogenen Varianzkomponente um 15 bzw. 20 Prozent zeigen, dass die Prädiktoren besser zur Erklärung der unmittelbaren Bewertung von Mappus geeignet sind als der einfache Faktor Lagerzugehörigkeit. Bei einer multivariaten Betrachtung der Einflüsse aller Variablen (M0.4) stellen sich die mit den Skalometer-Fragen gemessenen

Voreinstellungen gegenüber Mappus und der SPD als die beiden wichtigsten Prädiktoren heraus (vgl. Tabelle 6.7). Der Vergleich des Modells mit allen L2-Prädiktoren mit einem Modell, das nur diese beiden Variablen enthält (Mo.5), zeigt, dass das sparsamere Modell nicht signifikant schlechter abschneidet. Zudem ist der Wert des *AICc* des sparsameren Modells niedriger, was dafür spricht, dass die zusätzliche Erklärungsleistung der weiteren Prädiktoren in Mo.4 den Anstieg der Modellkomplexität nicht rechtfertigt. Die beiden für das beste Modell Mo.5 ausgewählten L2-Prädiktoren können 21 Prozent der personenbezogenen Varianzkomponente erklären.

Tabelle 6.6: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Voreinstellungen (Mo)

	Mo.0	Mo.1	Mo.2	Mo.3	Mo.4	Mo.5
AICc	29810	29795	29790	29783	29786	29781
Deviance	29804	29784	29779	29772	29766	29771
ΔDeviance (χ^2)		19	24	31	37	−5
Freiheitsgrade		2	2	2	6	−4
Signifikanz		<.001	<.001	<.001	<.001	.335
$R^2_{2_{\text{Intercept}}}$		.12	.15	.20	.21	.21

Anmerkungen
Informationskriterium *AICc*, Kenngrößen der LR-Tests und Reduzierung der personenbezogenen Varianzkomponente ($R^2_{2_{\text{Intercept}}}$) im Vergleich zu Mo.0. Die Modelle Mo.1 bis Mo.4 werden mit LR-Tests mit Mo.0 verglichen. Das finale Modell Mo.5 wird mit dem vollen Modell Mo.4 verglichen, um zu zeigen, dass der Verzicht auf die entsprechenden Koeffizienten das Modell nicht signifikant verschlechtert. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

In Tabelle 6.7 sind die Koeffizienten der Fixed Effects verzeichnet, mit denen sich die gerichteten Hypothesen 2 und 3 prüfen lassen. Die Modelle, in denen nur die Voreinstellungen zu den Kandidaten (Mo.2) bzw. den Parteien (Mo.3) untersucht werden, stützen die Hypothesen. Zwischen der Einstellung zu Mappus bzw. der CDU und der unmittelbaren Bewertung von Mappus während der Antwort besteht ein positiver Zusammenhang. Zwischen der Einstellung zu Schmid bzw. der SPD und der unmittelbaren Bewertung besteht ein negativer Zusammenhang. Die gleichzeitige Analyse der Voreinstellungen gegenüber den Kandidaten und Parteien sowie der Lagerzugehörigkeit offenbart die Einstellung gegenüber Mappus und der SPD als entscheidende Einflüsse. Die Wirkungsrichtung ist hypothesenkonform: Je positiver Mappus und je negativer die SPD vor dem Duell gesehen wurden, desto besser wird Mappus während seines Angriffs auf die finanzpolitischen Pläne der Oppositionsparteien bewertet.

Tabelle 6.7: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (Mo)

	Mo.0	Mo.1	Mo.2	Mo.3	Mo.4	Mo.5
Intercept	1.87*	3.45	4.74***	3.36***	4.22*	5.80***
	(1.03)	(2.13)	(1.11)	(1.05)	(2.24)	(1.13)
Lg. Mappus		4.67			1.93	
		(2.83)			(2.82)	
Lg. Schmid		−5.46*			−0.39	
		(2.52)			(2.74)	
Sk. Mappus			1.45***		0.80*	1.47***
			(0.33)		(0.46)	(0.32)
Sk. Schmid			−1.51**		0.32	
			(0.57)		(0.76)	
Sk. CDU				1.38***	0.72	
				(0.32)	(0.46)	
Sk. SPD				−1.97***	−1.91**	−1.72***
				(0.43)	(0.62)	(0.43)

Anmerkungen
Dargestellt sind die REML-Koeffizienten β mit ihren Standardfehlern (s.e) und Signifikanzniveau (Einseitige T-Tests mit Freiheitsgraden nach Satterthwaite-Approximation): $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Lg.: Lager; Sk.: Skalometer. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

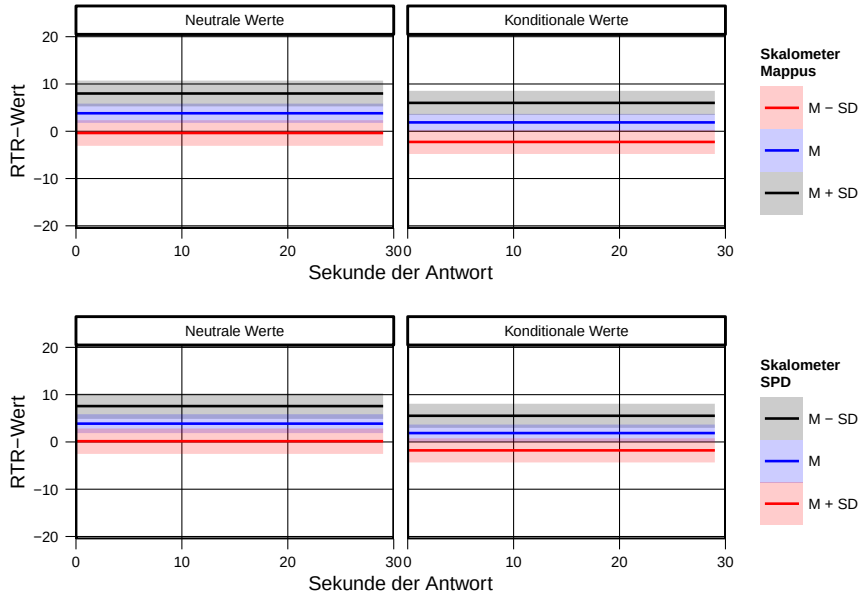
Die Ergebnisse lassen sich in dieser einfachsten L1-Spezifikation noch relativ gut anhand der in der Tabelle verzeichneten Koeffizienten als der mittlere Unterschied zwischen zwei Rezipienten interpretieren, die sich in einem L2-Prädiktor gleichen und im anderen L2-Prädiktor um einen Skalenpunkt unterscheiden. Die Bedeutung dieses Ergebnisses für die vorgesagten RTR-Werte können anhand deren grafischer Darstellung jedoch besser erfasst werden – dies gilt besonders für die noch folgenden komplexeren Modellspezifikationen. Abbildung 6.9 vermittelt einen Eindruck des Effekts der Voreinstellungen zu Mappus und zur SPD auf die unmittelbaren Urteile während der Antwort.

Die Ergebnisdarstellungen in Abbildung 6.9 stehen für zwei Möglichkeiten, über Fixed-Effects-Koeffizienten von Mehrebenenmodellen bzw. allgemein über die Bedeutung von Regressionskoeffizienten nachzudenken. Die einfache L1-Spezifikation bietet sich an, um die grundlegende Logik der Ergebnisdarstellungen, die auch im Folgenden vor der Präsentation der Koeffizienten in Tabellen bevorzugt werden, ausführlicher zu erläutern. Die drei Linien mit unterschiedlichen Farben stehen für die vorhergesagte RTR-Bewertung für Rezipienten, die auf dem jeweiligen L2-Prädiktor auf dem Mittelwert sowie je eine Standardabweichung über und unter dem Mittelwert liegen. Die Wahl dieser Ausprägungen stellt sicher, dass die vorhergesagten RTR-Werte für Werte der

L2-Prädiktoren stehen, die die Bandbreite der Einstellungen in der Stichprobe charakterisieren. Dabei wird auch berücksichtigt, dass sich die L2-Prädiktoren in der Stichprobe unterschiedlich verteilen. Mappus wird mit $M = -1.36$ $(SD = 2.85)$ im Durchschnitt relativ negativ bewertet. Die SPD wird dagegen mit $M = 1.12$ $(SD = 2.16)$ recht positiv wahrgenommen. Die Entscheidung für positive und negative Werte eine Standardabweichung unterhalb und oberhalb des Mittelwerts führt dazu, dass die Interpretation der Grafiken angelehnt an die Interpretation standardisierter Regressionskoeffizienten erfolgt (Gelman & Hill, 2006): Wenn sich zwei Rezipienten auf der Skala des Prädiktors um eine Standardabweichung unterscheiden, dann unterscheidet sich ihre unmittelbare Bewertung um die dargestellte Differenz.

Die farbliche Codierung der Linien nehmen wir entsprechend ihres Bezugs zu den Kandidaten Mappus und Schmid vor. Der Mittelwert wird blau dargestellt, die Abweichung in die Richtung, die auf der jeweiligen Variable dem Lager Mappus bzw. Schmid entspricht, in schwarz bzw. rot. Daher ist die Linie für eine negative Voreinstellung gegenüber Mappus, die eine Standardabweichung unter der mittleren Einstellung liegt, in rot dargestellt, eine überdurchschnittlich positive Voreinstellung von einer Standardabweichung über dem Mittelwert in schwarz. Genau umgekehrt ist die Farbzuordnung für die Einstellung der SPD angeordnet. Eine negative Bewertung der SPD wird nach der Logik der politischen Lagerzugehörigkeit mit einer schwarzen Linie, eine positive Bewertung mit einer roten Linie und eine durchschnittliche Bewertung mit einer blauen Linie dargestellt.

In den Facetten zu einem L2-Prädiktor sind zwei Möglichkeiten visualisiert, die Bedeutung der Koeffizienten für die unmittelbare Bewertung durch die Rezipienten zu interpretieren. Sie unterscheiden sich darin, welche Werte für die übrigen L2-Prädiktoren (in diesem Fall für den zweiten L2-Prädiktor) angenommen werden. Die Vorhersage der RTR-Bewertung auf Basis *neutraler Werte* in der linken Facette geht davon aus, dass alle übrigen L2-Prädiktoren einen inhaltlich neutralen Wert aufweisen. Am Beispiel des in Abbildung 6.9 oben dargestellten Effekts der Voreinstellung zu Mappus bedeutet dies, dass alle drei RTR-Verläufe unter der Annahme des neutralen Werts 0 auf dem Skalometer zur SPD vorhergesagt werden. Der Unterschied zwischen dem roten und dem schwarzen RTR-Verlauf kann damit als der Unterschied der Bewertungen durch zwei Rezipienten verstanden werden, deren Voreinstellungen gegenüber Mappus um zwei Standardabweichungen auseinander liegen, die aber beide gegenüber der SPD eine neutrale Einstellung haben. Der so dargestellte Unterschied kann als eine Annäherung an den Effekt verstanden werden, der ausschließlich auf die Voreinstellung zu Mappus zurückgeht und



Anmerkungen
Bewertung von Mappus auf einer Skala von -50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell 0.5 in Tabelle 6.7.
Vorhergesagte RTR-Werte beim $M \pm 1SD$) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf den Skalenmittelpunkt 0 gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt. Die Flächen zeigen 95%-Konfidenzintervalle.

Abbildung 6.9: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (Mo)

nicht von der ebenfalls wirksamen Einstellung gegenüber der SPD sowie der Verteilung der L2-Prädiktoren in der Stichprobe konfundiert wird.

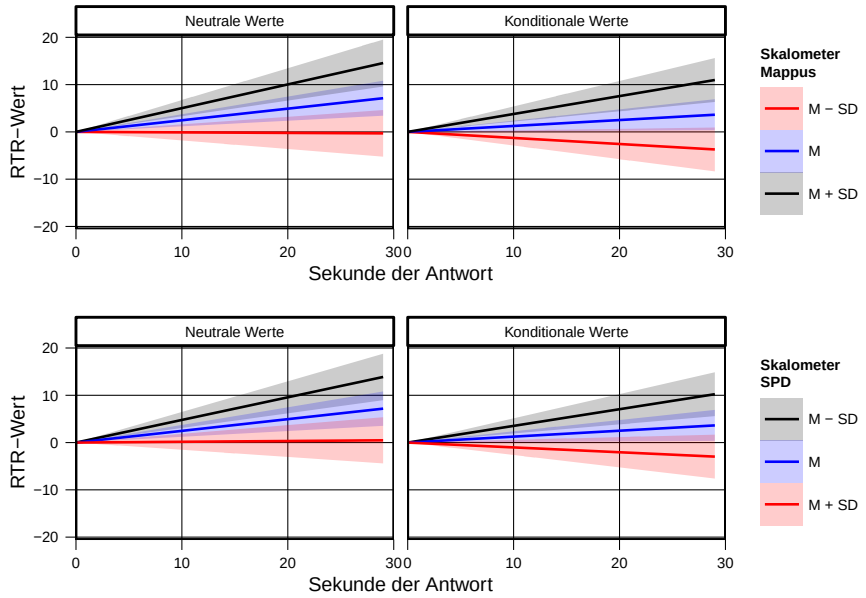
In der rechten Facette werden die vorhergesagten RTR-Verläufe auf Grundlage von Werten der anderen L2-Prädiktoren geschätzt, die für die Ausprägung des Prädiktoren von Interesse typisch sind. Im genannten Beispiel wird dazu ermittelt, wie Rezipienten, die eine unterdurchschnittliche ($M - 1SD$), durchschnittliche (M) und überdurchschnittliche ($M + 1SD$) Einstellung zu Mappus

haben, die SPD bewerten.[64] Die so vorhergesagten RTR-Verläufe vermitteln ein realistisches Bild davon, wie sich der Effekt eines Prädiktors in Abhängigkeit der Verteilung der übrigen Prädiktoren in der Stichprobe zeigt.

Im hier dargestellten Beispiel sind sich beide Betrachtungsweisen recht ähnlich. Dies liegt vor allem daran, dass zwischen der Bewertung von Mappus und der SPD mit $r = .01$ kein Zusammenhang besteht. Der wesentliche Unterschied besteht darin, dass die vorhergesagten RTR-Verläufe, die konditional der Verteilungen in der Stichprobe geschätzt werden, etwas negativer ausfallen als die Vorhersagen auf Basis neutraler Werte. Darin zeigt sich die tendenziell negative Voreinstellung der Rezipienten zu Mappus bzw. die tendenziell positive Einstellung zur SPD. Inhaltlich bestätigt sich das in Hypothesen 2 und 3 erwartete Muster. Die Visualisierung hilft uns, zu erkennen, wie sich die Effekte der Voreinstellungen in den Bewertungen durch für die Stichprobe typische Rezipienten zeigen. Wieder spielen sich alle Unterschiede im positiven Wertebereich der RTR-Skala ab. Eine positive Einstellung zu Mappus und eine negative Einstellung zur SPD führen zu einer positiven RTR-Bewertung der Antwort. Dagegen bewerten selbst Rezipienten, die Mappus vor dem Duell sehr negativ oder die SPD sehr positiv sahen, die Antwort eher neutral. Dies zeigt sich besonders deutlich, wenn wir die Bewertung des jeweils anderen Prädiktors auf den neutralen Wert 0 festsetzen und so nur die Vorhersage betrachten, die auf unterschiedlichen Werten des einen Prädiktors beruhen.

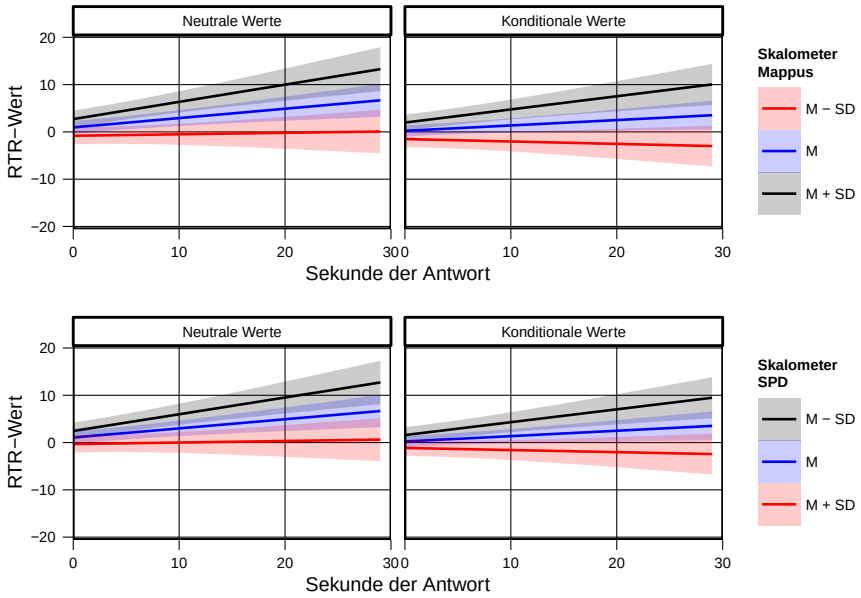
Auf den folgenden Seiten finden sich die vorhergesagten RTR-Verläufe auf Basis der Wachstumskurvenmodelle M1 bis M3. Das Vorgehen bei der Identifikation der finalen Modelle erfolgt analog zum gerade beschriebenen Vorgehen. Die Tabellen zur Modellwahl (Tabellen A.1, A.3, A.5) und der Fixed Effects (Tabellen A.2, A.4, A.6) finden sich im Anhang.

64 Analytisch werden diese Werte wie folgt ermittelt: In einer bivariaten Regression wird die Bewertung der SPD durch die Bewertung von Mappus erklärt. Anhand der Regressionsgerade lassen sich die Werte bestimmen, die für die Ausprägungen M und $M \pm 1SD$ des Skalometers Mappus im Skalometer SPD vorhergesagt werden.



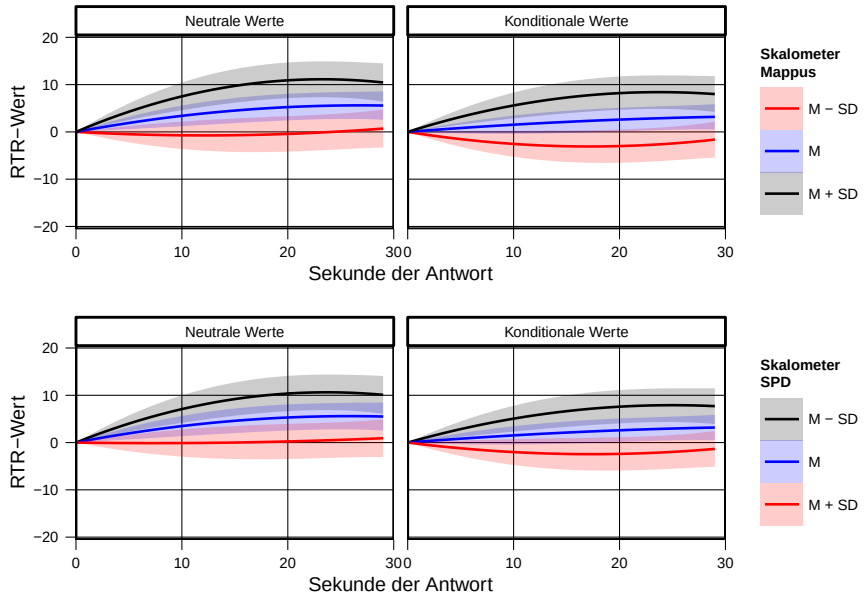
Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell 1.5 in Tabelle A.2.
Vorhergesagte RTR-Werte bei $M \pm 1SD$ des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf den Skalenmittelpunkt 0 gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt. Die Flächen zeigen 95%-Konfidenzintervalle.

Abbildung 6.10: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M1)



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell 2.5 in Tabelle A.4.
Vorhergesagte RTR-Werte bei $M \pm 1SD$ des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf den Skalenmittelpunkt 0 gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt. Die Flächen zeigen 95%-Konfidenzintervalle.

Abbildung 6.11: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M2)



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell 3.5 in Tabelle A.6.
Vorhergesagte RTR-Werte bei $M \pm 1SD$ des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf den Skalenmittelpunkt 0 gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt. Die Flächen zeigen 95%-Konfidenzintervalle.

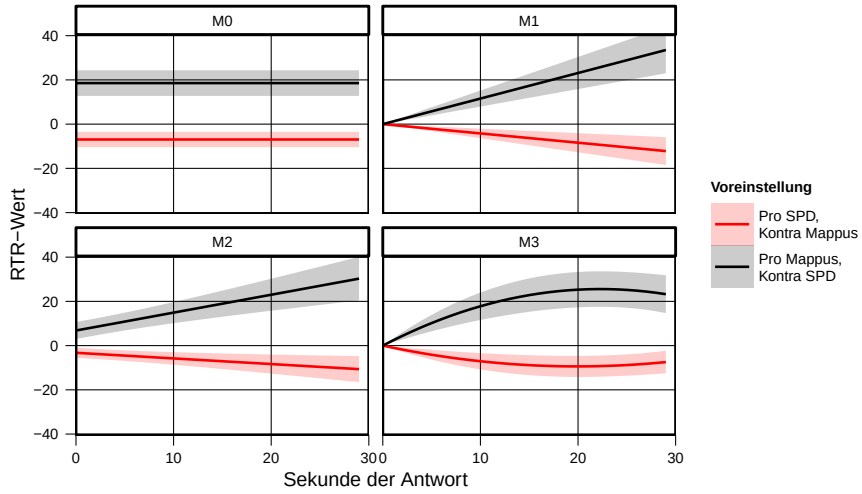
Abbildung 6.12: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M3)

In allen L1-Spezifikationen werden die Voreinstellungen zu Mappus und zur SPD als wichtige Prädiktoren der unmittelbaren Bewertung von Mappus während der Antwort identifiziert. Die personenbezogene Varianzaufklärung steigt jeweils gegenüber den einfachen Modellen der Lagerzugehörigkeit substantiell an, wenn auch alle Modelle weiter einen verhältnismäßig geringen Erklärungsbeitrag liefern. Dies gilt nicht nur für die finalen Modelle, sondern auch für die Teilmodelle, die jeweils die mit Skalometer gemessenen Voreinstellungen zu den Kandidaten bzw. Parteien berücksichtigen. Die Effektrichtung der Voreinstellungen entspricht den Annahmen in Hypothesen 2 und 3. Beide Prädiktoren in den finalen Modellen M.5 haben in etwa dieselbe Effektstärke: Mit der Veränderung des L2-Prädiktors um eine Standardabweichung geht jeweils eine etwa gleich große Veränderung der vorhergesagten RTR-Verläufe einher. Auffällig ist, dass sich die Veränderungen immer im positiven Spektrum der RTR-Skala zeigen. Eine positive Haltung zu Mappus und eine negative Haltung zur SPD führen zu Zustimmung zur Antwort von Mappus. Die umgekehrten Voreinstellungen schlagen sich jedoch nicht in negativen unmittelbaren Urteilen nieder. Sie führen für durchschnittliche Ausprägungen der L2-Prädiktoren lediglich zu einer neutralen Haltung.

Nur wenn wir sehr extreme Ausprägungen beider Voreinstellungen annehmen und ihren additiven Effekt berücksichtigen, erhalten wir eine signifikant negative Reaktion auf die Antwort. In Abbildung 6.13 sind in Rot die vorhergesagten RTR-Verläufe für Rezipienten dargestellt, die vor der Debatte Mappus sehr negativ und die SPD sehr positiv bewerteten. Diese Kombination, die in unserer Stichprobe so oder noch extremer mit immerhin acht Fällen vertreten ist, führt zu einer signifikanten Ablehnung der Antwort. In ihrem Umfang fällt diese Ablehnung jedoch noch immer vergleichsweise schwach aus. Wie zuerst Faas und Maier (2004a) feststellten, fällt den Rezipienten der Duelle offenbar eine Zustimmung leichter als eine Ablehnung.

6.1.3 Zwischenfazit zu den Wachstumskurvenmodellen

Inhaltlich können wir festhalten, dass sich der zu erwartende Effekt der Voreinstellungen auf die Bewertung eines Kandidaten während seiner Antwort im TV-Duell klar nachweisen lässt. Dies gilt sowohl für die auch in den aggregatbasierten Studien übliche Lagerzuordnung als auch für die quasi-metrisch gemessenen Voreinstellungen zu Kandidaten und Parteien. Bemerkenswert ist, dass in den finalen Modellen die Voreinstellung zum Sprecher der Antwort, Stefan Mappus, und zur Partei des Gegenkandidaten, der SPD, am besten zur Vorhersage der unmittelbaren Antworten geeignet sind. *Ex post* bieten sich hierfür zwei plausible Erklärungen an. In der hier untersuchten Antwort greift



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch die finalen Modelle M.5. *Pro SPD, Kontra Mappus*: Skalometer SPD = 4, Skalometer Mappus = −4; *Pro Mappus, Kontra SPD*: Skalometer SPD = −4, Skalometer Mappus = 4. Die Flächen zeigen 95%-Konfidenzintervalle.

Abbildung 6.13: Effekte extremer Voreinstellungen auf die Bewertung von Mappus während der Antwort

Mappus die Pläne der Oppositionsparteien in der Finanz- und Steuerpolitik an. Obwohl der Angriff sich zum Teil auch direkt gegen Schmid richtet, bezieht er sich doch hauptsächlich auf die politischen Pläne des rot-grünen Lagers. Daher ist es naheliegend, dass die Einstellung zum Angreifer und die Einstellung zur hauptsächlich angegriffenen SPD beeinflussen, wie der Angriff bewertet wird. Der im Vergleich zur Einstellung zu Schmid wichtigere Effekt der Einstellung zu seiner Partei könnte aber auch damit zusammenhängen, dass die Rezipienten vor dem Duell zum Kandidaten Schmid noch keine sehr prägnante Meinung hatten (Bachl, 2013b). Für diese Erklärung spricht auch, dass in den Modellen M.2, in denen nur die Einstellungen zu den Kandidaten als L2-Prädiktoren enthalten sind, die Effektstärke der Voreinstellung zu Schmid relativ gering ist. In der vorliegenden Analyse einer einzelnen Antwort muss diese Frage ungeklärt bleiben. Indem in den folgenden kreuzklassifizierten

(Wachstumskurven-) Modellen viele Aussagen der Kandidaten simultan und zudem die Aussagen von Schmid untersucht werden, können wir dieser Frage weiter nachgehen. Diese Modelle erlauben es dann, über die untersuchten Aussagen hinweg verallgemeinerbare Interpretationen zu finden.

Bezüglich des analytischen Nutzens der Wachstumskurvenmodellierung fallen die Ergebnisse gemischt aus. Das Verwenden eines Mehrebenenmodells hat sich generell bewährt. Die (zugegebenermaßen einfachen) Hypothesen zu den Effekten der Voreinstellungen auf die Bewertung der Antwort können unter Berücksichtigung der Varianz innerhalb und zwischen den individuellen Rezipienten korrekt getestet werden. Dabei stützen alle vorgestellten L1-Spezifikationen konsistent die Annahmen der Hypothesen. Die Mehrebenenmodelle erlauben es zudem, den Einfluss mehrerer und quasi-metrisch skalierter L2-Prädiktoren zu den Personencharakteristika zu modellieren, ohne die Messungen der Antwort zu einem einzelnen Mittelwert zusammenfassen zu müssen. Die Berücksichtigung der informationsreicheren L2-Prädiktoren erhöht die Aufklärung der personenbezogenen Varianzkomponente substantiell. Einschränkend muss angemerkt werden, dass die inhaltliche Spezifikation des finalen Modells mit den allgemeinen Prädiktoren Voreinstellung zu Mappus und Voreinstellung zur SPD nur wenig an den konkreten Inhalt dieser Antwort angepasst ist. Entsprechend gering fallen die L2-Varianzaufklärungen absolut betrachtet aus. Ginge es in der zentralen Forschungsfrage einer Arbeit um die Erklärung der Publikumsreaktionen auf genau diese Antwort, so sollten weitere, im Idealfall themenspezifische Prädiktoren berücksichtigt werden.

Offen bleibt an dieser Stelle, inwiefern sich die Spezifikation der Wachstumskurvenmodelle auf Ebene der individuellen Messungen (L1) „lohnt“. Dagegen spricht, dass die inhaltlichen Schlussfolgerungen für alle L1-Spezifikationen, also auch für das einfache Intercept-Only-Modell ohne Berücksichtigung der zeitlichen Dynamik, konsistent sind. Es macht in dieser Hinsicht also keinen Unterschied, ob die individuellen RTR-Messungen als dynamische Verlaufskurven oder als einfache Konstanten modelliert werden. Zwar liegen leichte Hinweise darauf vor, dass die Erfassung größerer Teile der Varianz in den individuellen Messungen durch Wachstumskurven mit zwei Parametern (M2 & M3) die Fähigkeit der Modelle verbessert, auch schwächere Effekte identifizieren zu können. So scheitert die Aufnahme der Voreinstellung zur CDU als weiterer L2-Prädiktor in diesen Modellen nur knapp, während sich in den Modellen mit nur einem L1-Parameter keinerlei Verbesserung durch die Aufnahme dieses Prädiktors zeigt. Letztendlich ergeben sich für die inhaltlichen Interpretationen der Einflüsse auf die unmittelbare Bewertung dieser Antwort aber keine Vorteile aus komplexeren L1-Spezifikationen. Bestünde der Zweck einer Studie nur aus der Erklärung der unmittelbaren Bewertung *dieser* Antwort, so wäre

es pragmatisch angemessen, in der Ergebnisdarstellung das Intercept-Only-Modell zu berichten, da sich seine Koeffizienten direkt interpretieren und die Befunde einfacher kommunizieren lassen. Allerdings sollte diese Entscheidung immer erst nach dem Vergleich mit komplexeren Modelle getroffen werden, um abzusichern, dass die Beschränkung auf das einfache Intercept-Only-Modell keine wichtigen Befunde verschleiert.

Es sollte aber nicht vernachlässigt werden, dass die Dynamik der individuellen RTR-Verläufe durch die Wachstumskurven wesentlich realistischer abgebildet wird. Je komplexer die L2-Parametrisierung, desto geringer sind die Abweichungen zwischen den beobachteten und den vorhergesagten Verläufen. Bei der detaillierten Beschreibung der Effekte der Voreinstellungen liefern die Wachstumskurven mit zwei Parametern so auch zusätzliche Informationen. So können wir anhand der visuellen Inspektion der Vorhersagen der Modelle M2 und M3 feststellen, dass sich die positiven Effekte von Voreinstellungen, die sich zu Gunsten Mappus bzw. zu Ungunsten von Schmid einordnen lassen, sowohl in einem früheren als auch in einem rapideren Anstieg der unmittelbaren Bewertungen zeigen. Zum Ende der Antwort hin kumuliert sich so ein beträchtlicher Unterschied in Abhängigkeit von den Voreinstellungen. Allgemeiner gesprochen geben Wachstumskurvenmodelle auch Informationen über den Verlauf und die Bewertung am Ende einer Antwort, während Intercept-Only-Modelle nur die durchschnittliche Bewertung während der gesamten Antwort quantifizieren.

Diese detaillierte Beschreibung kann bereits bei der Interpretation der Befunde zu einer einzelnen Antwort von Interesse sein. Vor allem bei der Analyse von Reaktionen auf mehrere Antworten bieten sich die komplexeren Modelle an, da sie vielfältigere Möglichkeiten des Vergleichs bieten. Besonders vielversprechend ist in dieser Hinsicht das Modell M2 mit einem latenten Intercept und einem latenten linearen Slope. Das lineare Modell bietet im Vergleich zur Spezifikation eines nicht-linearen Wachstums zunächst den Vorteil, dass sich die Koeffizienten relativ einfach direkt interpretieren lassen. Durch seine Sparsamkeit besteht eine recht geringe Gefahr der Überanpassung an die Daten, und die Berücksichtigung der über die Zeit variablen interindividuellen Varianz erlaubt eine angemessene Schätzung der statistischen Unsicherheit um die vorhergesagten Werte. Neben diesen datenanalytischen Vorzügen ist jedoch vor allem die theoretisch-konzeptionelle Eignung dieser Spezifikation herauszuheben. Effekte auf den latenten Intercept und auf den latenten linearen Slope können in der Interpretation der Befunde voneinander abgegrenzt werden. So wird ein Rückbezug zu Theorien der Informationsverarbeitung ermöglicht. Wie solche Analysen mit kreuzklassifizierten Wachstumskurvenmo-

dellen, die auf L1 entsprechend des hier vorgestellten Modells M2 spezifiziert sind, durchgeführt werden, zeigen wir im übernächsten Teilkapitel 6.3.

6.2 Das kreuzklassifizierte Modell

In diesem Teilkapitel beschreiben wir den Einsatz kreuzklassifizierter Modelle zur Erklärung der unmittelbaren Kandidatenbewertungen während eines TV-Duells. Der besondere Nutzen von kreuzklassifizierten Modellen ist es, die RTR-Messungen gleichzeitig als Reaktionen individueller Rezipienten und als Reaktionen infolge unterschiedlicher Debatteninhalte aufzufassen. Die Modelle ermöglichen es, Eigenschaften der Rezipienten – wie z.B. ihre Voreinstellungen gegenüber den Kandidaten – und Eigenschaften der Kandidatenaussagen – wie z.B. deren Themen – zur Erklärung der unmittelbaren Kandidatenbewertungen heranzuziehen. Noch wichtiger ist jedoch ihre Fähigkeit, Interaktionen zwischen Rezipienten- und Inhaltsmerkmalen explizit zu modellieren und so z.B. Hypothesen über das Zusammenwirken von Voreinstellungen der Rezipienten und Themen der Kandidatenaussagen zu testen. Dabei werden in den Modellen nicht nur Eigenschaften von Rezipienten und Inhalten zur Erklärung herangezogen, auch die statistische Unsicherheit auf Rezipienten- und Inhaltsebene wird angemessen in der Schätzung berücksichtigt. Dies trägt dem Umstand Rechnung, dass die Probanden eine realisierte Stichprobe aus einer Grundgesamtheit von Personen und die Inhalte der Debatte Realisationen aus einer Grundgesamtheit möglicher (politischer) Inhalte sind. Für viele der relevanten Forschungsfragen zu TV-Duell-Studien, die Inferenzschlüsse auf andere Personen *und* andere Inhalte erfordern (vgl. Kapitel 2), sind kreuzklassifizierte Modelle damit angemessene Analyseverfahren.

Wie im vorangegangen Kapitel führen wir zuerst in die grundsätzliche Logik der Modellklasse ein. Dann folgt eine praktische Demonstration: Wir untersuchen, wie die Kandidaten in Abhängigkeit von den Voreinstellungen der Rezipienten und von den Themen der Turns bewertet werden.

6.2.1 Grundlagen der Modellklasse

Kreuzklassifizierte Datenstrukturen haben mit einfachen hierarchischen Strukturen gemein, dass die einzelnen Messungen (L1) verschiedenen Gruppen auf einer höheren Ebene (L2) zugeordnet werden können. Der Unterschied besteht darin, dass die Messungen kreuzklassifizierter Strukturen in zueinander nicht hierarchisch angeordneten Ebenen gruppiert sind. Solche Datenstrukturen werden in der methodologischen Forschung zu Mehrebenenmodellen bereits seit

längerer Zeit diskutiert (Rasbash & Goldstein, 1994; Raudenbush, 1993). Da ihre numerische Evaluation allerdings komplex ist, verbreitet sich ihre Umsetzung in der angewandten Forschung mit Hilfe geläufiger Statistik-Software erst seit einigen Jahren (Hox, 2010, S. 186). Das *R* Paket *lme4*, das wir in dieser Arbeit einsetzen, ist besonders gut geeignet, komplexe kreuzklassifizierte Modelle zu schätzen (Bates, 2013b; Luo, 2013).

In der Grundlagenliteratur zu Mehrebenenmodellen wird die kreuzklassifizierte Datenstruktur häufig anhand von Beispielen aus der Bildungsforschung beschrieben (z.B. Beretvas, 2010, S. 314-316; Hox, 2010, S. 173-177; Snijders & Bosker, 2011, S. 206-207): So sind in einem einfachen hierarchischen Datensatz Schüler (L1) in Schulen (L2) gruppiert. In einer Querschnittserhebung gehört jeder Schüler zu genau einer Schule. Auf L2-Ebene kann dann beispielsweise untersucht werden, welche Eigenschaften der Schulen die Lernleistung der Schüler beeinflussen. Zusätzlich könnte aber z.B. auch von Interesse sein, ob Eigenschaften der Wohnorte der Schüler ihre Leistung ebenfalls beeinflussen. Damit wird eine weitere L2-Ebene eingeführt, da jeder Schüler (L1) genau einem Wohngebiet (L2) zugeordnet ist. Die Schüler sind also sowohl in Schulen als auch in Wohngebieten gruppiert. Schulen und Wohngebiete sind jedoch nicht hierarchisch angeordnet, da Schüler aus verschiedenen Wohngebieten dieselbe Schule besuchen und umgekehrt Schüler aus verschiedenen Schulen in dem selben Gebiet wohnen können. Die Austauschbarkeit der Formulierungen verdeutlicht, dass die Ebenen der Schulen und der Wohngebiete keine hierarchische Beziehung zueinander aufweisen.

Die Gruppierung der L1-Einheiten in mehrere L2-Ebenen hat auch Konsequenzen für die Konzeptualisierung der Stichprobenziehung und der darauf aufbauenden Inferenzen. Wenn im beschriebenen Beispiel eine Inferenz über den Effekt einer Eigenschaft der Schulen (z.B. dem durchschnittlichen Lehrer-Schüler-Verhältnis) auf den Lernerfolg der Schüler angestrebt wird, so beziehen wir uns auf die Stichprobe von Schulen, die aus einer Grundgesamtheit von Schulen stammt. Wenn uns darüber hinaus auch der Effekt einer Eigenschaft der Wohnorte (z.B. der Zahl an lernunterstützenden Einrichtungen in den Wohnorten) interessiert, wollen wir von einer Stichprobe von Wohnorten auf eine Grundgesamtheit von Wohnorten schließen. Um solche Inferenzschlüsse zu ermöglichen, müssen die Analyseverfahren nicht nur die Stichprobenziehung der Schüler als L1-Einheiten, sondern auch die Stichprobenziehungen der Schulen und der Wohnorte als L2-Einheiten angemessen berücksichtigen. Die Schulen und die Wohnorte müssen außerdem zumindest so ausgewählt sein, dass von beiden Einheiten eine ausreichende Fallzahl vorliegt und dass sie sich

in den Eigenschaften, für deren Effekte wir uns interessieren, in ausreichendem Maße unterscheiden.[65]

Diese Überlegungen können wir auf die RTR-Messungen während eines TV-Duells übertragen. Bereits im vorangegangen Kapitel haben wir dargestellt, dass die einzelnen RTR-Messungen (L1) innerhalb der Rezipienten, bei denen die Messungen vorgenommen werden (L2), gruppiert sind. Dieses Modell ist statistisch äquivalent zu einem latenten (Wachstumskurven-) Modell aus dem Strukturgleichungskontext, in dem je nach L1-Parametrisierung eine oder mehrere latente Variablen geschätzt werden, um die individuellen RTR-Verläufe zu beschreiben. Wenn wir jedoch nicht nur eine Antwort, sondern alle Turns[66] eines Kandidaten während des Duells untersuchen, führen wir eine zweite L2-Gruppierung ein. Die durch die Messmodelle beschriebenen individuellen RTR-Bewertungen sind auch in den Turns, zu denen sie abgegeben wurden, gruppiert. Wie im Beispiel der Schulen und Wohngebiete sind die Formulierungen, mit denen das Verhältnis der RTR-Verläufe zu Rezipienten und Turns beschrieben wird, austauschbar: Die RTR-Bewertungen eines Rezipienten beziehen sich auf verschiedene Turns, und die RTR-Bewertungen eines Turns kommen von verschiedenen Rezipienten.

Auch hier können wir einfach den Bezug zu der doppelten Stichprobenziehung einer TV-Duell-Studie – auf Ebene der Rezipienten und auf Ebene der Turns – herstellen. Wenn wir, wie es in TV-Duell-Studien häufig der Fall ist, allgemeine Aussagen über die Effekte von Eigenschaften der Rezipienten und Turns sowie ihrer Interaktionen prüfen wollen, betrachten wir die Probanden als Stichprobe aus einer Grundgesamtheit möglicher Rezipienten und die Turns als eine Stichprobe aus einer Grundgesamtheit möglicher Kandidatenaussagen (vgl. Kapitel 2). Wie wir im Folgenden zeigen werden, wird dieses Konzept der doppelten Stichprobenziehung in den kreuzklassifizierten Modellen der unmit-

65 Welcher Art die Stichprobenziehung auf den verschiedenen Ebenen sein muss, ist mehr eine philosophische als eine datenanalytische Frage. In einer idealen Welt lägen auf allen Ebenen Zufallsstichproben vor. In Studien der Bildungsforschung, in denen z.B. zuerst eine Zufallsstichprobe aus Schulen und dann innerhalb der Schulen Zufallsstichproben aus Schülern gezogen werden, ist dies möglich, wenn auch mit erheblichem Aufwand verbunden. Wenn wir die Inhalte der untersuchten TV-Debatte als Stichprobe aus einer Grundgesamtheit möglicher Debatteninhalte oder aus einer Grundgesamtheit möglicher politischer Aussagen der Kandidaten in diesem Wahlkampf betrachten, handelt es sich selbstverständlich nicht um eine Zufallsstichprobe, sondern um eine systematische Auswahl durch die Kandidaten und Journalisten als Gestalter des Duells. So betrachtet lassen sich die inferenzstatistischen Tests nicht im Sinne repräsentativer Inferenzen auf bestimmte Grundgesamtheiten verstehen (vgl. auch Kapitel 2).

66 Unter einem Turn verstehen wir die Zeitdauer, in der ein Kandidat das Wort hat, von dem Zeitpunkt, an dem ein Kandidat das Wort ergreift, bis zu dem Zeitpunkt, an dem die Sprecherrolle wechselt (vgl. ausführlicher die Erläuterung der Untersuchungseinheiten auf Ebene des Debatteninhalts auf S. 173f).

telbaren Kandidatenbewertungen umgesetzt. Den Probanden werden die vor dem Duell gemessenen Rezipienteneigenschaften zugeordnet, um Unterschiede zwischen ihren individuellen Bewertungen zu erklären. Den Turns werden Merkmale aus der systematischen Inhaltsanalyse der Debatte zugeordnet, um Unterschiede zwischen den Bewertungen während der verschiedenen Turns zu erklären. Und schließlich werden Interaktionen zwischen Voreinstellungen und Inhaltsmerkmalen modelliert, um ihr spezifisches Zusammenwirken abzubilden. So testen wir allgemeine Aussagen über die Effekte der Voreinstellungen, der Debatteninhalte und ihrer Interaktion. Beim Inferenzschluss auf die Rezipienten wird berücksichtigt, dass nicht alle Rezipienten mit einer Voreinstellung die Kandidaten genau gleichartig beurteilen und auf wie vielen Personen die Schätzung beruht. Genauso berücksichtigt der Inferenzschluss auf die Kandidatenaussagen die Zahl der untersuchten Turns und die Unterschiede der Kandidatenbewertungen zwischen ihnen.

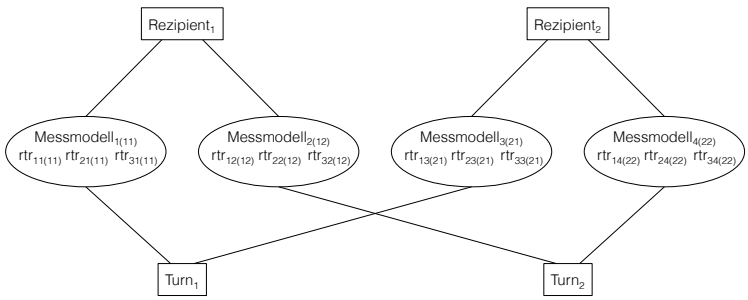
Tabelle 6.8: Auszug aus einem kreuzklassifiziertem Datensatz

	rtr	mm_id	rez_id	turn_id	lager	thema
1	−7	1	1	1	Unent.	Atom
2	−15	1	1	1	Unent.	Atom
3	−17	1	1	1	Unent.	Atom
4	12	2	1	2	Unent.	Schule
5	16	2	1	2	Unent.	Schule
6	22	2	1	2	Unent.	Schule
7	30	3	2	1	Lg. Mappus	Atom
8	30	3	2	1	Lg. Mappus	Atom
9	45	3	2	1	Lg. Mappus	Atom
10	40	4	2	2	Lg. Mappus	Schule
11	50	4	2	2	Lg. Mappus	Schule
12	50	4	2	2	Lg. Mappus	Schule

Anmerkungen
rtr: RTR-Messungen; mm_id, rez_id, turn_id: Identifikationsnummer des individuellen RTR-Verlaufs (Messmodell), des Rezipienten, des Turns; lager: Lagerzugehörigkeit des Rezipienten; thema: Thema des Turns.

Tabelle 6.8 stellt die kreuzklassifizierte Datenstruktur für zwei Rezipienten und zwei Turns dar. Um einen besseren Überblick zu gewährleisten, gehen wir vereinfachend davon aus, dass jeder Turn nur drei Sekunden dauert und damit drei RTR-Messungen jedes Rezipienten in jedem Turn vorliegen. In der Spalte der RTR-Messungen (rtr) wird deutlich, dass nur die einzelnen RTR-Messungen für jeden Fall variieren können. Die Ausprägungen aller anderen

Variablen wiederholen sich. Die einzelnen RTR-Messungen sind zuerst in einem Verlauf zusammengefasst, der durch ein Messmodell abgebildet wird (mm_id). Jedes Messmodell gehört zu genau einem Rezipienten (rez_id) und einem Turn (turn_id). Rechts sind beispielhaft die Rezipientenvariable Lagerzugehörigkeit (lager) und die Turnvariable Thema (thema) dargestellt. Sie sind für jedes Messmodell konstant und unterscheiden sich nur zwischen den Rezipienten bzw. Turns. Der kreuzklassifizierte Datensatz enthält damit im Gegensatz zum Panel-Datensatz oder zum Zeitreihen-Datensatz (vgl. Tabellen 5.1, S. 117 und 5.2, S. 123) sowohl Personen- als auch Inhaltscharakteristika. Auch bleiben die Unterschiede der RTR-Messungen sowohl innerhalb als auch zwischen Personen und Turns erhalten. Dies wird ermöglicht, indem jeder Rezipient bzw. jeder Turn (die Formulierungen sind austauschbar) so häufig im Datensatz vorkommt, wie er RTR-Messungen abgegeben bzw. erhalten hat. Dadurch entsteht in der typischen Betrachtung, in der die Zeilen eines Datensatzes die Fälle der Auswertung sind, eine sehr große Fallzahl. Diese Replikation der „wirklichen“ Fälle (je nach Analyserichtung die Zahl der Rezipienten, Turns oder Messmodelle) innerhalb des Datensatzes muss in den Modellen berücksichtigt werden – dazu sind kreuzklassifizierte Mehrebenenmodellen geeignet. Abbildung 6.14 visualisiert die Datenstruktur aus Tabelle 6.8 in einem Klassifikationsdiagramm, um die Mehrebenenstruktur der Kreuzklassifikation zu verdeutlichen.



Eigene Darstellung in Anlehnung an Snijders und Bosker (2011, S. 206).

Abbildung 6.14: Klassifikationsdiagramm der kreuzklassifizierten Datenstruktur für zwei Rezipienten und zwei Turns

Auf L1 werden die individuellen RTR-Messungen eines Rezipienten zu einem Turn mit einem Messmodell erfasst. Da wir in diesem Teilkapitel lediglich die Zuordnung der Messungen zu einem Turn betrachten und ihre zeitliche Abfolge innerhalb des Turns vernachlässigen, besteht das Messmodell aus einer einfachen Intercept-Only-Spezifikation. Es wird eine latente Variable geschätzt, die Auskunft über die mittlere RTR-Bewertung während eines Turns durch einen Rezipienten gibt. Die Messmodelle sind auf L2 in Rezipienten und Turns kreuzklassifiziert. Jedes Messmodell gehört sowohl zu genau einem Rezipienten, von dem die Bewertung stammt, als auch zu genau einem Turn, während dem die Bewertung abgegeben wird.

Formal ist das unkonditionale kreuzklassifizierte Modell, das noch keine Charakteristika der Rezipienten und Turns als L2-Prädiktoren berücksichtigt, damit wie folgt definiert.[67]

Auf Ebene der individuellen RTR-Messungen (L1) durch das Messmodell:

$$rtr_{th(ij)} = \pi_{0h(ij)} + e_{th(ij)} \tag{6.12}$$

$$\pi_{0h(ij)} = \pi_{0(ij)} + f_{h(ij)} \tag{6.13}$$

Wie wir anhand der zwei Gleichungen erkennen können, besteht die Schätzung des Messmodells statistisch betrachtet aus zwei Ebenen. Da es sich aber konzeptionell um ein Messmodel zur Beschreibung der individuellen RTR-Verläufe handelt, behalten wir die Begrifflichkeit L1 für die Messmodelle der RTR-Verläufe und L2 für die Ebenen der Rezipienten und Turns bei. Zuerst werden nach Gleichung 6.12 die individuellen RTR-Messungen beschrieben. Dabei ist $rtr_{th(ij)}$ eine individuelle RTR-Messung zum Zeitpunkt t in einem individuellen RTR-Verlauf h, der genau einem Rezipienten i und einem Turn j zugeordnet ist. $\pi_{0h(ij)}$[68] ist die mittlere Bewertung des individuellen RTR-Verlaufs h durch den Rezipienten i im Turn j und $e_{th(ij)}$ ist die Abweichung der Messung von der mittleren Bewertung $\pi_{0h(ij)}$ zum Zeitpunkt t. Das Subskript $h(ij)$ bedeutet, dass wir davon ausgehen, dass sich die mittleren Bewertungen der einzelnen Verläufe, der Rezipienten und der Turns voneinander unterscheiden. (ij) ist in

67 Die Notation folgt wiederum Hox (2010, S. 179-180).

68 Nach der Empfehlung von Hox (2010, S. 179) notieren wir alle Koeffizienten, die zwischen den Ebenen variieren, mit demselben griechischen Buchstaben und geben durch das Subskript an, zwischen welchen Ebenen sie variieren. Um eine Konsistenz zur Notation der Wachstumskurvenmodelle in Kapitel 6.1 herzustellen, behalten wir π als Bezeichnung für Koeffizienten zur Beschreibung der individuellen RTR-Messungen bei. Die L2-Koeffizienten bezeichnen wir analog zur obigen Darstellung mit β.

Klammern gesetzt, um anzuzeigen, dass die L2-Gruppierungen in Rezipienten und Turns nicht hierarchisch zueinander stehen.

Nach Gleichung 6.13 gehen wir davon aus, dass die mittleren RTR-Bewertungen $\pi_{0(ij)}$ in Abhängigkeit der Rezipienten i und der Turns j variieren. Die Fehlerkomponente $f_{h(ij)}$ ist die Abweichung zwischen den mittleren RTR-Bewertungen $\pi_{0h(ij)}$, die nicht alleine auf Eigenschaften der Rezipienten oder der Turns zurückgeht.

Auf den L2-Ebenen der Rezipienten und Turns können die in den Messmodellen erfassten mittleren Bewertungen der Rezipienten-Turn-Kombinationen $\pi_{0h(ij)}$ erklärt werden durch:

$$\pi_{0(ij)} = \beta_{00} + u_{0i} + v_{0j} \tag{6.14}$$

Dabei ist β_{00} die mittlere Bewertung aller Turns durch alle Rezipienten, u_{0i} die Abweichung der mittleren Bewertung des Rezipienten i vom Gesamtmittelwert aller Rezipienten und v_{0j} die Abweichung der mittleren Bewertung des Turns j vom Gesamtmittelwert aller Turns.

Durch Einsetzen von Gleichungen 6.13 und 6.14 in Gleichung 6.12 erhalten wir für das unkonditionale kreuzklassifizierte Intercept-Only-Modell zur Erklärung der individuellen RTR-Messungen:

$$rtr_{th(ij)} = \beta_{00} + u_{0i} + v_{0j} + f_{h(ij)} + e_{th(ij)} \tag{6.15}$$

Eine individuelle RTR-Messung ergibt sich demnach durch den Gesamtmittelwert, die durch Rezipienteneigenschaften bedingte Abweichung, die durch Turneigenschaften bedingte Abweichung, die durch die spezifische Kombination von Rezipient und Turn bedingte Abweichung und die Abweichung der einzelnen RTR-Messung vom Mittelwert des RTR-Verlaufs in der Rezipienten-Turn-Kombination.

Ein großer Nutzen des kreuzklassifizierten Intercept-Only-Modells ist es, die Varianz der Fehlerterme in mehrere Komponenten zergliedern und den einzelnen Ebenen zuordnen zu können. Die Varianz $\sigma^2_{e_{th(ij)}}$ ist die Streuung der einzelnen Messungen um die jeweiligen Mittelwerte der RTR-Verläufe. Sie kann als die Messfehlervarianz verstanden werden, die sich aus der Wahl der L1-Spezifikation ergibt. In diesem ersten kreuzklassifizierten Modell erfassen wir einen individuellen RTR-Verlauf in einem Turn durch eine Konstante, die

über den gesamten Verlauf des Turns denselben Wert annimmt (vgl. auch die Abbildungen zum Intercept-Only-Modell in Kapitel 6.1). Die Streuung um diesen Intercept herum kann in diesen Modellen nicht erklärt werden, da wir nur Prädiktoren berücksichtigen, deren Merkmale den L2-Einheiten Rezipienten und Turns zugeordnet sind. Sie variieren, wie in Tabelle 6.8 ersichtlich, nicht innerhalb der Messmodelle und können somit auch nicht erklären, warum eine einzelne RTR-Messung von der durch das Messmodell beschriebenen Konstanten abweicht. Wir können die Varianzkomponente $\sigma^2_{e_{th(ij)}}$ damit als modellbedingte Residualvarianz bezeichnen.

Ebenfalls auf Ebene der Messmodelle liegt die Varianzkomponente $\sigma^2_{f_{h(ij)}}$. Sie bezieht sich jedoch nicht auf die einzelnen RTR-Messungen, sondern ist die Fehlervarianz der durch die Messmodelle abgebildeten mittleren RTR-Verläufe, die weder den Rezipienten noch den Turns zugerechnet werden kann. In ihr finden sich zufällige Messfehler, situationsbedingte Abweichungen und alle nicht modellierten Interaktionen zwischen Rezipienten und Turns (Hox, 2010, S. 179). Die inhaltliche Bedeutung dieser Varianzkomponente ist vergleichsweise schwierig in wenige prägnante Worte zu fassen. Am nächsten kommt ihr die Beschreibung als Varianzkomponente, die dem Umstand Rechnung trägt, dass individuelle Rezipienten unterschiedlich auf verschiedene Turns reagieren.[69] Einen (geringen) Teil dieser Komponente können wir in den folgenden Modellen erklären, indem wir „cross-level interactions“ (Hox, 2010, S. 7, 20), das heißt Interaktionseffekte zwischen L2-Prädiktoren zu Rezipienten und Turns, modellieren. Allerdings sind für modellierte Interaktionen zwischen einzelnen Merkmalen keine allzu großen Varianzaufklärungen zu erwarten. Wir müssen bedenken, dass hier – neben dem im Umfang schwer einzuschätzenden zufälligen Messfehler – sämtliche Interaktionen zwischen allen latenten Merkmalen der Rezipienten und Turns enthalten sind. Hierzu gehören nicht nur Interaktionen zwischen den in der folgenden Analyse relevanten Merkmalen wie z.B. die Interaktion zwischen den politischen Voreinstellungen der Rezipienten und den Themen der Turns. Ebenso sind sämtliche denkbaren Interaktionen zwischen den distinkten Inhalten der einzelnen Turns (von rhetorischer Ausgestaltung über spezielle Formulierungen, die genauen in diesem Turn geäußerten Positionen bis hin zum visuellen und vokalem Auftreten der Kandidaten, vgl. z.B. Nagel, 2012) und den Eigenschaften der individuellen Rezipienten (z.B. „Vorlieben“ für bestimmte Formulierungen, Voreinstellungen zu konkreten Einzelpositionen) enthalten. Diese Interaktionen können für andere Forschungsfragen interessant sein, erfordern jedoch die Erfassung

69 Diese Umschreibung vernachlässigt jedoch den zufälligen Messfehler, der ja gerade *nicht* von Merkmalen der Rezipienten und Turns abhängt.

weiterer Rezipienten- und Turnmerkmale – teils in einem Detailgrad, der in der praktischen Umsetzung kaum realisierbar erscheint. Um dieses Argument greifbarer zu machen: Selbst wenn wir davon ausgehen, dass sich die Turns mit 20 Turnmerkmalen und die Rezipienten mit 20 Rezipientenmerkmalen perfekt beschreiben ließen, ergeben sich 400 potenzielle Interaktionseffekte. Von diesen Interaktionen können und wollen wir zur Beantwortung einer konkreten Forschungsfrage jedoch nur wenige explizit modellieren. Wir müssen dann auch nicht überrascht sein, wenn die Varianzaufklärung durch diese Modellierung verhältnismäßig klein ist, da wir nur einen Bruchteil aller potenziellen Interaktionen berücksichtigen.

Schließlich werden in dieser Varianzkomponente situationsbedingte Einflüsse, die sich auf einzelne Rezipienten-Turn-Kombinationen beziehen und die daher nicht durch die auf L2 gemessenen Variablen abgebildet werden können, wirksam. Darunter fällt z.B. die Aufmerksamkeit, die ein bestimmter Rezipient während eines bestimmten Turns dem Stimulus gegenüber aufbringt. Diese könnte z.B. gestört werden, wenn der Sitznachbar niest oder ein Getränk umstößt. Solche Störungen treten im Verlauf der gesamten Debatte bei verschiedenen Rezipienten während verschiedener Turns auf. Wir können sie als situationsbedingte Messfehler bezeichnen, die im vorliegenden Design weder erklärt noch kontrolliert werden können. Hierzu wäre eine zweite rezeptionsbegleitende Messung, etwa mit physiologischen Messverfahren oder einer systematischen Beobachtung der Probanden im Versuchsraum, notwendig. Die Gesamtheit der aufgezählten Einflüsse wird durch die latente Modellierung der Messmodellebene kontrolliert. Innerhalb dieser Varianzkomponente ist es jedoch unmöglich, *a priori* zu bestimmen, welchen Anteil wir durch die explizite Modellierung von Cross-Level-Interaktionen aufklären können.

Vergleichsweise einfach, inhaltlich jedoch ebenso relevant ist dagegen die Interpretation der beiden weiteren Varianzkomponenten, die den L2-Einheiten zugeordnet sind. Die Varianz $\sigma^2_{u_{0i}}$ kann durch Eigenschaften der Rezipienten (z.B. die Voreinstellungen zu den Kandidaten und Parteien), die Varianz $\sigma^2_{v_{0j}}$ durch Eigenschaften der Turns (z.B. das Thema oder die Relation) erklärt werden. Durch die Intraklassen-Korrelationen ρ (vgl. Formel (6.10), S. 196), die die Anteile der jeweiligen Komponenten an der gesamten Varianz quantifizieren, können wir eine Aussage über ihre relative Bedeutung zur Erklärung der unmittelbaren Kandidatenbewertungen treffen. Aus diesen Anteilen können wir auch ableiten, ob es sich lohnt, vermehrt Merkmale der Rezipienten oder der Debatteninhalte in die Modelle zur Erklärung der unmittelbaren Bewertungen aufzunehmen.

6.2.2 Bewertung der Kandidaten während aller Turns

Im Folgenden untersuchen wir mit kreuzklassifizierten Modellen, wie die Kandidaten in Abhängigkeit von den Themen der Turns und den Voreinstellungen der Rezipienten bewertet werden. Dazu ziehen wir die Turns in neun thematischen Blöcken des TV-Duells heran. Nicht untersucht werden die Bewertungen des Blocks zu Koalitionen und die Schlussstatements. Insgesamt stellen wir drei Auswertungen vor. Zuerst analysieren wir getrennt die Turns von Schmid und Mappus. Dann betrachten wir in einer kandidatenvergleichenden Perspektive das relative Abschneiden beider Kandidaten in den Themenblöcken. Auf Ebene des Inhalts werden alle Turns berücksichtigt, die mindestens 15 Sekunden dauern. Tabelle 6.9 stellt die Verteilung der Turns auf die Themenblöcke dar.

Tabelle 6.9: Verteilung der Turns auf die Themenblöcke

	Sprecher		
Thema	Schmid	Mappus	Gesamt
Atomkraft	5	6	11
EnBW	2	3	5
Arbeitsmarkt	4	5	9
Kiga/Kita	2	3	5
Schule	4	3	7
Studiengebühren	1	1	2
Finanzen	5	5	10
Persönliches	3	2	5
Stuttgart 21/Bürgerbeteiligung	6	6	12
Gesamt	32	34	66

Anmerkungen
Anzahl der Turns in einem Themenblock. Zu einer detaillierten Beschreibung der Verteilung der Themen auf die Sprecher vgl. Bachl, Käfferlein und Spieker (2013b).

Auf Ebene der Rezipienten werden alle Personen berücksichtigt, die in allen im Folgenden verwendeten Variablen zur Voreinstellung gültige Werte aufweisen (n = 172). Auf Ebene der Messmodelle werden die Verläufe aus der Analyse ausgeschlossen, in denen aufgrund technischer Störungen einzelne RTR-Messungen fehlen. Der so bereinigte Datensatz enthält insgesamt 10643 Messmodelle zu Rezipienten-Turn-Kombinationen, die insgesamt 455435 RTR-Messungen abbilden. Für jeden Turn liegen zwischen 152 und 169 Messmodelle vor ($M = 161$, $SD = 3$). Von jedem Rezipienten sind zwischen 15 und 66 Messmodelle enthalten ($M = 62$, $SD = 9$). 80 Prozent der Rezipienten sind

jedoch mit mehr als 60 von 66 Messmodellen vertreten, sodass wir zu den meisten Turns die RTR-Bewertungen von fast allen Rezipienten vorliegen haben. Der Ausfall von Messmodellen weist weder bezüglich der Charakteristika der betroffenen Rezipienten noch bezüglich der Themen der Turns eine Systematik auf.

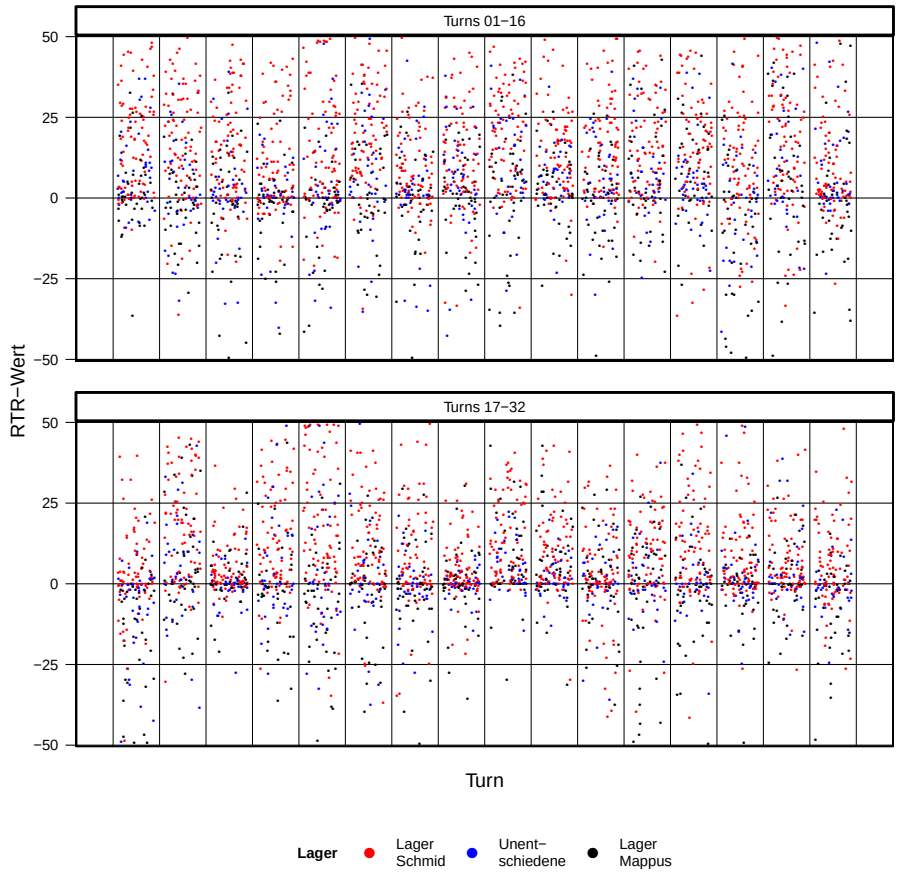
Für die Analyse der RTR-Bewertungen getrennt nach Sprecher wird die RTR-Skala für Schmid gedreht, um die Werte über die Sprecher hinweg leichter vergleichbar zu machen. Dadurch reicht die Skala für beide Sprecher von -50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Bei der Formulierung der L2-Spezifikationen gehen wir wie folgt vor: Zuerst schätzen wir ein unkonditionales Intercept-Only-Modell, um die Varianzkomponenten, die den Rezipienten, den Turns und deren Interaktion zuzuordnen sind, zu quantifizieren. Dann präsentieren wir Modelle, in denen die Lagerzugehörigkeit und die Themen als L2-Prädiktoren berücksichtigt werden. Schließlich nehmen wir weitere Rezipientencharakteristika in das Modell auf.

Erstes Beispiel: Bewertung von Nils Schmid

Varianzdekomposition Zuerst modellieren wir die Bewertung von Schmid als Funktion der Rezipienten- und Turnmerkmale. Die Vorgehensweise bei der Modellformulierung wird in diesem ersten Beispiel ausführlich beschrieben. Um einen weiteren Einblick in die Logik der kreuzklassifizierten Analyse zu erhalten, sind in den Abbildungen 6.15 und 6.16 die durch das unkonditionale Modell geschätzten RTR-Bewertungen $\pi_{0(ij)}$ aller Turns durch alle Rezipienten dargestellt. Beide Abbildungen enthalten dieselben Datenpunkte, die jedoch entsprechend der beiden Analyserichtungen des kreuzklassifizierten Modells unterschiedlich angeordnet sind.

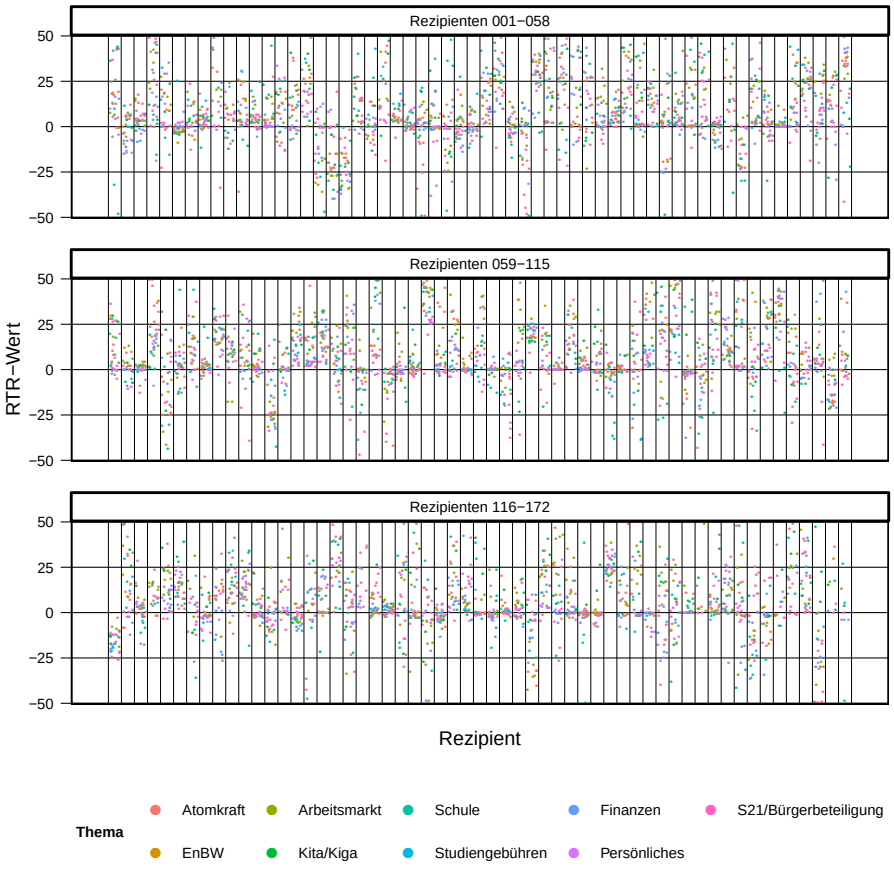
Abbildung 6.15 zeigt die Gruppierung der Bewertungen in den Turns. Um eine grobe Orientierung zu geben, um die Bewertungen welcher Rezipienten es sich handelt, sind sie farblich nach politischem Lager codiert. In dieser Darstellung können wir sehen, dass zwischen den Bewertungen der Rezipienten innerhalb der einzelnen Turns große Varianz besteht. Die farbliche Codierung lässt erahnen, dass die Lagerzugehörigkeit zur Erklärung dieser rezipientenspezifischen Varianz beiträgt. Die Anhänger Schmids bewerten ihn in der Regel besser, die Anhänger der Regierungsparteien schlechter. Doch auch zwischen den Bewertungen aus einem Lager ist einige Streuung zu erkennen. Hierbei handelt es sich um die Varianz, die bei der häufig angewendeten Aggregation über die Rezipienten in einem Lager hinweg verloren ginge.

In Abbildung 6.16 sind die Bewertungen den Rezipienten zugeordnet, die sie abgegeben haben. Die farbliche Codierung zeigt, welches Thema die einzelnen



Anmerkungen
Die Abbildung zeigt die durch das unkonditionale Modell geschätzten RTR-Bewertungen $\pi_{0(ij)}$ von Schmid während aller Turns durch alle Rezipienten auf einer Skala von -50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Entlang der x-Achse sind die Bewertungen nach Turns geordnet. Die Flächen innerhalb von zwei vertikalen Linien zeigen jeweils alle Bewertungen zu einem Turn. Innerhalb dieser Fläche sind die Bewertungen zufällig horizontal gestreut, um Überlagerungen zu reduzieren. Die Bewertungen sind farblich nach der Lagerzugehörigkeit der Rezipienten, die sie abgegeben haben, codiert.

Abbildung 6.15: Bewertung von Schmid während 32 Turns durch 172 Rezipienten, geordnet nach Turns



Anmerkungen
Die Abbildung zeigt die durch das unkonditionale Modell geschätzten RTR-Bewertungen $\pi_{0(ij)}$ von Schmid während aller Turns durch alle Rezipienten auf einer Skala von -50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Entlang der x-Achse sind die Bewertungen nach Rezipienten geordnet. Die Flächen innerhalb von zwei vertikalen Linien zeigen jeweils alle Bewertungen von einem Rezipienten. Innerhalb dieser Fläche sind die Bewertungen zufällig horizontal gestreut, um Überlagerungen zu reduzieren. Die Bewertungen sind farblich nach dem Thema des Turns, zu dem sie abgegeben wurden, codiert.

Abbildung 6.16: Bewertung von Schmid während 32 Turns durch 172 Rezipienten, geordnet nach Rezipienten

Turns haben. Im Vergleich zur Streuung innerhalb der Turns ist die Streuung innerhalb der Rezipienten verhältnismäßig gering. Die meisten Bewertungen des größten Teils der Rezipienten gehen in die gleiche Richtung. Abweichungen zwischen den Bewertungen eines Rezipienten finden sich vor allem auf derselben Seite der RTR-Skala. Die Abbildung vermittelt damit den Eindruck, dass das Thema (oder auch andere Eigenschaften des Turns) vergleichsweise wenig Unterschiede zwischen den Bewertungen eines Rezipienten verursachen.

Diese visuelle Inspektion entspricht der Logik einer kreuzklassifizierten Analyse. Wir betrachten zum einen, wie stark die RTR-Bewertungen in Abhängigkeit von den Rezipienten, von denen sie stammen, streuen, wenn sie jeweils denselben Turn bewerten. Zum anderen untersuchen wir die Streuung innerhalb der Bewertungen verschiedener Turns durch jeweils dieselben Rezipienten, um den relativen Einfluss der Turnmerkmale einzuschätzen. Insgesamt können wir aus dem Vergleich der Abbildungen den Eindruck gewinnen, dass den Merkmalen der Rezipienten eine relativ große Bedeutung zur Erklärung der Unterschiede in den individuellen RTR-Bewertungen zukommt. Die Effekte der Turneigenschaften sind dagegen eher begrenzt. Zwischen den Bewertungen der Turns durch einen Rezipienten bestehen nur recht geringe Unterschiede. Visuell kaum einschätzen lässt sich schließlich die Bedeutung der Interaktionen von Rezipienten- und Turncharakteristika.

Den Eindruck der Visualisierung können wir mithilfe der Varianzdekomposition des unkonditionalen kreuzklassifizierten Modells quantifizieren. Dazu dient die Intraklassen-Korrelation ρ, die den Anteil einer Varianzkomponente an der gesamten Varianz der individuellen RTR-Messungen angibt. Die Varianzkomponenten des unkonditionalen Modells M0 sind in der ersten Spalte von Tabelle 6.10 (S. 231) dargestellt. Es bestätigt sich, dass die Eigenschaften der Turns für nur wenig Streuung zwischen den Bewertungen derselben Rezipienten sorgen. Die Varianzkomponente der Turns macht gerade einmal vier Prozent aus. Im Vergleich dazu haben die Eigenschaften der Rezipienten einen recht großen Einfluss auf die Bewertungen. Die rezipientenbezogene Varianzkomponente kommt auf einen Anteil von immerhin 24 Prozent. Der Varianzanteil der Bewertungen zwischen den Messmodellen, der sich nicht den Personen- oder Turneigenschaften zuordnen lässt, beträgt 40 Prozent. Auf die L1-Residuen, also die Streuung der individuellen RTR-Messungen um die durch das Messmodell geschätzten mittleren RTR-Bewertungen in den Rezipienten-Turn-Kombinationen, entfallen schließlich die verbleibenden 32 Prozent der Varianz. Dieser Varianzanteil kann nicht durch die L2-Prädiktoren erklärt werden, was sich auch daran erkennen lässt, dass $\sigma^2_{\text{L1-Residuum}}$ in allen folgenden Modellen M1 bis M4, die um den Einfluss von Thema, Lagerzugehörigkeit und deren Interaktion erweitert werden, unverändert bleibt. Wenn wir

diese modellbedingt nicht erklärbare Residualvarianz ausschließen, verhalten sich die durch L2-Prädiktoren erklärbaren Varianzkomponenten zueinander wie folgt: Rezipientenbezogene Varianzkomponente 36 Prozent, turnbezogene Varianzkomponente fünf Prozent, messmodellbezogene Varianzkomponente 59 Prozent.

Effekte des Themas und der Lagerzugehörigkeit In den folgenden Analysen zeigen wir, wie mit kreuzklassifizierten Modellen die gleichzeitige Berücksichtigung von Rezipienten- und Turncharakteristika sowie ihrer Interaktionen möglich ist. Zuerst ziehen wir das Thema des Turns und die Lagerzugehörigkeit der Rezipienten als L2-Prädiktoren zur Erklärung der Bewertungen heran. Im ersten Teil von Tabelle 6.10 sind die Kennzahlen zum Vergleich der Modelle dargestellt. Dabei ist M0 das unkonditionale Nullmodell. In M1 wird nur das Thema des Turns, in M2 nur die Lagerzugehörigkeit berücksichtigt. M3 enthält die einfachen Effekte beider L2-Prädiktoren, M4 zusätzlich ihre Interaktion. Durch passende Vergleiche der Modellfits können Hypothesen zum Einfluss der beiden Faktoren getestet werden. Zum Einfluss des Themas des Turns auf die unmittelbaren Bewertungen stellen wir zunächst nur die einfache, ungerichtete Unterschiedshypothese auf:

> **H1**: Die unmittelbaren Bewertungen von Schmid unterscheiden sich in Abhängigkeit vom Thema des Turns.

Hypothese 1 kann für den bivariaten Zusammenhang zwischen Thema und Bewertung durch einen Vergleich des unkonditionalen Modells M0 mit Modell M1, das nur das Thema als L2-Prädiktor enthält, untersucht werden. Als angemessenes Informationskriterium für den Vergleich der Modelle kann zum einen das $AICc_m$ herangezogen werden.[70] Durch die Aufnahme des Faktors Thema sinkt der Wert des $AICc_m$ nicht ab. Demzufolge wird die Verbesserung der Modellgüte, die durch den L2-Prädiktor erreicht wird, durch den Anstieg der Modellkomplexität vollständig aufgehoben. In Abwägung von Datenpassung und Sparsamkeit des Modells lohnt sich die Aufnahme des Themas der Turns zur Erklärung der unmittelbaren Bewertungen nicht, demnach wäre Hypothese 1 zu verwerfen. Zum anderen kann die Modellverbesserung durch

70 Die in Kapitel 6.1 beschriebenen Probleme zur Wahl der „korrekten" Fallzahl für die Korrektur des AIC sind in der kreuzklassifizierten Datenstruktur nochmals komplexer, da auf mehreren (hier: 4) Ebenen unterschiedliche Fallzahlen vorliegen. Eine Simulationsstudie von Beretvas und Murphy (2013) zeigt, dass die Korrektur um die Fallzahl m der kreuzklassifizierten Ebene mit der größeren Fallzahl die geringste Fehlerrate bei der Identifikation des korrekten Modells aufweist. Demzufolge korrigieren wir den AIC-Wert hier um die Fallzahl von 172 Rezipienten und bezeichnen die Kennzahl als $AICc_m$.

den neuen Faktor durch einen LR-Test auf Signifikanz getestet werden. Eine Reduzierung der Deviance um 17 ergibt bei 8 zusätzlichen Freiheitsgraden einen Signifikanz-Wert von $p = .029$. Bei einer Orientierung am konventionellen Signifikanzniveau von $p < .05$ widerspricht dieser Befund dem Vergleich der $AICc_m$. Das Thema des Turns scheint signifikant zur Erklärung der Bewertungen beizutragen.

Tabelle 6.10: Vergleich der Modelle zur Erklärung der Bewertung von Schmid während seiner Turns durch Thema und Lagerzugehörigkeit

	M0	M1	M2	M3	M4
Modellvergleich					
$AICc_m$	1785746	1785746	1785688	1785689	1785650
Deviance	1785735	1785718	1785674	1785656	1785574
ΔDeviance (χ^2)		17	62	79	161
Freiheitsgrade		8	2	10	26
Signifikanz		.029	<.001	<.001	<.001
Varianzkomponenten					
σ^2_{Turns}	13.92	10.67	13.93	10.67	10.67
$\sigma^2_{\text{Rezipienten}}$	93.29	93.30	64.34	64.33	64.49
$\sigma^2_{\text{Messmodelle}}$	153.86	153.86	153.85	153.85	151.77
$\sigma^2_{\text{L1-Residuum}}$	123.69	123.69	123.69	123.69	123.69
Reduzierung der Varianzkomponenten					
$R^2_{2\text{Turns}}$		.23	.00	.23	.23
$R^2_{2\text{Rezipienten}}$		.00	.31	.31	.31
$R^2_{1\text{Messmodelle}}$		.00	.00	.00	.01

Anmerkungen
M0: unkonditionales Modell; M1, M2, M3: Modelle mit L2-Prädiktoren Thema, Lagerzugehörigkeit, bzw. beiden Prädiktoren; M4: Modell mit beiden L2-Prädiktoren und ihrer Interaktion.
Informationskriterium $AICc_m$, Kenngrößen von LR-Tests, Varianzen der Random Effects (σ^2) und Reduzierung der Varianzkomponenten (R^2_1 bzw. R^2_2) im Vergleich zu M0.
$n_{\text{Turns}} = 32$, $n_{\text{Rezipienten}} = 172$, $n_{\text{Messmodelle}} = 5187$, $n_{\text{RTR-Messungen}} = 230512$.

Da sich LR-Test und $AICc_m$-Modellvergleich widersprechen und der LR-Test bei der Evaluation von Fixed Effects eine erhöhte α-Fehlerrate aufweisen kann (z.B. Manor & Zucker, 2004; Pinheiro & Bates, 2000), testen wir den Effekt des Faktors Thema zur Absicherung zusätzlich mit einem konditionalen F-

Test.[71] Der Test ergibt, dass das Thema keinen signifikanten Effekt auf die unmittelbaren Bewertungen hat: $F(8, 19) = 2.08, p = .090$. Insgesamt sprechen die Daten gegen Hypothese 1. Der Effekt des Themas ist zu klein oder zu variabel, um ihn mit einer akzeptablen statistischen Sicherheit nachzuweisen. Da das Thema in diesem Modell eine Eigenschaft der Turns ist und alle Merkmale der Turns gemeinsam ohnehin nur für fünf Prozent der erklärbaren Varianz verantwortlich sind, verwundert dieses Ergebnis nicht. Das Thema kann mit einem $R^2_{2\text{Turns}} = .23$ zwar auf den ersten Blick beachtliche 23 Prozent der turnbezogenen Varianzkomponente erklären. Da sich diese Varianzreduzierung jedoch nur auf einen fast schon zu vernachlässigenden Anteil an der gesamten Varianz bezieht, fällt sie global betrachtet kaum ins Gewicht.

Zum Einfluss der Lagerzugehörigkeit nehmen wir wiederum zuerst die einfache, ungerichtete Hypothese an:

> **H2**: Die unmittelbaren Bewertungen von Schmid unterscheiden sich in Abhängigkeit von der Lagerzugehörigkeit der Rezipienten.

Der Vergleich der Modelle M2 und M3 mit Modellen M0 bzw. M1 offenbart, dass die Lagerzugehörigkeit wesentlich zur Erklärung der RTR-Bewertungen von Schmid beiträgt. Die Werte des $AICc_m$ sind jeweils deutlich geringer, und die LR-Tests zeigen hochsignifikante Modellverbesserungen. Auch der Erklärungsbeitrag der Lagerzugehörigkeit ist substantiell bedeutsam. Die Varianz in der rezipientenbezogenen Komponente wird um 31 Prozent reduziert. Da diese Varianzkomponente immerhin 36 Prozent der erklärbaren Varianz ausmacht, können wir der Lagerzugehörigkeit auch in Bezug auf die gesamte Streuung der unmittelbaren Bewertungen einen wesentlichen Effekt zusprechen. Hypothese 2 wird damit von den Daten gestützt.

Schließlich ist davon auszugehen, dass die Kandidaten, hier Nils Schmid, mit Aussagen zu verschiedenen Themen von den Rezipienten in Abhängigkeit von ihrer Lagerzugehörigkeit unterschiedlich bewertet werden. Wir erwarten also einen Interaktionseffekt zwischen der Lagerzugehörigkeit der Rezipienten und dem Thema des Turns:

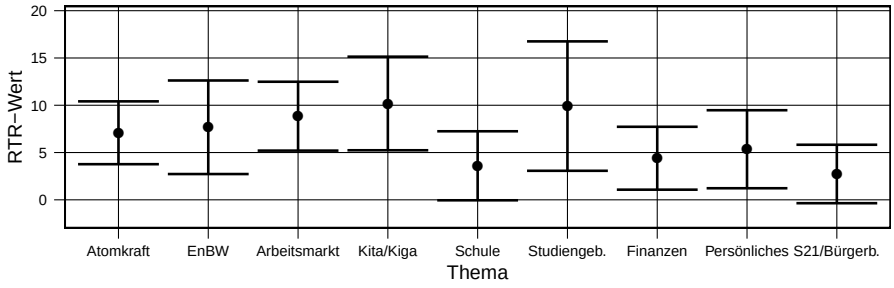
> **H3**: Die unmittelbaren Bewertungen von Schmid unterscheiden sich in Abhängigkeit von Kombinationen von Thema und Lagerzugehörigkeit.

71 F-Tests (und ebenso T-Tests) der Fixed Effects sind im Kontext von nicht hierarchischen Mehrebenenmodellen nicht statistisch trivial, da die Nennerfreiheitsgrade nicht exakt zu bestimmen sind. Wir nutzen hier eine Satterthwaite-Approximation der Nennerfreiheitsgrade, die nach einer Simulationsstudie von Manor und Zucker (2004) verlässliche Ergebnisse liefert. Die Approximation der Nennerfreiheitsgrade wird mit dem *R* Paket *lmerTest* (Kuznetsova, Brockhoff & Christensen, 2013a, 2013b) vorgenommen.

Bereits die Varianzdekomposition zeigt, dass der größte Teil der Streuung in den unmittelbaren Bewertungen auf spezifische Kombinationen von Turn- und Rezipientenmerkmalen zurückgeht. Modell M4, das mit der Cross-Level-Interaktion von Thema und Lagerzugehörigkeit eine solche Kombination enthält, ist nach den $AICc_m$-Werten das angemessenste Modell für die Daten. Der LR-Test von M4 im Vergleich zu M3 weist eine signifikante Modellverbesserung aus. Allerdings ist die Varianzaufklärung mit einem Prozent recht gering. Wir müssen bei der Interpretation des $R^2_{1\,\text{Messmodelle}}$ jedoch auch die vielen möglichen Quellen dieser Varianzkomponente bedenken (vgl. ausführlich S. 222ff.), von der wir gerade einmal *eine* Interaktion explizit modelliert haben. In Anbetracht dessen ist eine geringe Reduzierung der mit 59 Prozent der erklärbaren Varianz größten Varianzkomponente nicht überraschend.

Bisher haben wir nur geklärt, ob und in welchem Umfang die L2-Prädiktoren Lagerzugehörigkeit und Thema zur Erklärung der unmittelbaren Bewertungen von Schmid beitragen. Um zu untersuchen, in welche Richtung die Prädiktoren wirken, könnten wir zum einen die Koeffizienten der Fixed Effects der Modelle M3 und M4 in Tabelle A.7 betrachten. Besser geeignet ist jedoch eine visuelle Inspektion der durch das Modell vorhergesagten Kandidatenbewertungen. Um die absoluten Bewertungen in Abhängigkeit vom Thema des Turns einschätzen zu können, müssen die vorhergesagten RTR-Bewertungen unter konstanten Werten der Lagerzugehörigkeit ermittelt werden. Diese sind in Abbildung 6.17 dargestellt. Vereinfachend nehmen wir für diese Darstellung an, dass die Prädiktoren *Lg. Schmid* und *Lg. Mappus* jeweils die Ausprägung $\frac{1}{3}$ annehmen. Zwar existiert dieser Wert in den Daten nicht, er repräsentiert jedoch die relative Zugehörigkeit zu einer der drei Gruppen (Gelman & Hill, 2006, S. 56). Damit können Vorhersagen auf dieser Basis als plausible Schätzung für die Bewertung Schmids unabhängig von der Zugehörigkeit bzw. Nicht-Zugehörigkeit zu einem Lager betrachtet werden. Die so vorhergesagten Werte entsprechen konzeptionell der Bildung eines gewichteten Mittelwerts, bei dem die Bewertungen durch alle drei Gruppen gleich gewichtet sind.

Abbildung 6.17 offenbart, dass die Bewertung von Schmid bei einigen Themen etwas besser, bei anderen Themen etwas schlechter ist, die Unterschiede sich jedoch in Grenzen halten. Die Bewertungen in den thematischen Blöcken zu *Schule*, *Finanzen* und *Stuttgart 21/Bürgerbeteiligung* fallen etwas weniger positiv aus als die in anderen Themenblöcken. Der Vergleich der Konfidenzintervalle deutet an, dass einzelne Unterschiede durchaus auf einem Niveau von $p < .05$ signifikant sein könnten. Jedoch müsste für solche *a posteriori* Vergleiche die α-Fehler-Kumulation, die durch das multiple Testen auftritt, korrigiert werden, was zu deutlich weiteren Konfidenzintervallen führte (Bortz & Schuster, 2010,



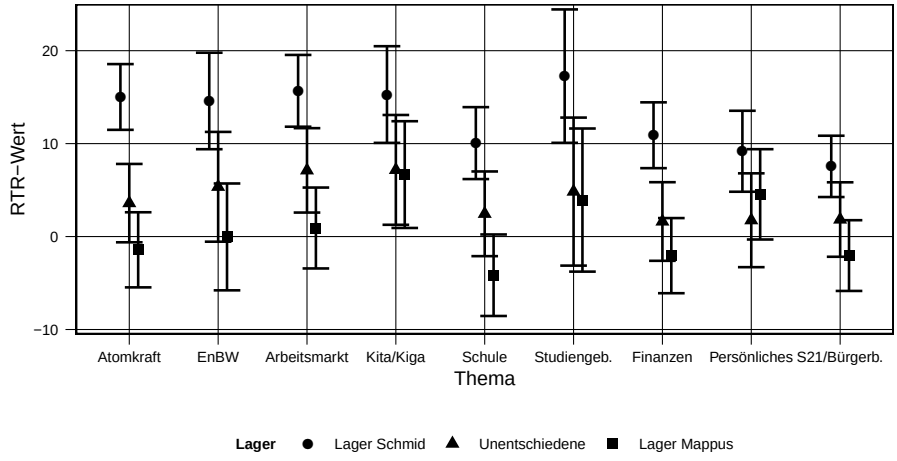
Anmerkungen
Bewertung von Schmid auf einer Skala von −50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch Modell M3 in Tabelle A.7. Die Fehlerbalken zeigen 95%-Konfidenzintervalle. Die Ausprägungen der Prädiktoren der Lagerzugehörigkeit *Lg. Schmid* und *Lg. Mappus* sind auf $\frac{1}{3}$ gesetzt.

Abbildung 6.17: Effekt des Themas auf die Bewertung von Schmid während seiner Turns (M3)

S. 232-235) und sämtliche Unterschiede jenseits des konventionellen Signifikanzniveaus zeigte. Später werden wir zeigen, wie durch eine Neuspezifikation des Modells *a priori* formulierte Hypothesen zum Effekt bestimmter Themen auf die Bewertung des Kandidaten gezielt getestet werden können.

Auch die Bewertung Schmids in Abhängigkeit von den Interaktionen von Thema und Lagerzugehörigkeit ließen sich aus Tabelle A.7 entnehmen (M4). Allerdings ist die direkte Interpretation der Koeffizienten in einem Modell mit Interaktionseffekten kaum zielführend. Weder bezeichnen sie in den meisten Fällen eine Kenngröße, die uns inhaltlich interessiert, noch lassen sie auf einen Blick erkennen, wie die Bewertung Schmids bei einem Thema durch eine Rezipientengruppe im Verhältnis zur RTR-Skala ausfällt (Brambor et al., 2006). Daher bietet sich hier die visuelle Inspektion der durch das Modell vorhergesagten RTR-Bewertungen umso mehr an (Abbildung 6.18).

Die Abbildung offenbart zunächst das erwartete Muster: Schmid wird von seinen Anhängern durchgehend am besten bewertet, die Anhänger von Mappus bewerten ihn am negativsten. Die Wertungen der Unentschiedenen sind dazwischen angeordnet. Hierin zeigt sich der starke Effekt des Faktors Lager. Die Bewertungen schwanken jedoch innerhalb der Gruppen zwischen den Themen, und auch die Differenzen zwischen den Gruppen bei einem Thema



Anmerkungen
Bewertung von Schmid auf einer Skala von −50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch Modell M4 in Tabelle A.7. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.18: Effekte des Themas und der Lagerzugehörigkeit auf die Bewertung von Schmid während seiner Turns (M4)

unterscheiden sich teils deutlich. In den ersten vier thematischen Blöcken des Duells wird Schmid von seinen Anhängern etwa gleich positiv bewertet. Die Bewertung durch die Unentschiedenen und die Anhänger von Mappus wird im Verlauf dieser vier Blöcke besser, die Differenz zwischen den Gruppen nimmt dadurch ab. Im vierten Themenblock zur *Kinderbetreuung* liegen die vorhergesagten Bewertungen durch alle drei Gruppen jenseits des neutralen Skalenmittelpunkts. Bei den Themen *Schule*, *Finanzen* und *Stuttgart 21/Bürgerbeteiligung* ist eine Art negativer „Haupteffekt“ des Themas zu beobachten (vgl. auch Abbildung 6.17). Die Bewertungen durch sämtliche Gruppe liegen niedriger als in den ersten vier Themenblöcken. Doch auch hier sind die Differenzen innerhalb des Themas unterschiedlich. Kaum Unterschiede nach der Lagerzugehörigkeit finden sich dagegen in der Phase, in der über die persönlichen Charakteristika der Kandidaten gesprochen wird. Einen Sonderfall aus datenanalytischer Sicht stellt schließlich das Thema Studiengebühren dar. Da

von Schmid (und ebenso von Mappus) nur ein Turn zu diesem Thema vorliegt, muss die gesamte Varianz dieses Turns bei der Schätzung dem Standardfehler des Fixed Effects zugeschlagen werden. Daher sind die Schätzungen hier nur sehr unpräzise, was sich an den weiten Konfidenzintervallen zeigt.

Effekte des Issue Ownership und der Lagerzugehörigkeit Mit den bisher präsentierten kreuzklassifizierten Modellen haben wir – wenn auch mit angemessener Repräsentation der Datenstruktur – lediglich die einfachen Auswertungen nachvollzogen, die auch mit einer einfachen Zusammenfassung der RTR-Messungen zu Themenblöcken zu leisten sind (Bachl, 2013a, S. 154-158). Die flexible Struktur der Mehrebenenmodelle erlaubt es jedoch auch auf einfache Weise, konkrete, gerichtete Hypothesen über die Wirkung bestimmter Themen auf die Bewertung der Kandidaten zu testen. So argumentieren wir in unserer ausführlichen Analyse der unmittelbaren Bewertung der Kandidaten im TV-Duell 2011 auf Basis des Issue-Ownership-Ansatzes, dass Schmid vor allem dann besonders gut bewertet wird, wenn er zu Themen spricht, bei denen er den Issue Owner vertritt (Bachl, 2013a, S. 138-139). Anhand der traditionell von den Parteien besetzten Themen (Franzmann, 2006), einer Analyse der politischen Stimmung vor dem TV-Duell (Bachl & Brettschneider, 2013; Vögele, 2013) und der Inhaltsanalyse der Debatte (Bachl, Käfferlein & Spieker, 2013b) vermuten wir, dass Schmid vor allem bei den Themen *Arbeitsmarkt* und *Kita/Kiga* punkten kann. Beide Themen wurden im Duell vor allem unter dem Gesichtspunkt der sozialen Gerechtigkeit diskutiert. In Abwesenheit eines Kandidaten der Grünen dürfte er zudem als Stellvertreter des rot-grünen Lagers als Issue Owner in der Diskussion über *Atomkraft* angesehen werden.

Zur Interaktion mit der Lagerzugehörigkeit erwarten wir, dass Schmid durch sein eigenes Lager bei „seinen" Themen nochmals besser bewertet wird. Zudem erwarten wir wie bisher, dass die Anhänger des rot-grünen Lagers Schmid besser bewerten als die Unentschiedenen, und die Anhänger des Regierungslagers Schmid schlechter bewerten als die Unentschiedenen. Damit formulieren wir die Hypothesen:[72]

> **H4a**: Schmid wird von den Anhängern des Lager Mappus schlechter bewertet als von den Unentschiedenen (technisch: unkonditionaler negativer Effekt der Zugehörigkeit zum Lager Mappus).

72 Angesichts der umfangreichen induktiven Analysen, die wir bereits zu dieser Debatte durchgeführt haben, ist die Herleitung dieser Erwartungen sicherlich durch unser Wissen um die empirischen Ergebnisse verzerrt. Da es hier aber vor allem um die Demonstration der Eignung des analytischen Vorgehens für die Überprüfung solcher Hypothesen geht, wollen wir diesen theoretischen Mangel an dieser Stelle vernachlässigen.

H4b: Schmid wird bei Themen, bei denen er *nicht* den Issue Owner vertritt, von seinen Anhängern besser bewertet als von den Unentschiedenen (technisch: konditionaler positiver Effekt der Zugehörigkeit zum Lager Schmid unter der Bedingung, dass das Thema kein Issue-Owner-Thema von Schmid ist).

H4c: Schmid wird bei Themen, bei denen er den Issue Owner vertritt, von den Unentschiedenen und den Anhängern des Lager Mappus besser bewertet als bei Themen, bei denen er *nicht* den Issue Owner vertritt (technisch: konditionaler positiver Effekt des Issue Ownership unter der Bedingung, dass der Rezipient *nicht* zum Lager Schmid gehört).

H4d: Der positive Effekt des Issue Ownership ist bei Angehörigen des Lagers Schmid größer als bei Rezipienten, die nicht dem Lager Schmid angehören (technisch: konditionaler positiver Effekt des Issue Ownership unter der Bedingung, dass der Rezipient zum Lager Schmid gehört).

Formal werden die Annahmen des Issue-Ownership-Ansatzes in die drei Hypothesen H4b bis H4d übersetzt. Zusätzlich wird dem Forschungsstand und den bisherigen Ergebnissen Rechnung getragen, indem ein vom Issue Ownership unabhängiger negativer Effekt der Zugehörigkeit zum Lager des Gegenkandidaten angenommen wird. Tabelle 6.11 und Abbildung 6.19 fassen das Modell zum Test der Hypothesen zusammen.

Hypothese 4a wird durch den Koeffizienten von *Lg. Mappus* gestützt: Angehörige des Regierungslagers bewerten Schmid durchschnittlich um 3.8 Skalenpunkte schlechter als die Unentschiedenen. Der Koeffizient von *Lg. Schmid* quantifiziert den konditionalen Effekt der Zugehörigkeit zum Oppositionslager in Kontrast zu den Unentschiedenen, wenn Schmid sich zu einem Thema äußert, bei dem er nicht den Issue Owner vertritt. Seine Anhänger bewerten ihn in Übereinstimmung mit Hypothese 4b bei diesen Themen um 7.6 Skalenpunkte besser als die Referenzgruppe der Unentschiedenen. Der Effekt des Issue Ownership auf die Rezipienten, die keine Schmid-Anhänger sind (H4c), kann dem Koeffizienten von *Issue Owner Schmid* entnommen werden. Schmid wird als Issue Owner von diesen Rezipienten mit 2.5 Skalenpunkten leicht, aber statistisch signifikant positiver bewertet. Schließlich stützt der Koeffizient der Interaktion *Lg. Schmid X Issue Owner Schmid* auch Hypothese 4d. Zusätzlich zu den 2.5 Skalenpunkten, welche das Issue Ownership schon bei den übrigen Rezipienten bringt, steigert sich die Bewertung Schmids in seinem eigenen Lager nochmals um 2.7 Skalenpunkte. Die Annahmen auf Basis

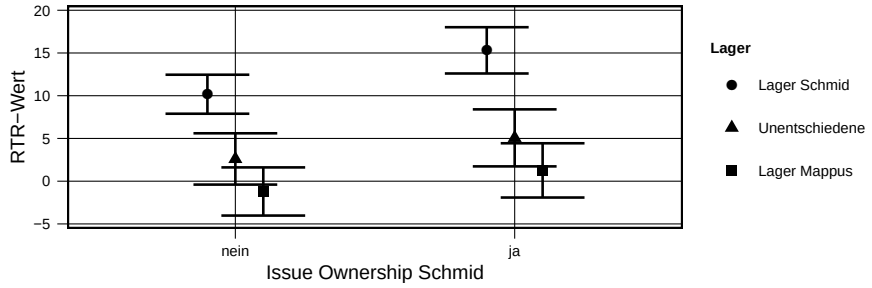
Tabelle 6.11: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die Bewertung von Schmid während seiner Turns

	β	s.e.	t	df	p
Intercept	2.60	1.53	1.70	184	.045
Lager Mappus	−3.81	1.82	−2.09	169	.019
Lager Schmid	7.57	1.63	4.64	178	<.001
Issue Owner Schmid	2.47	1.34	1.85	35	.036
Lg. Schmid X Issue Owner Schmid	2.67	0.73	3.64	4984	<.001

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade; p: einseitige Tests.
Modell: $n_{\text{Turns}} = 34$, $n_{\text{Rezipienten}} = 172$, $n_{\text{Messmodelle}} = 5456$, $n_{\text{RTR-Messungen}} = 224923$.
$\sigma^2_{\text{Turns}} = 10.92$, $\sigma^2_{\text{Rezipienten}} = 64.39$, $\sigma^2_{\text{Messmodelle}} = 153.47$, $\sigma^2_{\text{L1-Residuum}} = 123.69$.
$AICc_m = 1785671$.

des Issue-Ownership-Ansatzes werden damit für die Bewertung von Schmid von den Daten gestützt. Allerdings müssen wir wiederum anmerken, dass die Reduzierung der messmodellbezogenen Varianzkomponente im Vergleich zum unkonditionalen Mo mit $R^2_{1\text{Messmodelle}} = .003$ sehr gering ist. Dies ist nicht überraschend, wenn wir bedenken, auf welch einfache Weise die unterschiedliche Wahrnehmung verschiedener Turns durch individuelle Rezipienten hier modelliert wird. Wir müssen uns bewusst machen, dass wir durch die hier vorgeschlagenen Erklärungsansätze eben nur einen kleinen Teil der individuellen Variationen erklären können – auch wenn wir einen substantiellen Beitrag zur Erklärung der durchschnittlichen Bewertung der Kandidaten leisten.

Effekte des Issue Ownership und der Voreinstellungen Als letzten Schritt nehmen wir nun eine Erweiterung des rezipientenseitigen Modells vor. Die Varianzdekomposition hat ergeben, dass die Charakteristika der Rezipienten bessere Ansatzpunkte für die Erklärung der unmittelbaren Bewertung von Schmid liefern. Daher berücksichtigen wir mit der Voreinstellung der Rezipienten gegenüber den Kandidaten und ihren Parteien hierzu weitere L2-Prädiktoren. Zur Erweiterung wählen wir das gerade berichtete Modell zum Issue Ownership, in dem wir prüfen, ob Schmid als Issue Owner besser bewertet wird und ob der Effekt des Issue Ownership von den Voreinstellungen ihm und seiner Partei gegenüber moderiert wird. Konkret nehmen wir als weitere L2-Prädiktoren zum einen die mit Skalometer-Fragen gemessenen Voreinstellungen gegenüber Schmid und der SPD auf. Zum anderen berücksichtigen



Anmerkungen
Bewertung von Schmid auf einer Skala von −50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch das Modell in Tabelle 6.11. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.19: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die Bewertung von Schmid während seiner Turns

wir analog zum obigen Modell die Voreinstellungen gegenüber Mappus und der CDU. Beim Test eines vollen Modells (nicht dargestellt) erweisen sich die Zugehörigkeit zum Lager Mappus und die Voreinstellung zur CDU als zur Voreinstellung gegenüber Mappus redundante Prädiktoren – sie werden daher im finalen Modell nicht berücksichtigt. Tabelle 6.12 stellt das finale Modell vor.

Die Voreinstellung zum Gegenkandidaten Mappus hat den erwartet negativen Effekt auf die Bewertung von Schmid: Je besser (schlechter) ein Rezipient Mappus vor dem Duell bewertet, desto schlechter (besser) ist der Eindruck, den er während des Duells von Schmid hat. Die Zugehörigkeit zum Lager von Schmid sowie die Voreinstellungen ihm und seiner Partei gegenüber weisen erwartungskonform einen positiven Zusammenhang mit seiner Beurteilung während der Turns, in denen er nicht den Issue Owner vertritt, auf. Der Einfluss der Voreinstellung zur SPD ist jedoch nicht statistisch signifikant. Die Interpretation des Koeffizienten des Issue Ownership ist in Präsenz der konditionalen Effekte der quasi-metrischen Skalometer-Variablen nochmals komplexer. Der Koeffizient von $\beta = 2.7$ gibt an, dass Schmid als Issue Owner von Rezipienten, die nicht dem rot-grünen Lager angehören *und* sowohl Schmid als auch der SPD neutral gegenüberstehen (also in diesen Variablen die Ausprägung 0 haben), um 2.7 Punkte auf der RTR-Skala besser bewertet wird.

Tabelle 6.12: Effekte des Issue Ownership und der Voreinstellungen auf die Bewertung von Schmid während seiner Turns

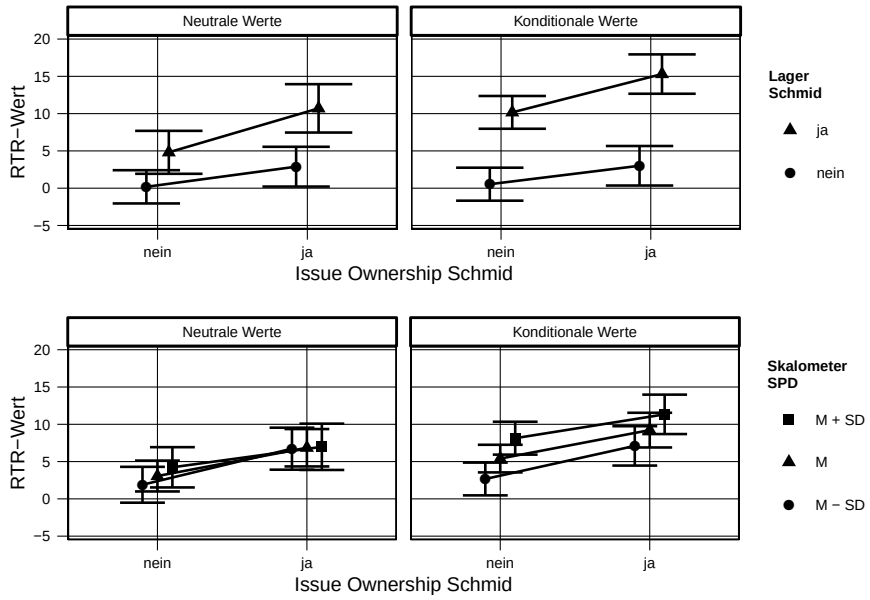
	β	s.e.	t	df	p
Intercept	0.18	1.14	0.16	118	.873
Skalometer Mappus	−1.23	0.25	−4.85	165	<.001
Lager Schmid	4.63	1.51	3.06	176	.003
Skalometer Schmid	0.95	0.45	2.13	190	.034
Skalometer SPD	0.53	0.35	1.53	191	.128
Issue Owner Schmid	2.70	1.34	2.02	35	.050
Lg. Schmid X Issue Owner Schmid	3.19	0.76	4.20	4982	<.001
Sk. Schmid X Issue Owner Schmid	0.01	0.28	0.05	5001	.960
Sk. SPD X Issue Owner Schmid	−0.48	0.22	−2.20	4997	.028

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.
Modell: $n_{\text{Turns}} = 34$, $n_{\text{Rezipienten}} = 172$, $n_{\text{Messmodelle}} = 5456$, $n_{\text{RTR-Messungen}} = 224923$.
$\sigma^2_{\text{Turns}} = 10.94$, $\sigma^2_{\text{Rezipienten}} = 55.86$, $\sigma^2_{\text{Messmodelle}} = 153.30$, $\sigma^2_{\text{L1-Residuum}} = 123.69$.
$AICc_m = 1785649$.

Der konditionale Effekt der Zugehörigkeit zum Oppositionslager auf die Bewertung von Schmid als Issue Owner ist auch in diesem erweiterten Modell positiv. Der Effekt der Voreinstellung gegenüber Schmid wird nicht signifikant vom Issue Ownership moderiert. Der Koeffizient der Interaktion *Sk. SPD X Issue Owner Schmid* ist mit $\beta = -0.48$ signifikant negativ. Daraus kann jedoch nicht analog der Interpretation eines einfachen additiven Regressionsmodells geschlossen werden, dass eine positive Voreinstellung zur SPD zu einer negativen Bewertung von Schmid führt. Als konditionaler Effekt entzieht er sich dieser einfachen Interpretation und muss stattdessen im Verhältnis zu den Effekten erster Ordnung betrachtet werden. Um die substantielle Interpretation von konditionalen Effekten in Regressionsmodellen zu vereinfachen, empfiehlt sich die Berechnung und grafische Darstellung der Effekte für realistische Wertebereiche der Variablen von Interesse (Brambor et al., 2006). Abbildung 6.20 visualisiert die beiden signifikanten Interaktionseffekte.[73]

Der konditionale Effekt der Zugehörigkeit zum Lager Schmid entspricht den Befunden des oben dargestellten Modells, das nur die Lagerzugehörigkeit als Rezipientenmerkmal berücksichtigt. Schmid wird von seinen Anhängern besser bewertet als von den übrigen Rezipienten. Dieser positive Effekt steigert

73 Zur Vorhersage der RTR-Werte auf Basis neutraler und konditionaler Werte der übrigen L2-Prädiktoren vgl. die ausführliche Erläuterung zu Abbildung 6.9 (S. 207).



Anmerkungen
Bewertung von Schmid auf einer Skala von -50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch das Modell in Tabelle 6.11. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.
Vorhergesagte RTR-Werte beim Mittelwert (M) und ± 1 Standardabweichung (SD) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf einen neutralen Wert gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt.

Abbildung 6.20: Effekte des Issue Ownership und der Voreinstellungen auf die Bewertung von Schmid während seiner Turns

sich nochmals, wenn Schmid zu einem Thema spricht, bei dem er den Issue Owner vertritt. Die untere Grafik in Abbildung 6.20 hilft, die Bedeutung des negativen Koeffizienten der Interaktion der Voreinstellung zur SPD mit dem Issue Ownership einzuordnen. Inhaltlich entspricht der negative Koeffizient einem Dämpfungseffekt, der den positiven Einfluss der Voreinstellung zur SPD auf die Bewertung von Schmid bei Turns, in denen er nicht als Issue Owner auftritt, nivelliert. Gut beobachten lässt sich dies in der linken Darstellung, in der die Einflüsse der übrigen Prädiktoren ausgeblendet werden. Unter der Bedingung „Schmid nicht Issue Owner" besteht ein – wenn auch recht schwacher – positiver Zusammenhang zwischen der Voreinstellung zur SPD und der Bewertung von Schmid. Wenn Schmid als Issue Owner spricht, liegen die vorhergesagten RTR-Werte für alle Ausprägungen der Voreinstellung zur SPD fast exakt auf einer Linie. Hier macht es also keinen Unterschied mehr, wie die Rezipienten zur SPD stehen. Betrachten wir diesen Zusammenhang allerdings unter für die Stichprobe typischen Verteilungen aller L2-Prädiktoren (Konditionale Werte in Abbildung 6.20), so stellen wir fest, dass diese Interaktion im Verhältnis zu den weiteren Effekten substantiell von untergeordneter Bedeutung ist.

Während die übrigen Effekte den Erwartungen des Issue-Ownership-Ansatzes entsprechen, passt der konditionale Effekt der Voreinstellung zur SPD auf den ersten Blick nicht zum erwarteten Muster. Berücksichtigen wir jedoch, dass mit der Zugehörigkeit zum Lager Schmid und der Voreinstellung zum Kandidaten zwei weitere Effekte in diese Richtung wirken, lässt er sich als partieller Effekt sinnvoll interpretieren. Die Zugehörigkeit zum rot-grünen Lager und die Voreinstellung zu Schmid bewirken bereits eine bessere Bewertung von Schmid während des Duells, der erstgenannte Effekt wird auch wie erwartet vom Issue Ownership verstärkt. Die Voreinstellung zur SPD wirkt nach diesen Effekten nur noch als „Bonus", der für eine etwas bessere Bewertung von Schmid sorgt, wenn nicht ein ohnehin für ihn vorteilhaftes Thema besprochen wird.

Insgesamt verbessert sich durch die Aufnahme der zusätzlichen Rezipientenmerkmale die Passung des Modells zu den Daten. Das $AICc_m$ sinkt im Vergleich zum vorangegangen Modell um 22 Einheiten. Entsprechend der Modellerweiterung kann vor allem die rezipientenbezogene Varianzkomponente reduziert werden. Es ergibt sich im Vergleich zum unkonditionalen Null-Modell ein $R^2_{2\text{Rezipienten}}$ von .40, was einem $\Delta R^2_{2\text{Rezipienten}}$ von .13 gegenüber dem Vormodell zum Issue Ownership entspricht. Es lohnt sich also, auf Seiten der Rezipientenmerkmale eine über die einfache Lagerzuordnung hinausgehende Modellspezifikation zu wählen. Auch die Erklärungskraft des Modells bezüglich der Interaktionskomponente steigt leicht auf $R^2_{1\text{Messmodelle}} = .004$ ($\Delta R^2_{1\text{Messmodelle}} = .001$) an, bleibt jedoch absolut betrachtet weiterhin sehr gering.

Zweites Beispiel: Bewertung von Stefan Mappus

Im folgenden Abschnitt erklären wir analog zum gerade beschrieben Vorgehen die Bewertung von Mappus während des Duells durch Rezipienten- und Turnmerkmale. Da die grundsätzliche Logik der kreuzklassifizierten Analyse bereits ausführlich beschrieben wurde, fassen wir diesen Teil etwas knapper und beschränken uns auf eine Darstellung der inhaltlich bedeutsamen Modelle. Tabelle 6.13 fasst die Kennzahlen der Modellpassung und die Varianzkomponenten der Random Effects zusammen. Dabei ist M0 das unkonditionale Null-Modell. M1 enthält die Faktoren Thema und Lagerzugehörigkeit als L2-Prädiktoren. M2 testet den Issue-Ownership-Ansatz mit dem Faktor Lagerzugehörigkeit. In M3 wird M2 um weitere Rezipientenmerkmale ergänzt.

Varianzdekomposition Die Varianzdekomposition des unkonditionalen Modells der Bewertung von Mappus bestätigt in ihrer Struktur die Befunde zur Bewertung von Schmid. Die turnbezogene Varianzkomponente macht wiederum nur vier Prozent der gesamten Varianz bzw. fünf Prozent der erklärbaren Varianz aus. Alle Merkmale der Rezipienten sind dagegen für immerhin 30 Prozent der gesamten bzw. 42 Prozent der erklärbaren Varianz verantwortlich. Dieser im Vergleich zur Bewertung Schmids noch etwas größere Varianzanteil, der den Rezipienten zuzurechnen ist, spricht für eine stärker von der Voreinstellung geprägte Beurteilung von Mappus während des Duells. Mit 37 Prozent der gesamten bzw. 53 Prozent der erklärbaren Varianz hat die Komponente der einzelnen Messmodelle wiederum den größten Anteil. Die modellbedingt nicht erklärbare L1-Residualvarianz stellt einen Anteil von 30 Prozent an der gesamten Varianz.

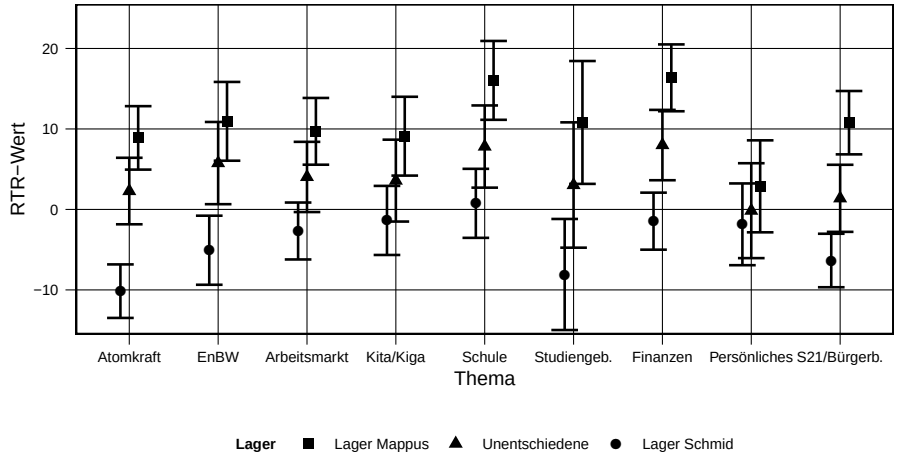
Effekte des Themas und der Lagerzugehörigkeit Modell M1 mit den L2-Prädiktoren Thema und Lagerzugehörigkeit ist nach $AICc_m$ und LR-Test eine Verbesserung gegenüber dem unkonditionalen Modell M0. Sowohl das Thema des Turns ($F(8, 22) = 3.13$, $p = .016$) als auch die Lagerzugehörigkeit ($F(2, 175) = 41.30$, $p < .001$) tragen signifikant zur Erklärung der Bewertung von Mappus bei. Sie erklären jeweils ca. ein Drittel der Varianz in der ihnen zugeordneten Varianzkomponente. Hier ist wieder zu beachten, dass eine Varianzreduzierung um ein Drittel der rezipientenbezogenen Komponente wesentlich bedeutsamer ist als bei der turnbezogenen Komponente. Die Interaktion der Faktoren ist ebenfalls hochsignifikant ($F(16, 5132) = 6.56$, $p < .001$), kann aber nur einen geringen Teil (2%) der Varianz zwischen den Messmodellen aufklären. Abbildung 6.21 zeigt den Einfluss der L2-Prädiktoren auf die Bewertung von Mappus.

Tabelle 6.13: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während seiner Turns durch Turn- und Rezipientenmerkmale

	M0	M1	M2	M3
Modellvergleich				
$AICc_m$	1752073	1751943	1752003	1751970
Deviance	1752063	1751867	1751983	1751939
ΔDeviance (χ^2)		196	79	123
Freiheitsgrade		26	4	9
Signifikanz		$<.001$	$<.001$	$<.001$
Varianzkomponenten				
σ^2_{Turns}	14.98	9.53	12.81	12.79
$\sigma^2_{\text{Rezipienten}}$	125.90	84.08	84.04	66.34
$\sigma^2_{\text{Messmodelle}}$	156.37	153.71	156.23	156.20
$\sigma^2_{\text{L1-Residuum}}$	128.57	128.57	128.57	128.57
Reduzierung der Varianzkomponenten				
$R^2_{2\text{Turns}}$		.36	.14	.15
$R^2_{2\text{Rezipienten}}$		.33	.33	.47
$R^2_{1\text{Messmodelle}}$		.02	.001	.001

Anmerkungen
M0: unkonditionales Modell; M1: L2-Prädiktoren Thema und Lagerzugehörigkeit; M2: L2-Prädiktoren Issue Ownership Mappus und Lagerzugehörigkeit; M3: L2-Prädiktoren Issue Ownership Mappus und verschiedene Rezipientenmerkmale
Informationskriterium $AICc_m$, Kenngrößen der LR-Tests des Vergleichs mit M0, Varianzen der Random Effects (σ^2) und Reduzierung der Varianzkomponenten (R^2_1 bzw. R^2_2) im Vergleich zu M0.
$n_{\text{Turns}} = 34$, $n_{\text{Rezipienten}} = 172$, $n_{\text{Messmodelle}} = 5456$, $n_{\text{RTR-Messungen}} = 224923$.

Der Effekt des politischen Lagers tritt deutlich hervor. In fast allen Themenblöcken (mit Ausnahme des Debattenabschnitts zu Persönlichem) liegen die Bewertungen von Anhängern der Regierung und der Opposition recht weit auseinander. Die Urteile der Unentschiedenen liegen durchweg zwischen diesen Gruppen und tendieren mal stärker zu der einen, mal stärker zu der anderen Gruppe. Auf einen Blick wird der zentrale Unterschied zur Bewertung von Schmid (vgl. Abbildung 6.18) deutlich. Mappus wird von den Anhängern der Oppositionsparteien bei einer Reihe von Themen klar negativ bewertet – für Schmid trifft dies nur für das Thema *Schule* zu, und selbst hier nur knapp.



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell M2. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.21: Effekte des Themas und der Lagerzugehörigkeit auf die Bewertung von Mappus während seiner Turns (M2)

Die Interaktion zwischen den Faktoren drückt sich in den variablen Distanzen zwischen den Gruppen in einem Themenblock und den variablen Bewertungen der Themenblöcke durch die Gruppen aus. So sind die Wertungen der Gruppen bei den Themen *Atomkraft* und *Stuttgart 21/Bürgerbeteiligung* besonders stark polarisiert – die Bewertung durch die Anhänger der Opposition ist hier klar negativ, während Mappus' eigenes Lager positiv urteilt. Auch bei den *finanz- und schulpolitischen* Diskussionen ist der Abstand zwischen den Gruppen beträchtlich – hier jedoch vor allem wegen der besonders großen Zustimmung des eigenen Lagers im Kontrast zu neutralen Wertungen der Oppositionsanhänger.

Effekte des Issue Ownership und der Lagerzugehörigkeit Neben der Interaktion der Voreinstellungen mit dem neunstufigen Faktor Thema soll auch der Einfluss des Issue Ownership auf die Bewertung von Mappus getestet werden. Aus unseren Vorüberlegungen (Bachl, 2013a; Bachl & Brettschneider,

2013; Vögele, 2013) geht hervor, dass im Themenblock *Finanzpolitik* Vorteile für Mappus zu erwarten sind, da die CDU hier traditionell als Issue Owner gilt (Franzmann, 2006). Das starke Abschneiden von Mappus im Themenblock zur Schulpolitik (Bachl & Vögele, 2013) war dagegen weder auf Grundlage langjähriger Themenbesetzung noch wegen kurzfristiger Kompetenzzuschreibungen in den Umfragen vor der Debatte zu erwarten. Der Faktor *Issue Ownership Mappus* entspricht daher in den folgenden Modellen lediglich dem Vorkommen des Themas *Finanzpolitik*. Immerhin fallen fünf der 34 Turns in diesen Themenblock, sodass zumindest eine ausreichende Präzision bei der Schätzung des Effekts gegeben sein sollte. Wie oben für Schmid nehmen wir nun für Mappus an, dass er von den Anhänger der Oppositionsparteien negativer bewertet wird als von den Unentschiedenen (H1a); auch als Nicht-Issue-Owner von seinen Anhängern positiver bewertet wird als von den Unentschiedenen (H1b); als Issue Owner von den Unentschiedenen und den Oppositionsanhängern besser bewertet wird (H1c); und der positive Effekt des Issue Ownership in seinem Lager noch stärker zu tragen kommt (H1d). Die zur Prüfung der Hypothesen notwendigen Koeffizienten und ihre Tests sind in Tabelle 6.14 berichtet.

Tabelle 6.14: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die Bewertung von Mappus während seiner Turns (M3)

	β	s.e.	t	df	p
Intercept	3.47	1.65	2.10	195	.019
Lager Schmid	−8.40	1.82	−4.61	168	<.001
Lager Mappus	6.58	2.06	3.19	170	.001
Issue Owner Mappus	3.80	1.82	2.08	33	.023
Lg. Mappus X Issue Owner Mappus	2.50	1.11	2.26	5258	.012

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade; p: einseitige Tests.
Die Angaben zu Varianzkomponenten und Modellpassung finden sich in Tabelle 6.13 (M3).

Wie erwartet wird Mappus von den Oppositionsanhängern deutlich negativer und von den Regierungsanhängern auch als Nicht-Issue-Owner deutlich positiver bewertet als von den Unentschiedenen (H1a und H1b gestützt). Ebenso zeigt sich, dass die Anhänger des rot-grünen Lagers und die Unentschiedenen Mappus etwas besser bewerten, wenn er zu „seinem“ Thema Finanzpolitik spricht (H1c gestützt). Schließlich fällt die Zustimmung seines eigenen Lagers bei diesem Thema nochmals stärker aus (H1d gestützt). Die Annahmen des Issue-Ownership-Ansatzes helfen also auch bei der Erklärung der unmittelba-

ren Urteile über Mappus weiter. Modell M3 ist nach $AICc_m$ und LR-Test eine Verbesserung gegenüber dem unkonditionalen Modell M0. Bemerkenswert ist, dass alleine durch die Berücksichtigung der Themenausprägung *Finanzpolitik* 14 Prozent der turnbezogenen Varianzkomponente erklärt werden können. Auf dieser Klassifikationsebene ist zwar insgesamt nur wenig Varianz vorhanden. Diese lässt sich jedoch mit wenigen einfachen Prädiktoren zu erheblichen Teilen erklären.

Effekte des Issue Ownership und der Voreinstellungen Schließlich soll auch das Issue-Ownership-Modell zur Erklärung der Bewertung von Mappus um weitere Voreinstellungen der Rezipienten erweitert werden. Das ermittelte Modell M4 umfasst unkonditional die Voreinstellungen zu Schmid und zur SPD sowie in Interaktion mit dem Issue Ownership die Zugehörigkeit zum Lager Mappus und die Voreinstellungen zu Mappus und zur CDU. Die Erweiterung des Modells führt zu einer weiteren Reduzierung der rezipientenbezogenen Varianz um 21 Prozentpunkte, sodass nun fast die Hälfte dieser Varianzkomponente erklärt wird. Dieser Befund deutet an, dass die politischen Voreinstellungen bei der Bewertung des Duellauftritts von Mappus eine wichtigere Rolle spielen als bei der Bewertung von Schmid. Viele allgemeine politische Urteile sind offenbar schon vor dem Duell mit der Person Mappus verknüpft und werden von den Rezipienten abgerufen, wenn sie den Kandidaten während des Duells bewerten. Wie die einzelnen Voreinstellungen die Bewertung von Mappus (in Interaktion mit dem Issue Ownership) beeinflussen, ist in Tabelle 6.15 dargestellt.

Eine positivere Voreinstellung zu Schmid und zur SPD führt zu einer negativeren Bewertung von Mappus während des Duells. Es muss allerdings beachtet werden, dass die Koeffizienten der unkonditionalen Effekte das Signifikanzniveau von $p < .05$ nicht erreichen. Da das Entfernen eines der beiden Prädiktoren zu einer signifikanten Verschlechterung des Modells führt, werden beide Koeffizienten im Modell beibehalten. Die Voreinstellungen zu Mappus und zur CDU haben einen signifikant positiven Einfluss auf die RTR-Bewertung. Nicht mehr signifikant ist nach der Modellerweiterung dagegen der Koeffizient des Issue Ownership auf die Bewertung durch die Unentschiedenen und Oppositionsanhänger. In Anwesenheit der Interaktionen mit den Skalometer-Variablen zu Mappus und der CDU bezieht sich dieser Koeffizient nur noch auf die Personen, die gegenüber diesen beiden eine neutrale Voreinstellung haben. Auch wenn die konditionalen Effekte nicht signifikant sind, so reicht ihre Aufnahme ins Modell trotzdem aus, um den ohnehin recht unpräzise geschätzten Effekt des Issue Ownership zu marginalisieren. Der einzige si-

Tabelle 6.15: Effekte des Issue Ownership und der Voreinstellungen auf die Bewertung von Mappus während seiner Turns (M4)

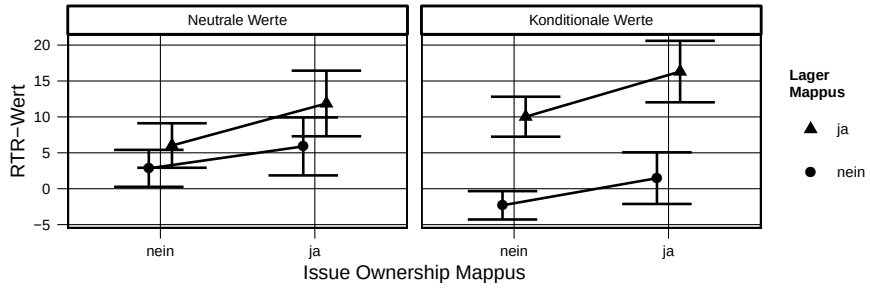
	β	s.e.	t	df	p
Intercept	2.83	1.31	2.16	174	.032
Skalometer Schmid	−0.94	0.48	−1.96	170	.052
Skalometer SPD	−0.73	0.38	−1.93	171	.055
Lager Mappus	3.18	1.83	1.74	168	.084
Skalometer Mappus	1.29	0.32	4.02	167	<.001
Skalometer CDU	1.09	0.31	3.57	168	<.001
Issue Owner Mappus	3.06	1.89	1.62	38	.113
Lg. Mappus X Issue Owner Mappus	2.80	1.30	2.16	5254	.031
Sk. Mappus X Issue Owner Mappus	−0.37	0.23	−1.58	5256	.114
Sk. CDU X Issue Owner Mappus	0.29	0.22	1.31	5256	.190

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.
Die Angaben zu Varianzkomponenten und Modellpassung finden sich in Tabelle 6.13 (M4).

gnifikante konditionale Effekt, die Interaktion der Zugehörigkeit zum Lager Mappus mit dem Issue Ownership, ist in Abbildung 6.22 illustriert. Wenn Mappus zu „seinem“ Thema Finanzpolitik spricht, wird er tendenziell von allen Rezipienten besser bewertet. Größer ist der Effekt bei seinen eigenen Anhängern. Von diesen wird er ohnehin positiver bewertet, das Issue Ownership sorgt dann für einen zusätzlichen Schub in der Zustimmung.

Drittes Beispiel: Die vergleichende Bewertung der Kandidaten

Als letztes Beispiel der einfachen kreuzklassifizierten Modelle untersuchen wir, wie die Kandidaten relativ zueinander bewertet werden. Damit kehren wir zu der auch in den verbreiteten Analysen der RTR-Zeitreihen angewandten Logik zurück: Die RTR-Skala wird als eine Differentialskala von −50 = „größter Vorteil Schmid“ bis 50 = „größter Vorteil Mappus“ interpretiert. Dabei beziehen wir alle Turns von Schmid und Mappus (n_{Turns} = 66) in die Analyse ein. Der Fokus der Betrachtung liegt auf dem relativen Abschneiden der Kandidaten. In der Analyse der Bewertungen in Abhängigkeit von Thema und Lagerzugehörigkeit replizieren wir unsere einfache Auswertung auf Basis der Mittelwert-Aggregation über die Themenblöcke hinweg (Bachl, 2013a, S. 157). Danach untersuchen wir die Annahmen des Issue-Ownership-Ansatzes. In einem zusätzlichen Modell nehmen wir den Sprecher des Turns



Anmerkungen
Bewertung von Mappus auf einer Skala von -50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle 6.15. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.
Vorhergesagte RTR-Werte beim Mittelwert (M) und ± 1 Standardabweichung (SD) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf einen neutralen Wert gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt.

Abbildung 6.22: Effekte des Issue Ownership und der Zugehörigkeit zum Lager Mappus auf die Bewertung von Mappus während seiner Turns (M4)

als erklärende Variable auf. Damit fassen wir die Analysen, die wir in den ersten beiden Beispielen präsentiert haben, in einem Modell zusammen und ermöglichen eine vergleichende Darstellung. Schließlich erweitern wir das Issue-Ownership-Modell wieder um weitere Merkmale der Rezipienten als Prädiktoren. Tabelle 6.16 fasst einführend die Modellpassung und die Varianzkomponenten der in diesem Abschnitt behandelten Modelle zusammen. M0 ist das unkonditionale Modell. M1 erklärt die Bewertung der Kandidaten durch die Faktoren Thema und Lagerzugehörigkeit. M2 prüft die Annahmen des Issue-Ownership-Ansatzes in Kombination mit der Lagerzugehörigkeit. In M3 wird zusätzlich berücksichtigt, welcher der beiden Kandidaten gerade das Wort hat. M4 erweitert M2 schließlich um weitere L2-Prädiktoren zu den Voreinstellungen der Rezipienten.

Varianzdekomposition Auch wenn wir die Bewertung beider Kandidaten gemeinsam betrachten, ergibt die Zergliederung der Varianz im unkonditionalen Modell M0 das bereits vertraute Bild. 21 Prozent der gesamten bzw. 30 Prozent der erklärbaren Varianz liegt zwischen den Rezipienten, nur acht bzw.

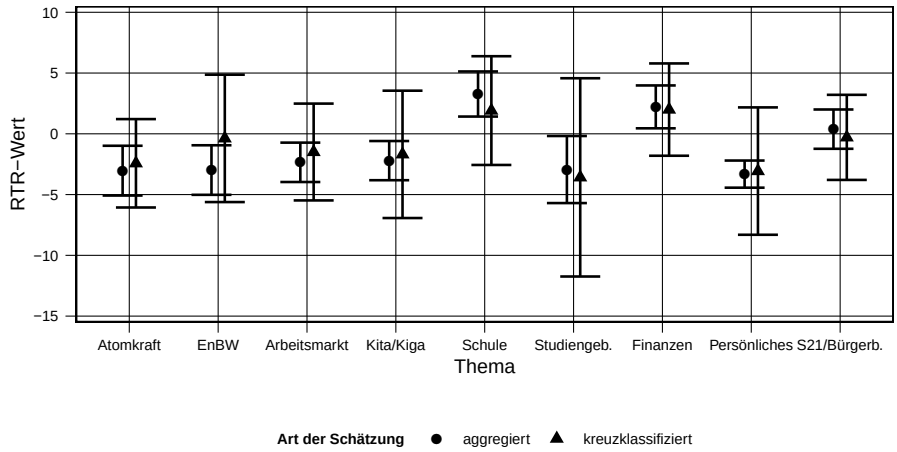
Tabelle 6.16: Vergleich der Modelle zur Erklärung der relativen Kandidatenbewertung während aller Turns durch Turn- und Rezipientenmerkmale

	M0	M1	M2	M3	M4
Modellvergleich					
$AICc_m$	3538715	3538553	3538630	3538559	3538582
Deviance	3538705	3538476	3538606	3538519	3538544
ΔDeviance (χ^2)		228	99	186	161
Freiheitsgrade		26	6	13	12
Signifikanz		<.001	<.001	<.001	<.001
Varianzkomponenten					
σ^2_{Turns}	31.86	32.71	31.10	12.09	31.11
$\sigma^2_{\text{Rezipienten}}$	89.92	54.59	54.55	54.55	41.08
$\sigma^2_{\text{Messmodelle}}$	174.89	172.84	174.73	174.37	174.56
$\sigma^2_{\text{L1-Residuum}}$	126.10	126.10	126.10	126.10	126.10
Reduzierung der Varianzkomponenten					
$R^2_{2\text{Turns}}$		−.03	.02	.62	.02
$R^2_{2\text{Rezipienten}}$		.39	.39	.39	.54
$R^2_{1\text{Messmodelle}}$		.01	.001	.003	.002

Anmerkungen
M0: unkonditionales Modell; M1: L2-Prädiktoren Thema und Lagerzugehörigkeit; M2: L2-Prädiktoren Issue Ownership und Lagerzugehörigkeit; M3: L2-Prädiktoren Issue Ownership, Lagerzugehörigkeit und Sprecher; M4: L2-Prädiktoren Issue Ownership, Lagerzugehörigkeit und verschiedene Rezipientenmerkmale.
Informationskriterium $AICc_m$, Kenngrößen der LR-Tests des Vergleichs mit M0, Varianzen der Random Effects (σ^2) und Reduzierung der Varianzkomponenten (R^2_1 bzw. R^2_2) im Vergleich zu M0.
n_{Turns} = 66, $n_{\text{Rezipienten}}$ = 172, $n_{\text{Messmodelle}}$ = 10643, $n_{\text{RTR-Messungen}}$ = 455435.

elf Prozent zwischen den Turns. Der etwas höhere Anteil der turnbezogenen Varianzkomponente erklärt sich, wie wir noch nachweisen werden, durch die Analyse der Turns beider Kandidaten in einem Modell. Die Varianzkomponente der Messmodelle ist mit einem Anteil von 41 bzw. 59 Prozent auch hier am größten. Die nicht durch L2-Prädiktoren erklärbare L1-Residualvarianz macht 30 Prozent der gesamten Varianz aus.

Effekte von Thema und Lagerzugehörigkeit Im ersten Modell M1 untersuchen wir, welcher Kandidat bei welchem Thema und welcher Zuschauergruppe im Vorteil ist. Dabei stellt sich heraus, dass das Thema kein bedeutsamer Prädiktor der relativen Kandidatenbewertung ist: $F(8, 53) = 0.8$, $p = .608$. Dies können wir anschaulich visualisieren, indem wir die RTR-Werte für die Themen bei neutralen Werten der Lagerzugehörigkeit vorhersagen (Abbildung 6.23).



Anmerkungen
Relative Bewertung der Kandidaten auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell M2 und mittlere Bewertung der Themenblöcke nach der Zusammenfassung über Messzeitpunkte in Bachl (2013a, S. 157). Die Ausprägungen der Prädiktoren der Lagerzugehörigkeit *Lg. Schmid* und *Lg. Mappus* sind auf $\frac{1}{3}$ gesetzt. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.23: Effekt des Themas auf die relative Kandidatenbewertungen während aller Turns (M2)

Daneben stellt diese Grafik einen Vergleich mit unserer einfachen, auf der Zusammenfassung der RTR-Messungen zu Themenblöcken basierenden Auswertung (Bachl, 2013a, gewichtete Gesamtmittelwerte in Abbildung 7, S. 157) dar.

Die durch das kreuzklassifizierte Modell geschätzten Mittelwerte weisen klar auf den geringen Einfluss des Themas hin. In keinem einzigen Themenblock weicht die Bewertung signifikant vom neutralen Mittelpunkt der Skala ab. Damit kann keiner der Kandidaten im Vergleich zu seinem Kontrahenten über alle Rezipienten hinweg bei einem Thema einen Vorteil erzielen. Auf Basis der einfachen Aggregationsauswertung kommen wir zu einem leicht abweichenden Urteil. Zwar unterscheiden sich die geschätzten Mittelwerte auch hier in den meisten Fällen nicht stärker vom neutralen Skalenmittelpunkt. Der einzige

wesentliche Unterschied besteht in dieser Hinsicht beim Thema *EnBW*. Die Differenzen in den Mittelwerten entstehen durch die unterschiedliche Gewichtung der einzelnen RTR-Messungen. Das kreuzklassifizierte Modell nimmt auf allen Ebenen eine Gewichtung vor. Vereinfacht dargestellt haben die einzelnen Messmodelle das Gewicht aller RTR-Messungen, auf die sie sich beziehen, und werden dann wiederum nach den Fallzahlen der L2-Einheiten Rezipienten und Turns gewichtet. Bei der Berechnung der vorhergesagten RTR-Werte wird dann die gleiche Gewichtung der drei Lager sichergestellt. Die einfache Aggregation gewichtet alle RTR-Messungen einer Person zu einem Thema gleich, eine weitere Gewichtung nach der Zahl der Turns zu einem Thema findet nicht statt. Bei der Bildung der Gesamtmittelwerte über die Personen hinweg werden dann alle drei Lager gleich gewichtet. Die Unterschiede in den Mittelwerten entstehen damit letztlich durch eine inhaltliche Entscheidung. Wir haben uns – wie eingangs des Kapitels ausführlich erläutert – dafür entschieden, die Effekte des Themas anhand der Turns zu untersuchen und damit die kumulierten Bewertungen der in sich geschlossenen Kandidatenaussagen zu einem Thema in den Fokus zu rücken. Hier wären auch andere Entscheidungen möglich gewesen, über „falsch“ oder „richtig“ kann eine inhaltliche Debatte geführt werden.

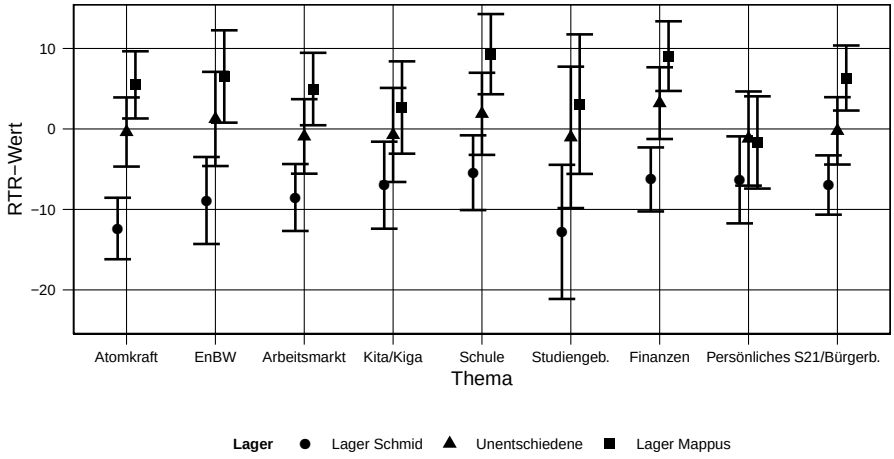
Auf einen datenanalytischen Mangel der aggregationsbasierten Vorgehensweise geht jedoch der zweite Unterschied zurück, der in Abbildung 6.23 deutlich zu sehen ist. Die Konfidenzintervalle der einfachen Mittelwertschätzung sind wesentlich enger als die der kreuzklassifizierten Schätzung. Dies liegt daran, dass beim ersten Schritt der Aggregation die Variation innerhalb der Rezipienten verloren geht. Damit wird implizit davon ausgegangen, dass ein Rezipient während eines Themenblocks kontinuierlich dieselbe RTR-Wertung abgibt, Unterschiede innerhalb und zwischen den Turns werden so vernachlässigt. In die Berechnung des Standardfehlers geht nur die Varianz zwischen den Rezipienten bei einem Thema ein (vgl. ausführlich Kapitel 5.3.2). Durch diese fälschlicherweise angenommene Präzision der Mittelwerte kommen wir zu dem Fehlschluss, dass die knappen Vorteile Schmids bei den Themen *Atomkraft*, *EnBW*, *Arbeit*, *Kiga/Kita* und *Persönliches* sowie Mappus' Vorteile bei *Schule* und *Finanzen* statistisch signifikant sind. Wie sich nun auf Basis einer der gesamten Unsicherheit in den Daten angemessenen Schätzung herausstellt, ist dieser Befund zu verwerfen. Das Thema hat, zumindest in der von uns gewählten groben Operationalisierung über die Themenblöcke, keinen signifikanten Einfluss auf die relative Bewertung der Kandidaten.

Wichtige Prädiktoren sind wiederum die Lagerzugehörigkeit ($F(2, 174) = 53.66$, $p < .001$) und die Interaktion zwischen Lagerzugehörigkeit und Thema ($F(16, 10390) = 8.59$, $p < .001$). Abbildung 6.24 stellt die Ergebnisse dar. Die

Trennung der Bewertungen nach den Lagern sticht deutlich ins Auge. Mit Ausnahme des *unpolitischen Abschnitts* sehen die Anhänger eines Lagers jeweils ihren Kandidaten im Vorteil. Nach der Bewertung der Unentschiedenen hat keiner der Kandidaten bei einem Thema einen klaren Vorteil. Zwischen den Themenblöcken zeigen sich einige Differenzen. Beim Thema *Atomkraft* liegen die Wertungen durch die beiden Lager besonders weit voneinander entfernt, was vor allem auf einen großen Vorteil Schmids im eigenen Lager zurückgeht. Einen ähnlich großen Vorteil bei seinen Anhängern kann er in den folgenden Themenblöcken nicht mehr erringen. Mappus schneidet in den Augen der Regierungsanhänger bei den Themen *Schule* und *Finanzen* am besten ab. Insgesamt fällt auf, dass über alle Themen hinweg betrachtet die Bewertungen durch die beiden Kandidatenlager leicht zum Vorteil Schmids verschoben sind. Sein Vorteil bei den eigenen Anhängern ist insgesamt etwas größer als der Vorteil von Mappus im schwarz-gelben Lager. Dies lässt sich durch die differenzielle Betrachtung der RTR-Skala in dieser Auswertung erklären. Wenn wir uns die separaten Auswertungen zu den beiden Kandidaten in Erinnerung rufen (vgl. Abbildungen 6.18, S. 235 und 6.21, S. 245), so stellen wir fest, dass Mappus von den Oppositionsanhängern bei vielen Themen relativ negativ bewertet wird, während Schmid als Sprecher von den Regierungsanhängern meist neutral beurteilt wird. Wenn diese Wertungen unbeachtet des Sprechers miteinander verrechnet werden, ergibt sich hierdurch ein Vorteil für Schmid.

Effekte des Issue Ownership und der Lagerzugehörigkeit Als nächstes wenden wir uns dem Effekt des Issue Ownership auf die Bewertung der Kandidaten zu. Entsprechend der Analysen zu den einzelnen Kandidaten untersuchen wir, ob die Kandidaten bei „ihren“ Themen einen größeren Vorteil haben, und ob dieser Effekt durch eine Zugehörigkeit zum Lager des Kandidaten verstärkt wird (Tabelle 6.17).

Bei der Interpretation der Vorzeichen der Koeffizienten muss nun die differentielle Logik der RTR-Skala bedacht werden, auf der negative Werte einen Vorteil für Schmid, positive Werte einen Vorteil für Mappus bedeuten. Einfach nachvollzogen werden kann dies anhand der Koeffizienten von *Lg. Mappus* und *Lg. Schmid*, die den Effekt der Lagerzugehörigkeit im Vergleich zu den Unentschiedenen quantifizieren, wenn der jeweilige Kandidat nicht als Issue Owner auftritt. Die Wertungen der Anhänger von Mappus weichen um 5.1 Punkte von den Unentschiedenen ab, die der Anhänger von Schmid um -7.9 Punkte. Jedes Lager sieht damit im Vergleich zur Referenzgruppe der Unentschiedenen seinen Kandidaten im Vorteil. Auch die Koeffizienten der konditionalen Effekte der Lagerzugehörigkeit unter der Bedingung des Issue Ownership des eige-



Anmerkungen
Relative Bewertung der Kandidaten auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Vorhersage durch Modell M2. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.24: Effekte des Themas und der Lagerzugehörigkeit auf die relative Kandidatenbewertung während aller Turns (M2)

nen Kandidaten sind signifikant und in ihrer Richtung erwartungskonform. Der konditionale Effekt des Issue Ownership auf Rezipienten, die nicht dem Lager des jeweiligen Kandidaten angehören, ist dagegen unbedeutend. Diese Rezipienten bewerten den Kandidaten bei „seinen" Themen nicht anders als in den übrigen Themenblöcken. Abbildung 6.25 vermittelt eine visuelle Zusammenfassung dieser Effekte.

In der differentiellen Betrachtung der relativen Vorteile der Kandidaten in Abhängigkeit vom Issue Ownership stellt sich die Wirkung dieser „eigenen" Themen damit als ein konditionaler Effekt heraus, der nur für die eigenen Anhänger in bedeutenderem Maße die Kandidatenbewertung beeinflusst. Ein Thema, bei dem die Kandidaten traditionell oder in Anbetracht aktueller Ereignisse als kompetent gelten, verschafft ihnen einen weiteren „Bonus" bei ihren eigenen Anhängern, über den Vorteil hinaus, den sie in dieser Gruppe ohnehin auch bei den übrigen Themen genießen. Dies widerspricht auf den ersten Blick den Befunden der kandidatenspezifischen Analysen, in denen

Tabelle 6.17: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die relative Kandidatenbewertung während aller Turns (M3)

	β	s.e.	t	df	p
Intercept	0.53	1.58	0.33	207	.371
Lager Mappus	5.10	1.66	3.08	170	.001
Issue Owner Mappus	1.57	2.08	0.76	63	.225
Lager Schmid	−7.89	1.48	−5.34	176	<.001
Issue Owner Schmid	−1.02	1.55	−0.66	66	.256
Lg. Mappus X Issue Owner Mappus	1.85	0.84	2.21	10405	.014
Lg. Schmid X Issue Owner Schmid	−1.53	0.55	−2.80	10417	.003

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade; p: einseitige Tests.
Die Angaben zu Varianzkomponenten und Modellpassung finden sich in Tabelle 6.16 (M3).

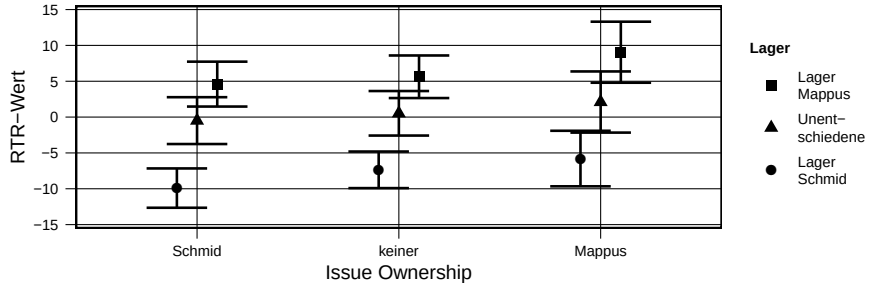
darüber hinaus ein – wenn auch schwacher – Effekt des Issue Ownership auf die Bewertung des Sprechers durch die Rezipienten, die nicht zu den Anhängern des Kandidaten gehören, identifiziert werden kann. Um diesen scheinbaren Widerspruch aufzuklären, erweitern wir das Modell M3 um den dichotomen L2-Prädiktor *Sprecher des Turns*. Dieses Modell trennt die Effekte für die beiden Sprecher und entspricht damit konzeptionell den oben präsentierten Analysen.

Die Koeffizienten in Tabelle 6.18 sind nun zusätzlich in Abhängigkeit des Sprechers zu interpretieren. Komplexer wird die Interpretation zudem durch die Präsenz der dreifachen Interaktionen *Sprecher X Lager X Issue Ownership*. Um unter der Berücksichtigung der dreifachen Interaktionen bedeutsame Tests der Effekte von Lagerzugehörigkeit, Issue Ownership und deren Interaktion für die beiden Sprecher zu erhalten, müssen zwei Spezifikationen desselben Modells betrachtet werden. Diese sind in Tabelle 6.18 dargestellt. Auf der linken Seite finden sich die Koeffizienten für das Modell, wenn die Sprechervariable mit den Ausprägungen 0 = Sprecher Schmid und 1 = Sprecher Mappus codiert ist. Die in den ersten sieben Zeilen angeordneten Koeffizienten dieses Modells beziehen sich auf die Turns, in denen Schmid spricht. Von besonderem Interesse sind hier die Effekte des Issue Ownership und ihre Interaktion mit der Lagerzugehörigkeit. Wenn Schmid im finanzpolitischen Themenblock, in dem Mappus als Issue Owner auftritt, spricht, hat dies keine negativen Konsequenzen für seine Bewertung. Weder seine eigenen Anhänger und die Unentschiedenen, noch die Anhänger von Mappus bewerten ihn hier signifikant negativer als

Tabelle 6.18: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die Bewertung beider Kandidaten während aller Turns (M4)

	Sprecher: 1 = Mappus			Sprecher: 1 = Schmid		
	β (s.e.)	t (df)	p	β (s.e.)	t (df)	p
Intercept	−2.57 (1.54)	−1.67 (226)	.048	3.92 (1.56)	2.52 (222)	.006
Lager Mappus	3.72 (1.70)	2.18 (191)	.015	6.42 (1.7)	3.78 (189)	<.001
Issue Owner Mappus	−0.38 (1.89)	−0.2 (61)	.421	3.29 (1.9)	1.73 (61)	.044
Lager Schmid	−7.67 (1.53)	−5.03 (200)	<.001	−8.20 (1.53)	−5.36 (202)	<.001
Issue Owner Schmid	−2.41 (1.48)	−1.63 (67)	.054	−1.01 (1.4)	−0.72 (67)	.237
Lg. Map. X I. O. Map.	1.48 (1.18)	1.26 (10357)	.104	2.29 (1.19)	1.92 (10441)	.027
Lg. Sch. X I. O. Sch.	−2.79 (0.79)	−3.52 (10376)	<.001	−0.27 (0.75)	−0.35 (10439)	.363
Sprecher	6.49 (1.41)	4.61 (78)	<.001	−6.49 (1.41)	−4.61 (78)	<.001
Sprecher X Lg. Map.	2.70 (0.78)	3.45 (10417)	<.001	−2.70 (0.78)	−3.45 (10417)	<.001
Sprecher X I. O. Map.	3.67 (2.68)	1.37 (61)	.088	−3.67 (2.68)	−1.37 (61)	.088
Sprecher X Lg. Sch.	−0.53 (0.78)	−0.69 (10426)	.245	0.53 (0.78)	0.69 (10426)	.245
Sprecher X I. O. Sch.	1.40 (2.04)	0.69 (67)	.246	−1.40 (2.04)	−0.69 (67)	.246
Sp. X Lg. Map. X I. O. Map.	0.81 (1.67)	0.48 (10398)	.316	−0.81 (1.67)	−0.48 (10398)	.316
Sp. X Lg. Sch. X I. O. Sch.	2.53 (1.09)	2.32 (10405)	.010	−2.53 (1.09)	−2.32 (10405)	.010

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade; p: einseitige Tests.
Die Angaben zu Varianzkomponenten und Modellpassung finden sich in Tabelle 6.16 (M4).



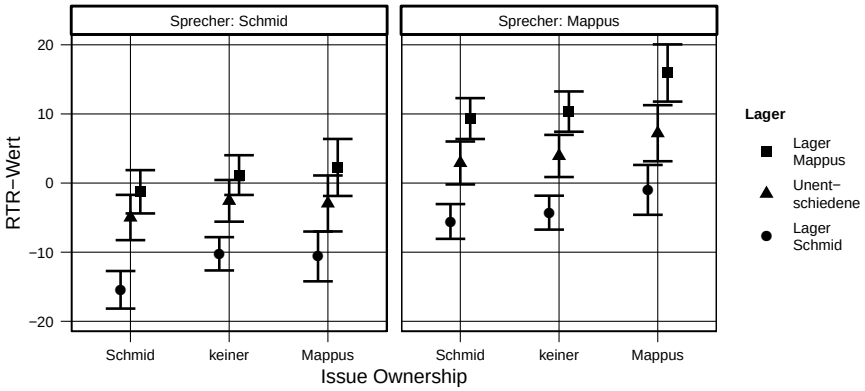
Anmerkungen
Relative Bewertung der Kandidaten auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle 6.17. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.25: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die relative Kandidatenbewertung während aller Turns (M3)

bei den Themen, bei denen keiner der Kandidaten Issue Owner ist. Mit Aussagen zu „seinen" Themen *Atomkraft*, *Arbeit* und *Kita/Kiga* erreicht er besonders große Zustimmung unter seinen eigenen Anhängern. Deskriptiv zeigen bei diesen Themen auch die Unentschiedenen und Regierungsanhänger eine etwas positivere Bewertung, die Differenz ist allerdings statistisch nicht signifikant.

Aus den ersten sieben Zeilen des rechten Teils der Tabelle können die Effekte für Mappus als Sprecher abgelesen werden. Auch hier zeigt sich, dass der Sprecher bei Themen, bei denen der Gegenkandidat als Issue Owner auftritt, nicht schlechter bewertet wird als bei den „neutralen" Themen. Dagegen erreicht Mappus mit Aussagen zu „seinen" *finanzpolitischen* Themen eine größere Zustimmung bei den Unentschiedenen und Oppositionsanhängern, und von seinen eigenen Anhänger wird er hier nochmals besser bewertet.

Die unteren sieben Zeilen geben an, wie sich die jeweiligen Effekte für die Turns von Mappus und Schmid unterscheiden. Sie sind technisch gesprochen die konditionalen Effekte in Abhängigkeit vom Sprecher des Turns. Die Äquivalenz der beiden in Tabelle 6.18 dargestellten Modelle können wir daran erkennen, dass sich die Beträge der Koeffizienten genau gleichen und sich nur das Vorzeichen ändert. Anhand des Koeffizienten der Sprechervariable lässt sich diese Logik recht einfach nachvollziehen. Die Bewertung des Kandidaten, der in der Sprechervariable die Ausprägung 0 hat, durch die Unentschiedenen



Anmerkungen
Bewertung von Schmid und Mappus auf einer Skala von −50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle 6.18. Die Fehlerbalken zeigen 95%-Konfidenzintervalle.

Abbildung 6.26: Effekte des Issue Ownership und der Lagerzugehörigkeit auf die Bewertung beider Kandidaten während aller Turns (M4)

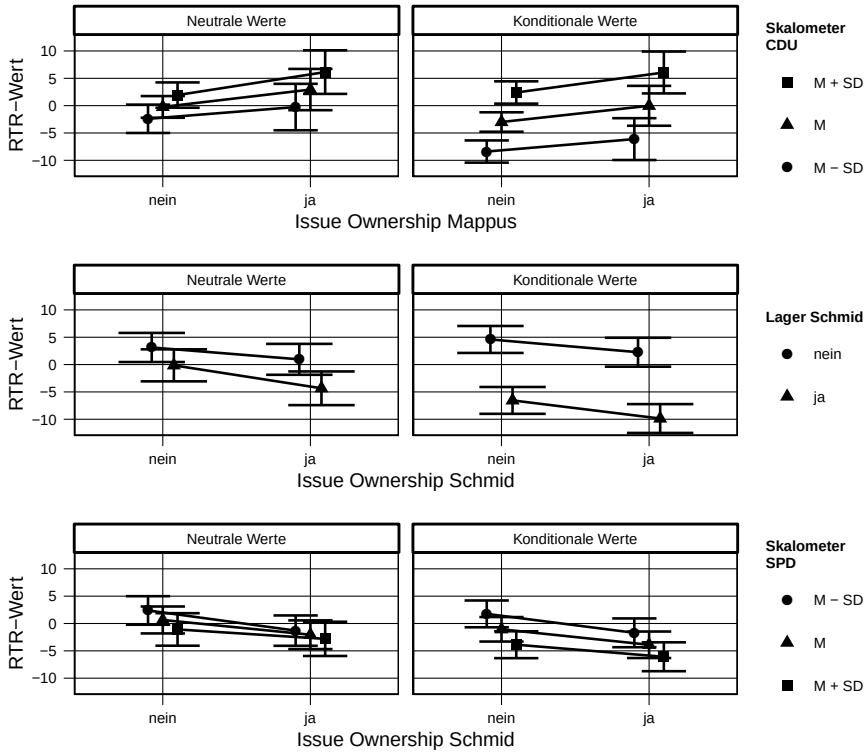
während der Themenblöcke, in denen keiner der Kandidaten als Issue Owner auftritt, findet sich im Koeffizienten des Intercept. Im linken Modell bewerten sie hier Schmid mit einem Koeffizienten von −2.6 leicht positiv. Der Koeffizient von *Sprecher* gibt hier an, dass die Bewertung von Mappus bei diesen Themen durch diese Gruppe um 6.5 Skalenpunkte von der Bewertung Schmids abweicht und damit in Verrechnung der Koeffizienten leicht auf der Seite der Skala liegt, die eine positive Bewertung für Mappus ausdrückt. Hier wird deutlich, dass alleine das Sprechen einem Kandidaten in Abwesenheit anderer Einflüsse (keine Lagerzugehörigkeit, kein Issue Ownership) eine positivere Bewertung einbringt (vgl. zu diesem Befund zuerst Faas & Maier, 2004a; Maurer & Reinemann, 2003). Diese Verschiebung der Bewertungen in Richtung des Sprechers verursacht auch die etwas größere Varianz, die in den die Kandidaten vergleichenden Modellen auf die Eigenschaften der Turns zurückgeht. Durch die Aufnahme des Prädiktors *Sprecher* wird die turnbezogene Varianzkomponente um 62 Prozent reduziert.

Neben dieser zu erwartenden Verschiebung der Bewertungen in Richtung des Kandidaten, der gerade das Wort hat, findet sich nur eine weitere signifi-

kante Interaktion mit der Sprechervariable. Der konditionale Effekt des Issue Ownership Schmids in Abhängigkeit der Zugehörigkeit zum Lager Schmid wird wiederum durch den Sprecher des Turns moderiert. Wenn Schmid zu einem „seiner" Themen spricht, wird er von seinen Anhänger um 2.8 Skalenpunkte besser bewertet, als es alleine durch den Effekt des Issue Ownership und der Zugehörigkeit zum Lager Schmid zu erwarten wäre. Diese Wirkung wird jedoch nur aufrecht erhalten, solange Schmid spricht. Wenn Mappus sich zu denselben Themen äußert, wird dieser zusätzliche „Bonus" fast komplett durch den entgegengesetzten konditionalen Interaktionseffekt von $\beta = 2.53$ ausgeglichen. Der verbleibende Vorsprung von Schmid in seinem Lager entspricht damit nur noch dem einfachen Effekt der Lagerzugehörigkeit.

Insgesamt sprechen die Befunde der Modelle dafür, dass der Effekt des Issue Ownership auf die unmittelbare Bewertung der Kandidaten während des Duells komplexer ist, als es eine einfache Unterscheidung nach „Gewinner- und Verliererthemen" erscheinen lässt. In erster Hinsicht zeigt sich der Effekt des Issue Ownership als ein konditionaler Effekt, der bei den eigenen Anhängern besonders deutlich in Erscheinung tritt. Dies zeigt sich sowohl in den separaten Analysen der Bewertungen von Schmid und Mappus als auch in den vergleichenden Analysen beider Kandidaten ohne (M3) und mit (M4) Berücksichtigung des Sprechers. Demzufolge eignet sich eine Auseinandersetzung zu den „eigenen" Themen vor allem dazu, die eigenen Anhänger zu mobilisieren. Leichte Hinweise liegen zudem dafür vor, dass während der Passagen, in denen ein Kandidat selbst zu einem „eigenen" Thema spricht, auch die übrigen Rezipienten einen etwas besseren Eindruck von ihm haben. Dieser Effekt ist allerdings nicht stark genug, um im Vergleich mit dem Gegenkandidaten über den gesamten Themenblock hinweg einen Vorteil zu erzielen.

Effekte des Issue Ownership und der Voreinstellungen Abschließend soll nun noch eine Erweiterung des Issue-Owner-Modells um weitere Rezipientenmerkmale vorgenommen werden. Um die Darstellungen überschaubar zu halten, beschränken wir uns auf eine Weiterentwicklung des Modells M3, in dem die Sprechervariable nicht enthalten ist. Aus den bereits bekannten Rezipientenmerkmalen können die Voreinstellungen gegenüber den beiden Kandidaten und ihren Parteien sowie die Zugehörigkeit zum Lager Schmid als bedeutsame L2-Prädiktoren identifiziert werden. Mit ihnen erhöht sich die Varianzaufklärung in der rezipientenbezogenen Komponente gegenüber dem einfachen Modell M3 der Lagerzugehörigkeit um 25 Prozentpunkte auf 54 Prozent. Tabelle 6.19 und Abbildung 6.27 fassen die Befunde zusammen.



Anmerkungen
Relative Bewertung der Kandidaten auf einer Skala von -50 (größter Vorteil Schmid) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle 6.19. Die Fehlerbalken zeigen 95%-Konfidenzintervalle. Vorhergesagte RTR-Werte beim Mittelwert (M) und ± 1 Standardabweichung (SD) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf einen neutralen Wert gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt.

Abbildung 6.27: Effekte des Issue Ownership und der Voreinstellungen auf die relative Kandidatenbewertung während aller Turns (M5)

Tabelle 6.19: Effekte des Issue Ownership und der Voreinstellungen auf die relative Kandidatenbewertung während aller Turns (M5)

	β	s.e.	t	df	p
Intercept	2.18	1.35	1.62	155	.107
Skalometer Mappus	1.11	0.25	4.49	167	<.001
Skalometer CDU	0.74	0.25	2.93	167	.004
Lager Schmid	−3.27	1.38	−2.37	172	.019
Skalometer Schmid	−0.86	0.38	−2.25	185	.026
Skalometer SPD	−0.79	0.31	−2.57	185	.011
Issue Owner Mappus	1.90	2.08	0.91	64	.366
Issue Owner Schmid	−1.22	1.55	−0.79	66	.432
I. O. Map. X Sk. Map.	−0.01	0.17	−0.04	10400	.968
I. O. Map. X Sk. CDU	0.35	0.16	2.15	10403	.032
I. O. Sch. X Lg. Sch.	−2.02	0.57	−3.57	10414	<.001
I. O. Sch. X Sk. Sch.	−0.04	0.20	−0.21	10425	.834
I. O. Sch. X Sk. SPD	0.45	0.16	2.82	10425	.005

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.
Die Angaben zu Varianzkomponenten und Modellpassung finden sich in Tabelle 6.16 (M5).

Als einfache Effekte ohne eine signifikante Interaktion mit dem Issue Ownership hängen die Voreinstellungen zu den Kandidaten in der erwarteten Richtung mit der differentiellen Bewertung der Kandidaten zusammen. Je besser (schlechter) Mappus vor dem Duell bewertet wird, desto stärker wird das Duell zum Vorteil von Mappus (Schmid) bewertet. Der Zusammenhang zwischen der Voreinstellung zu Schmid und der relativen Bewertung der Kandidaten während des Duells ist entsprechend der Logik der RTR-Skala negativ.

Die Voreinstellung zur CDU in Interaktion mit dem Issue Ownership von Mappus sowie die Zugehörigkeit zum Lager Schmid in Interaktion mit dem Issue Ownership von Schmid zeigen den bereits bekannten Verstärkereffekt: Rezipienten, die einem Kandidaten näherstehen, sehen ihn noch stärker im Vorteil, wenn „dessen“ Thema debattiert wird. Die Voreinstellung zur SPD wirkt gegen den Verstärkereffekt der Zugehörigkeit zum Oppositionslager. Auf den ersten Blick kontraintuitiv ist der Effekt der Voreinstellung zur SPD nur bei Themen wirksam, bei denen Schmid *nicht* als Issue Owner auftritt. Wie oben bereits ausführlicher erläutert, lässt sich dieser Befund als partieller

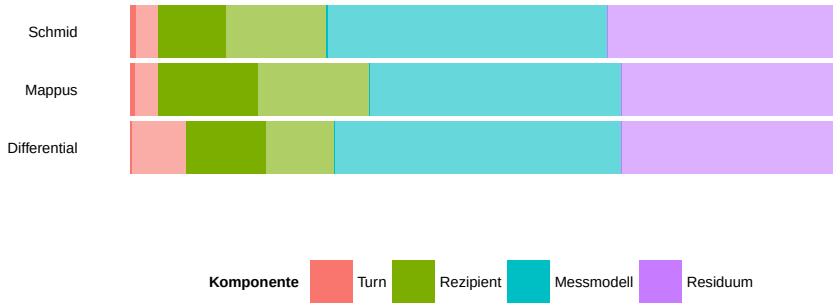
Effekt in Anwesenheit des wesentlich stärkeren Einflusses der Lagerzugehörigkeit plausibilisieren. Über die Wirkung der Zugehörigkeit zum Lager des Oppositionskandidaten und dessen Verstärkung durch das Issue Ownership enthält das Modell bereits einen wichtigen Mechanismus, der einen Vorteil von Schmid in den unmittelbaren Bewertungen erklärt. Ein zusätzlicher Vorteil durch eine überdurchschnittlich positive Einstellung zu seiner Partei macht sich nur noch dann bemerkbar, wenn nicht auch noch der Verstärkereffekt des Issue Ownership wirkt. Insgesamt sollte bei allen diesen Interpretationen berücksichtigt werden, dass die Interaktionseffekte – wenngleich statistisch signifikant – im Vergleich zu den einfachen Effekten der Voreinstellungen nur wenig ins Gewicht fallen. Dies wird vor allem bei der Betrachtung der vorhergesagten RTR-Werte auf Basis der Stichprobenverteilung aller L2-Prädiktoren (Konditionale Werte in Abbildung 6.27) deutlich. In Abhängigkeit vom Issue Ownership verändern sich diese starken Effekte in Maßen, der wesentliche Einfluss geht jedoch von den Voreinstellungen aus.

6.2.3 Zwischenfazit zu den kreuzklassifizierten Modellen

Aus datenanalytischer Sicht ermöglicht es das kreuzklassifizierte Modell, die Datenstruktur der RTR-Messungen (die Kandidaten werden während eines Turns von vielen Rezipienten bewertet; ein Rezipient bewertet die Kandidaten in vielen Turns) angemessen abzubilden. Knapp zusammengefasst bringt dies drei zentrale Nutzen für die Analyse der unmittelbaren Kandidatenbewertungen während des TV-Duells.

Erstens erlaubt die Modellierung der Datenstruktur eine Varianzdekomposition und damit die relative Quantifizierung der Wichtigkeit von Merkmalen der Rezipienten und des Inhalts für die Erklärung der Echtzeitbewertungen. Abbildung 6.28 fasst die Befunde der Varianzdekomposition der Modelle zur Erklärung der Bewertung von Schmid und Mappus sowie zur vergleichenden Bewertung zusammen. Durch die Farben sind die Varianzkomponenten gekennzeichnet, die nach der Intra-Klassen-Korrelation auf Turns, Rezipienten, Turn-Rezipienten-Kombinationen (Messmodelle) und modellbedingte Messfehler (L1-Residuen) entfallen. Dunkler schattiert sind dann jeweils die Anteile, die durch die L2-Prädiktoren Issue Ownership und Voreinstellungen sowie ihre Interaktionen erklärt werden können. Heller dargestellt sind die Varianzanteile, die in diesen Modellen als Fehlervarianz verbleiben.

Bereits aus diesen Ergebnissen können wir wichtige Erkenntnisse ableiten. So kommt den Eigenschaften der Rezipienten – und damit den Voreinstellungen, Vorwissen oder Persönlichkeitseigenschaften – eine deutlich wichtigere Rolle bei der Erklärung der unmittelbaren Kandidatenbewertungen während des



Anmerkungen
Ergebnisse der Varianzdekomposition der unkonditionalen Modelle sowie Varianzreduzierung durch die Modelle mit den L2-Prädiktoren Issue Ownership und Voreinstellungen. Die Farben kennzeichnen die Zuordnung der Varianzkomponenten zu den Ebenen. Dunkel schattiert: erklärte Varianz; hell schattiert: nicht erklärte Varianz.

Abbildung 6.28: Varianzdekomposition und Varianzerklärung im Vergleich

Duells zu als den Eigenschaften der Turns. Wenn es unser Ziel ist, einfache Erklärungen für die RTR-Bewertungen zu finden, müssen wir vor allem bei einer Modellierung der Rezipienteneigenschaften ansetzen. Für alle Rezipienten gleichgerichtete Befunde zur Wirkung bestimmter Debatteninhalte dürfen wir in diesem Modell kaum erwarten. Selbst wenn wir alle Merkmale der Turns in einer Inhaltsanalyse perfekt erfassen, könnten wir damit lediglich vier Prozent aller Varianz (bzw. fünf Prozent der erklärbaren Varianz) erklären. Zumindest bei einer zeitlichen Untergliederung der Debatte in Turns ist eine detaillierte Erfassung der rezipientenseitigen Merkmale damit erfolgsversprechender als eine aufwändige, detaillierte Inhaltsanalyse der Turns.

In allen Modellen entfällt die weitaus größte Varianzkomponente auf die Varianz zwischen den in den Messmodellen erfassten RTR-Bewertungen, die sich nicht durch einfache „Haupteffekte“ von Rezipienten- oder Turnmerkmalen, sondern nur durch deren spezifische Interaktionen erklären lassen.[74] Diese Feststellung ist kaum überraschend. Sie bedeutet nichts anderes, als dass individuelle Rezipienten unterschiedliche Debatteninhalte anders bewerten. Obgleich diese Erkenntnis spätestens mit der Erweiterung des einfachen

74 Hinzu kommen, wie oben ausführlich erläutert, zufällige und situationsbedingte Messfehler (vgl. S. 222ff).

Stimulus-Response-Modells zum S-O-R-Modell in den Yale-Studien der 1950er Jahren zu den Grundlagen der Medienwirkungsforschung zählt (Schenk, 2007, S. 77-137), findet sie in den verbreiteten Verfahren zur Analyse der unmittelbaren Kandidatenbewertung während der TV-Duelle keine Anwendung. Weder die Aggregation über die Rezipienten noch die Aggregation über Messzeitpunkte erlaubt es, solche Interaktionen zu modellieren (vgl. Kapitel 5.3). Besonders Analysen, die die Urteile aller Rezipienten unabhängig von ihrer Voreinstellung zusammenfassen und damit interindividuell unterschiedliche Reaktionen auf Debatteninhalte überhaupt nicht untersuchen (können), beschränken sich in dem, was sie erklären können, auf einen verschwindend geringen Teil der in den Daten enthaltenen Variation.

Einschränkend müssen wir festhalten, dass unsere Entscheidung, die Wirkung der Debatteninhalte an den recht weit gefassten Turns festzumachen, die inhaltsbezogene Varianz etwas unterschätzt. Doch selbst wenn wir annehmen, dass die relevanten Merkmale des Stimulus sekundengenau erfasst und den passenden Stellen in der Publikumsreaktion perfekt zugeordnet werden können, ändert sich der Befund nicht wesentlich. In Modellen, in denen die Inhaltswirkungen auf Sekundenebene eintreten können, liegen die Anteile der inhaltsbezogenen Varianzkomponenten bei sieben (Bewertungen von Mappus), acht (Bewertungen von Schmid) bzw. zwölf (differentielle Bewertungen der Kandidaten) Prozent. Die rezipientenbezogene Komponente macht dagegen immer noch 32, 27 bzw. 22 Prozent der gesamten Varianz aus. Die verbleibenden Anteile liegen jeweils in den spezifischen Kombinationen aus Rezipienten- und sekündlichen Inhaltsmerkmalen. Die Ergebnisse der Varianzdekomposition beziehen sich natürlich nur auf unseren Datensatz zur Bewertung von Schmid und Mappus während des TV-Duells vor der Landtagswahl 2011 in Baden-Württemberg. In Anbetracht der sehr klaren Befunde erscheinen aber Analysen, die mit sehr großem Aufwand sehr genaue Inhaltsanalysen der Debatte durchführen, um dann die über Personen aggregierten RTR-Messungen zu erklären, auch im Allgemeinen wenig effizient. Besonders wenn sich die Analysen auf die für das gesamte Publikum unbeachtet der Lagerzugehörigkeit oder anderer bedeutender Merkmale der politischen Voreinstellungen zusammengefassten RTR-Kurven beziehen, können sie nur einen sehr kleinen Anteil der zuvor aufwändig erhobenen unmittelbaren Kandidatenbewertungen erklären, da nicht einmal durch vergleichende Betrachtungen indirekt auf die interaktionsbezogene Komponente geschlossen werden kann. Allgemein können die Befunde der Varianzdekomposition als ein weiterer Hinweis verstanden werden, auf Analysen ohne Berücksichtigung zumindest einiger weniger wichtiger Rezipientenmerkmale wie beispielsweise der Lagerzugehörigkeit zu verzichten –

sei es in Form von deskriptiven Peak-Spike-Analysen, aggregationsbasierter Verfahren oder komplexeren Methoden.

Zweitens können wir exemplarisch zeigen, dass die Einbeziehung aller Quellen statistischer Unsicherheit zu angemesseneren Schätzungen der Effekte führt. So kommt eine Analyse des relativen Vorteils der Kandidaten in den Themenblöcken der Debatte auf Basis einer einfachen Zusammenfassung aller RTR-Messungen eines Rezipienten während der jeweiligen Blöcke fälschlicherweise zu dem Schluss, dass beide Kandidaten in einigen Themenbereichen im Vorteil sind (Bachl, 2013a). Dieser Fehlschluss kam zustande, da wir die Variation zwischen den Bewertungen eines Rezipienten durch die Aggregation über die Messzeitpunkte hinweg vernachlässigt haben. Nach den Ergebnissen des kreuzklassifizierten Modells, das sowohl die Variation zwischen als auch innerhalb der Rezipienten berücksichtigt, müssen wir diesen Befund korrigieren. Auch hier müssen wir einschränkend bemerken, dass ein Teil der sehr großen Unsicherheit um die mittlere Bewertung der Themen auf die Wahl der Turns als Analyseeinheit für den Debatteninhalt zurückgeht. Wenn wir annehmen, dass der Effekt eines Themas „häufiger" auftritt, d.h. sich die Bewertung eines Kandidaten z.B. auch durch einen Themenwechsel während des Turns ändern kann, so würden bei einem konsistenten Effekt präzisere Schätzungen möglich. Solche kleineren Analyseeinheiten für die Ermittlung des Themas erfordern jedoch eine theoretische Vorstellung davon, wann sich das Thema einer Aussage derart verändert, dass dies bei den Rezipienten ankommt und einen Effekt auslösen kann. Nur wenn diese Definition in der Inhaltsanalyse gelingt, würden tatsächlich präzisere Schätzungen der Themeneffekte (oder auch der Effekte anderer inhaltlicher Merkmale) erreicht.

Drittens ermöglicht das kreuzklassifizierte Modell die simultane Berücksichtigung von Eigenschaften der Rezipienten und Turns und damit das Überprüfen von Annahmen über den Effekt dieser Merkmale und ihrer Interaktionen. Auch wenn die Berechnung der entsprechenden Tests, wie in mehreren Anmerkungen festgehalten, nicht immer statistisch trivial ist, so stellt dies alleine einen großen Fortschritt gegenüber den bislang verwendeten Analysen deduktiver Studien zur Wirkung bestimmter Inhalts- und/oder Rezipientenmerkmale dar. Zudem hilft die Möglichkeit, im Gegensatz zu Aggregationen über die Personen auch mehrere und (quasi-) metrisch skalierte Prädiktoren in einem Modell zu berücksichtigen, gerade bei der Erklärung der rezipientenbezogenen Varianzkomponente weiter. Entgegen steht der Erweiterung der Modelle in dieser Hinsicht nur die mit der Ordnung der Interaktionen ansteigende Komplexität der Interpretationen. Gerade das Modell, das neben den Cross-Level-Interaktionen zusätzlich Interaktionen mit der Sprecher-Variable enthält, zeigt, dass dreifache Interaktionen selbst dann nur noch schwer zu interpre-

tieren sind, wenn sie lediglich aus dichotomen Variablen bestehen. Sollten jedoch Interaktionen höherer Ordnung und/oder mit (quasi-) metrischen Variablen theoretisch von Interesse sein, so können sie ohne weiteres mit dieser Modellklasse getestet werden.

Aus inhaltlicher Perspektive sind hier – neben der Bestätigung der großen Bedeutung der Voreinstellungen für die Bewertung der Kandidaten im Duell – die Erkenntnisse zum Effekt des Issue Ownership herauszuheben. Es zeigt sich, dass die Wirkung des Issue Ownership vor allem als konditionaler Effekt bei den eigenen Anhängern auftritt. In den Auseinandersetzungen zu Atomkraft, Arbeit und Kita/Kiga hat Schmid bei den Oppositionsanhängern einen besonders großen Vorteil. Die Anhänger der Regierungsparteien sehen Mappus beim Thema Finanzen besonders weit vor Schmid. Einen leichten positiven Effekt hat das Issue Ownership auf die Bewertung der Kandidaten bei den Unentschiedenen und den Anhängern des Gegenkandidaten – allerdings nur, während der Kandidat selbst zu „seinen" Themen spricht. Da dieser Effekt recht schwach ist und kein „Malus" für Aussagen eines Kandidaten zu den Themen, bei denen der andere Kandidat als Issue Owner auftritt, festgestellt werden kann, reicht der Effekt nicht aus, um bei den Rezipienten, die nicht dem eigenen Lager angehören, im Themenblock insgesamt einen Vorteil zu erzielen. Während die Annahmen des Issue-Ownership-Ansatzes übertragen auf das TV-Duell insgesamt von den Daten gestützt werden können, muss auch festgehalten werden, dass die identifizierten Effekte im Vergleich zum Einfluss der Voreinstellungen sehr begrenzt sind.

6.3 Das kreuzklassifizierte Wachstumskurvenmodell

Im letzten Teilkapitel zu Mehrebenenmodellen der unmittelbaren Kandidatenbewertungen kombinieren wir die Modellklassen der Wachstumskurven- und kreuzklassifizierten Modelle zur Klasse der kreuzklassifizierten Wachstumskurvenmodelle. In Kapitel 6.1 haben wir gezeigt, dass Wachstumskurvenmodelle dazu geeignet sind, die dynamischen Veränderungen der unmittelbaren Kandidatenbewertungen innerhalb einer Antwort zu erfassen. Mit einer Spezifikation des Wachstumskurvenmodells durch einen latenten Intercept und einen latenten linearen Slope erhalten wir zwei Parameter, die uns Auskunft über schnelle, heftige Reaktionen zu Beginn einer Antwort sowie Veränderungen über den weiteren Verlauf einer Antwort geben. Neben einer verbesserten Datenpassung haben diese Modelle den großen Vorteil, dass wir detaillierte theoretische Annahmen über die dynamische Natur der unmittelbaren Kandidatenbewertungen empirisch prüfen können.

Die einfachen Wachstumskurvenmodelle sind jedoch zunächst auf die Analyse der Reaktionen auf eine einzelne Antwort beschränkt. Um dem Umstand Rechnung zu tragen, dass in einem TV-Duell unmittelbare Kandidatenbewertungen von vielen individuellen Rezipienten zu vielen unterschiedlichen Kandidatenaussagen vorliegen, haben wir in Kapitel 6.2 das kreuzklassifizierte Modell eingeführt. Diese Modellklasse erlaubt es, die in der Forschung zu TV-Duellen häufig angestrebten Inferenzschlüsse in Richtung einer Grundgesamtheit von potenziellen Rezipienten *und* einer Grundgesamtheit von potenziellen (Debatten-) Inhalten korrekt durchzuführen. Dabei werden Effekte von Merkmalen der Rezipienten, der Debatteninhalte und ihrer Interaktionen zur Erklärung der unmittelbaren Bewertungen herangezogen. So wird es möglich, Annahmen über das Zusammenwirken von Voreinstellungen und Debatteninhalten systematisch für alle Rezipienten und alle Kandidatenaussagen (in diesem Beispiel als Turns operationalisiert) empirisch zu überprüfen. Die Dekomposition der Varianz hilft uns zudem dabei, die allgemeine Bedeutung von Eigenschaften der Rezipienten und Kandidatenaussagen für die unmittelbaren Kandidatenbewertungen einzuschätzen und somit geeignete Ansatzpunkte für weitere Modellverbesserungen und zukünftige Forschungsvorhaben zu finden.

Die einfachen kreuzklassifizierten Modelle vernachlässigen jedoch die dynamische Veränderung der unmittelbaren Kandidatenbewertungen innerhalb der Kandidatenaussagen. Diese Einschränkung ist für einige Forschungsfragen zu verkraften. Wenn wir beispielsweise wie in Kapitel 6.2 an der summarischen Bewertung der Positionen der Kandidaten zu bestimmten Themen interessiert sind, so ist die mittlere Bewertung während der Kandidatenaussagen zu einem Thema eine angemessene Operationalisierung. Wenn jedoch die dynamische Veränderung der unmittelbaren Kandidatenbewertungen infolge eines Merkmals des Debatteninhalts von Interesse ist, greift das einfache kreuzklassifizierte Modell zu kurz. Im Folgenden wollen wir untersuchen, ob Rezipienten auf Angriffe, Verteidigungen oder Selbstpräsentationen systematisch über alle Antworten der Kandidaten hinweg unterschiedlich reagieren – hier steht die Veränderung der Kandidatenbewertungen im Fokus. Für dieses Erkenntnisinteresse müssen die Klassen der Wachstumskurvenmodelle und der kreuzklassifizierten Modelle kombiniert werden. Es ist dann möglich, Annahmen über die Dynamiken innerhalb der einzelnen Antworten systematisch in Abhängigkeit von Eigenschaften der Rezipienten, der Antworten und deren Interaktion zu überprüfen. Wie im einfachen kreuzklassifizierten Modell wird dabei die statistische Unsicherheit auf beiden Ebenen berücksichtigt, um valide Inferenzen auf andere Rezipienten und andere Antworten zu ziehen. Ebenso bleibt die Möglichkeit zur Varianzdekomposition erhalten und wird um

die Unterscheidung zwischen den Parametern des Wachstumskurvenmodells erweitert.

Im Folgenden erläutern wir zuerst die Grundlagen der Modellklasse. Dann wenden wir sie in zwei exemplarischen Analysen zur Beantwortung der Frage an, ob sich die Bewertungen der Kandidaten in Abhängigkeit von Voreinstellungen der Rezipienten und Relationen der Antworten unterschiedlich verändern. Dabei wenden wir uns zuerst der Bewertung der Kandidaten während ihrer direkten Antworten auf Fragen der Moderatoren zu. Danach untersuchen wir, ob mit einem Wechsel der Relation auch eine Veränderung der unmittelbaren Bewertungen eintritt.

6.3.1 Grundlagen der Modellklasse

Auf Basis der ausführlichen Darstellungen des Wachstumskurvenmodells (Kapitel 6.1) und des kreuzklassifizierten Modells (Kapitel 6.2) kann die für sich genommen komplexe Modellklasse der kreuzklassifizierten Wachstumskurvenmodelle zumindest konzeptionell recht einfach nachvollzogen werden. Wir erinnern uns, dass im einfachen Wachstumskurvenmodell die individuellen Verläufe der RTR-Messungen auf L1 durch ein Messmodell aus einer oder mehreren latenten Variablen abgebildet wird, die auf L2 jeweils in einem Rezipienten gruppiert sind. Dabei beschränkt sich das Modell auf die Analyse der Bewertungen aller Rezipienten zu einer Antwort. Im einfachen kreuzklassifizierten Modell wird als L1-Messmodell für jede Rezipienten-Turn-Kombination ein Intercept-Only-Modell geschätzt, das den individuellen Verlauf in dieser Kombination durch einen einfachen Mittelwert erfasst. Diese L1-Messmodelle sind auf L2 jeweils genau einem Rezipienten und einem Turn zugeordnet. Hier können die Bewertungen während aller Turns berücksichtigt werden, allerdings wird die dynamische Veränderung der Bewertungen innerhalb der Turns vernachlässigt. Kombinieren wir nun diese beiden Modelle, so erhalten wir die komplexeren, die Dynamik erfassenden L1-Messmodelle der Wachstumskurvenmodelle, die wie im kreuzklassifizierten Modell auf L2 in Rezipienten und Turns gruppiert sind. Dadurch ist es möglich, die dynamische Veränderung der Bewertungen während aller Antworten[75] in Abhängigkeit von Merkmalen der Rezipienten und der Antworten zu analysieren.

75 Die im zweiten Anwendungsbeispiel als Analyseeinheiten des Debatteninhalts genutzten Relationswechsel sind in ihrer inhaltlichen Logik etwas komplexer, da sie keinen formal definierten Startpunkt besitzen. Daher beschränken wir uns in diesen einführenden Erklärungen auf die Einheit der Antworten. Die statistische Analyselogik der Relationswechsel ist äquivalent und wird anhand des Beispiels näher erläutert.

Die formale Repräsentation des kreuzklassifizierten Wachstumskurvenmodells ergibt sich, indem wir die L1-Spezifikation eines Wachstumskurvenmodells in das Messmodell des kreuzklassifizierten Modells einsetzen. Wir nutzen in den folgenden Analysen das einfache Wachstumskurvenmodell mit einem latenten Intercept und einem latenten linearen Slope, das sich bereits in den Analysen in Kapitel 6.1 bewährt hat. Anhand der exemplarischen Analysen werden wir weiter zeigen, dass diese Spezifikation ein angemessenes Verhältnis von inhaltlicher Interpretierbarkeit, Sparsamkeit und Datenpassung aufweist. Dieses Wachstumskurvenmodell ist formal definiert als (reproduziert von S. 180):

$$rtr_{ti} = \pi_{0i} + \pi_{1i} \times zeit_{ti} + e_{ti} \tag{6.5}$$

Das Messmodell im einfachen kreuzklassifizierten Modell, statistisch betrachtet auf zwei Ebenen angesiedelt, ist definiert als (reproduziert von S. 221):

$$rtr_{th(ij)} = \pi_{0h(ij)} + e_{th(ij)} \tag{6.12}$$

$$\pi_{0h(ij)} = \pi_{0(ij)} + f_{h(ij)} \tag{6.13}$$

Wenn wir im Messmodell des kreuzklassifizierten Modells Gleichung 6.12 durch die L1-Spezifikation des Wachstumskurvenmodells in Gleichung 6.5 ersetzen, erhalten wir für das Messmodell des kreuzklassifizierten Wachstumskurvenmodells mit einer L1-Spezifikation aus Intercept und linearem Slope:

$$rtr_{th(ij)} = \pi_{0h(ij)} + \pi_{1h(ij)} \times zeit_{th(ij)} + e_{th(ij)} \tag{6.16}$$

$$\pi_{0h(ij)} = \pi_{0(ij)} + f_{0h(ij)} \tag{6.17}$$

$$\pi_{1h(ij)} = \pi_{1(ij)} + f_{1h(ij)} \tag{6.18}$$

Da die einzelnen RTR-Messungen nach dieser L1-Spezifikation durch zwei L1-Parameter, den Intercept $\pi_{0h(ij)}$ und den Slope $\pi_{1h(ij)}$, beschrieben werden, folgen im zweiten Teil des Messmodells ebenfalls zwei Gleichungen 6.17 und 6.18, in denen wiederum diese Parameter beschrieben werden. Dieses Messmodell kann nun in die kreuzklassifizierte L2-Struktur eingesetzt werden. Diese ist ebenfalls in zwei Gleichungen definiert als (vgl. zur Erläuterung auch die L2-Spezifikation des einfachen kreuzklassifizierten Modells auf S. 222):

$$\pi_{0(ij)} = \beta_{00} + u_{0i} + v_{0j} \tag{6.19}$$

$$\pi_{0(ij)} = \beta_{10} + u_{1i} + v_{1j} \tag{6.20}$$

Durch Ersetzung erhalten wir so für die Beschreibung der RTR-Messungen durch das unkonditionale kreuzklassifizierte Wachstumskurvenmodell mit Intercept und linearem Slope die Gleichung:

$$\begin{aligned} rtr_{th(ij)} &= \beta_{00} + \beta_{10} \times zeit_{th(ij)} \\ &\quad + u_{0i} + v_{0j} + f_{0h(ij)} \\ &\quad + u_{1i} \times zeit_{th(ij)} + v_{1j} \times zeit_{th(ij)} + f_{1h(ij)} \times zeit_{th(ij)} \\ &\quad + e_{th(ij)} \end{aligned} \tag{6.21}$$

In der ersten Zeile wird die RTR-Messung rtr zum Zeitpunk t im RTR-Verlauf h durch den Rezipienten i zur Antwort j durch die Schätzer der Populationskoeffizienten (Fixed Effects) beschrieben. Dabei quantifiziert der Intercept-Koeffizient β_{00} im unkonditionalen Modell ohne weitere L2-Prädiktoren den geschätzten durchschnittlichen RTR-Wert aller Rezipienten in allen Antworten zu Beginn der Antworten. Der Koeffizient β_{10} gibt die durchschnittliche Steigung des RTR-Verlaufs aller Rezipienten und Antworten in einer Sekunde an. In der zweiten Zeile finden sich die rezipienten-, antwort- und messmodellspezifischen Abweichungen (Random Effects) vom durchschnittlichen Intercept β_{00}. Die beobachtete RTR-Messung wird demnach weiter angenähert durch die Abweichung u_{0i} des Rezipienten i von allen Rezipienten, die Abweichung v_{0j} der Antwort j von allen Antworten und die Abweichung $f_{0h(ij)}$ des Messmodells h von allen Messmodellen. Nach demselben Prinzip beziehen sich die Abweichungen u_{1i}, v_{1j} und $f_{1h(ij)}$ auf die Abweichungen vom Slope-Koeffizienten β_{10}. Entsprechend der Logik des kreuzklassifizierten Modells streuen die Fehlerterme u zwischen den Rezipienten, die Fehlerterme v zwischen den Antworten und die Fehlerterme f zwischen den Messmodellen innerhalb der Rezipienten und Antworten. Der Fehlerterm $e_{th(ij)}$ beschreibt schließlich die modellbedingte Abweichung der einzelnen RTR-Messung vom durch die übrigen Bestandteile der Gleichung geschätzten RTR-Verlauf. Da die L2-Prädiktoren nur auf die durch die L1-Messmodelle geschätzten RTR-Verläufe Einfluss nehmen, kann dieser letzte Fehlerterm in den hier präsentierten Modellen nicht erklärt werden.

Es wird deutlich, dass die grundsätzliche Logik der kreuzklassifizierten Modelle auch für die kreuzklassifizierten Wachstumskurvenmodelle gilt. Wir interessieren uns dafür, wie sehr die RTR-Verläufe der Rezipienten und Antworten im unkonditionalen Modell um die durchschnittlichen Verläufe streuen. Dann versuchen wir, diese Streuung mit der Erweiterung des Modells um Rezipienten- und Antwortmerkmale sowie ihre Interaktionen zu erklären. Die Untersuchung und Interpretation der Koeffizienten für diese Merkmale ist vergleichsweise einfach, da sie nur die erste Zeile der Regressionsgleichung 6.21 erweitert. Im Folgenden werden wir zuerst die RTR-Bewertungen der Kandidaten in Abhängigkeit der Lagerzugehörigkeit der Rezipienten (Lager Mappus: lg_map, Lager Schmid: lg_sch) und der Relation der Antwort (Angriff: rel_angr, Verteidigung: rel_vert) analysieren. Als Referenzausprägungen werden die Unentschiedenen bzw. die Selbstpräsentationen genutzt. Dieses Modell ist formal definiert als:

$$
\begin{aligned}
rtr_{th(ij)} = {} & \beta_{00} \\
& + \beta_{01} \times lg\_map_i + \beta_{02} \times lg\_sch_i \\
& + \beta_{03} \times rel\_angr_j + \beta_{04} \times rel\_vert_j \\
& + \beta_{05} \times lg\_map_i \times rel\_angr_j + \beta_{06} \times lg\_sch_i \times rel\_angr_j \\
& + \beta_{07} \times lg\_map_i \times rel\_vert_j + \beta_{08} \times lg\_sch_i \times rel\_vert_j \\
& + \beta_{10} \times zeit_{th(ij)} \\
& + \beta_{11} \times lg\_map_i \times zeit_{th(ij)} + \beta_{12} \times lg\_sch_i \times zeit_{th(ij)} \\
& + \beta_{13} \times rel\_angr_j \times zeit_{th(ij)} + \beta_{14} \times rel\_vert_j \times zeit_{th(ij)} \\
& + \beta_{15} \times lg\_map_i \times rel\_angr_j \times zeit_{th(ij)} + \beta_{16} \times lg\_sch_i \times rel\_angr_j \times zeit_{th(ij)} \\
& + \beta_{17} \times lg\_map_i \times rel\_vert_j \times zeit_{th(ij)} + \beta_{18} \times lg\_sch_i \times rel\_vert_j \times zeit_{th(ij)} \\
& + u_{0i} + v_{0j} + f_{0h(ij)} \\
& + u_{1i} \times zeit_{th(ij)} + v_{1j} \times zeit_{th(ij)} + f_{1h(ij)} \times zeit_{th(ij)} \\
& + e_{th(ij)}
\end{aligned}
\tag{6.22}
$$

Auch wenn der Teil der Fixed Effects auf den ersten Blick sehr umfangreich wirkt, handelt es sich lediglich um die übliche Repräsentation eines Regressionsmodells mit zwei dreistufigen Faktoren. Die Zeilen, in denen die Koeffizienten β_0 vorkommen, entsprechen genau der Spezifikation des einfachen kreuzklassifizierten Modells. Da zusätzlich die Zeit-Variable in den Modellen enthalten ist, quantifizieren sie den Einfluss der L2-Prädiktoren zum Beginn der Antwort.

Nach der Logik der multiplikativen Regressionsmodelle mit Interaktionen steht β_{00} für den geschätzten RTR-Wert zu Beginn der Antworten durch die Unentschiedenen, wenn die Antwort nur Selbstpräsentationen enthält. β_{01} zeigt, wie sich die Bewertung durch die Anhänger von Mappus von dieser Bewertung unterscheidet. Der Koeffizient β_{03} gibt an, wie die Bewertung von Antworten, die einen Angriff enthalten, von der Bewertung von Selbstpräsentationen durch die Unentschiedenen abweicht. Dem Wert von β_{05} kann entnommen werden, wie sich die Bewertungen von Unentschiedenen und Anhängern von Mappus während der Antworten unterscheiden, die einen Angriff enthalten. Die Koeffizienten β_1 können nach der selben Logik für die Unterschiede in der Veränderung der Bewertungen über die Zeit interpretiert werden. Die Schwierigkeit besteht nun darin, dass uns in der Regel nicht nur interessiert, ob sich die Bewertungen der Kandidaten im Intercept oder im Slope unterscheiden. Stattdessen wollen wir wissen, ob sich die Bewertungen der Antworten insgesamt unterscheiden. Dies können wir mit dem beschriebenen Modell nur testen, indem wir die Modellverbesserung durch alle Fixed Effects gemeinsam betrachten. Die inhaltliche Interpretation der gesamten Effekte kann nur durch die Betrachtung der durch das Modell vorhergesagten RTR-Bewertungen erfolgen.

Komplexer als im einfachen kreuzklassifizierten Modell ist auch die Interpretation der Varianzdekomposition und der relativen Reduzierung der Varianzkomponenten. In der Darstellung der Wachstumskurvenmodelle haben wir bereits erläutert, dass die Varianzen der Random Intercepts nicht mit den Varianzen der Random Slopes verrechnet werden können und dass in Anwesenheit von Random Slopes kein Vergleich mit der L1-Residualvarianz möglich ist (Kapitel 6.1). Hier wird besonders deutlich, dass es in der Analyselogik der Mehrebenenmodelle zurecht kein einfaches R^2-artiges Maß der gesamten Varianzaufklärung gibt. Betrachten wir nochmals genau die Varianzkomponenten des unkonditionalen kreuzklassifizierten Modells mit einem L1-Messmodell aus Intercept und linearem Slope in Gleichung 6.21, so können wir sieben Varianzkomponenten identifizieren: jeweils eine auf den Intercept und den Slope bezogene Komponente für Rezipienten, Antworten und Messmodelle sowie die L1-Residualvarianz. Um einen ungefähren Eindruck von der relativen Bedeutung der auf Rezipienten, Antworten, Messmodelle und L1-Residuen bezogenen Varianz zu erhalten, können wir ein unkonditionales kreuzklassifiziertes Modell schätzen, das nur die Random Intercepts enthält. Da die Random Slopes, wie noch zu zeigen sein wird, die Residualvarianz verringern, können wir diese Befunde als eine untere Grenze der relativen Bedeutung der Varianzkomponenten auffassen. In den unkonditionalen Modellen, die wie in Gleichung 6.21 Random Intercepts und Random Slopes enthalten, können

wir zumindest die Varianzkomponenten der Intercepts und Slopes unter Vernachlässigung der Residualvarianz jeweils untereinander vergleichen. Damit erhalten wir Maße der relativen Bedeutung der Komponenten für die durch die Modelle prinzipiell erklärbare Varianz. Durch den Vergleich der jeweils äquivalenten Varianzkomponenten zwischen dem unkonditionalen Modell und einem Modell mit L2-Prädiktoren können wir dann zumindest relativ zeigen, wie stark die L2-Prädiktoren zur Erklärung der jeweiligen Varianzkomponente beitragen. So erhalten wir für das in den Gleichungen 6.21 und 6.22 dargestellte Modellpaar sechs Quantitäten der Varianzreduzierung durch die Faktoren Lagerzugehörigkeit und Relation.

6.3.2 Bewertung der Kandidaten während aller Antworten

In diesem Teilkapitel untersuchen wir die Effekte von Voreinstellungen und Relation auf die unmittelbare Bewertung der Kandidaten während ihrer Antworten. Zuerst evaluieren wir die Datenpassung der gewählten Spezifikation des Messmodells mit einem latenten Intercept und einem latenten Slope im Vergleich zu den weiteren in Kapitel 6.1 vorgestellten L1-Spezifikationen. Dann untersuchen wir in einfachen Modellen die Bewertung von Schmid und Mappus in Abhängigkeit von der Lagerzugehörigkeit der Rezipienten und der Relation der Antworten. Schließlich erweitern wir diese Modelle um weitere L2-Prädiktoren zu den Voreinstellungen der Rezipienten.

Die Einheit der Antworten ist in diesen Analysen der Merkmalsträger der Debatteninhalte. Unter Antworten verstehen wir die ersten maximal 30 Sekunden der Aussagen der Kandidaten, nachdem ein Moderator das Wort hatte.[76] Berücksichtigt werden alle Antworten, die mindestens 15 Sekunden dauern. Aus der sekundengenauen Inhaltsanalyse der Debatte (vgl. Kapitel 2 sowie Bachl, Käfferlein & Spieker, 2013a, 2013b) ziehen wir die Ausprägungen Selbstpräsentation, Angriff und Verteidigung der Kategorie Relation heran. Alle Antworten, die in den ersten 20 Sekunden einen Angriff oder eine Verteidigung mit einer Dauer von mindestens fünf Sekunden enthalten, werden in dichotomen Variablen als Angriff bzw. Verteidigung eingeordnet. Vier Antworten (drei von Schmid, eine von Mappus) enthalten sowohl einen Angriff als auch eine Verteidigung. Damit ist die Relation streng genommen kein Faktor, da sich die Ausprägungen nicht gegenseitig ausschließen. Für die Datenanalyse ist dies unerheblich, da beide Relationen als getrennte dichotome Variablen in die Modelle eingehen und das gemeinsame Auftreten sehr selten ist. Es sei jedoch darauf hingewiesen, dass die Bezeichnung der Relation als Faktor, die

76 Vergleiche dazu ausführlich die Erläuterung der Analyseeinheiten auf S. 173.

wir teilweise aus Gründen der Prägnanz einsetzen, formal nicht korrekt ist. Alle Antworten, in denen weder ein Angriff noch eine Verteidigung vorkommen, bestehen zum größten Teil aus Selbstpräsentationen, enthalten teilweise jedoch auch nicht identifizierbare Passagen und Lagebeschreibungen. Ob deren Vorkommen einen Einfluss auf die unmittelbaren Bewertungen der Kandidaten hat, kann wegen ihrer geringen Fallzahl nicht sinnvoll geprüft werden. Der Einfachheit halber bezeichnen wir Referenzausprägung der Antworten, die weder Angriffe noch Verteidigungen enthalten, als Selbstpräsentation. Nach dieser Definition sind von den 34 Antworten Schmids 13 Angriffe und vier Verteidigungen. Von Mappus liegen in 34 Antworten sechs Angriffe und 12 Verteidigungen vor.

Auf Ebene der Rezipienten werden alle Probanden berücksichtigt, die in den verwendeten Variablen zu den politischen Voreinstellungen gültige Werte aufweisen (n = 172). Die abhängigen Variablen sind im Folgenden die Bewertungen von Schmid und Mappus während der direkten Antworten der Kandidaten auf Fragen der Moderatoren. Anhand dieser Ausschnitte der Debatte können wir besonders gut untersuchen, wie die Rezipienten auf die von den Kandidaten vorgetragenen Inhalte reagieren, da die meisten Probanden ihr RTR-Dial während der Sprechzeit der Moderatoren auf den neutralen Skalenmittelpunkt bewegen. Da wir in dieser Auswertung an der Veränderung der RTR-Wertungen ausgehend von diesem Null-Punkt interessiert sind, schließen wir individuelle Verläufe von der Analyse aus, die außerhalb des Wertebereichs $[-15; 15]$ starten. Ebenso verzichten wir auf RTR-Verläufe, die aufgrund technischer Messfehler ungültige Werte aufweisen. Als Folge dieser Auswahl ergeben sich für die Antwort-Rezipienten-Kombinationen unterschiedliche Fallzahlen. Für die einzelnen Antworten liegen von 75 bis 162 Rezipienten RTR-Verläufe vor ($M = 139$, $SD = 15$). Von den einzelnen Rezipienten liegen gültige Bewertungen zu 10 bis 68 Antworten vor ($M = 55$, $SD = 13$). Zwischen dem Sprecher der Antwort, der Relation der Antwort und den relevanten Voreinstellungen der Rezipienten besteht kein systematischer Zusammenhang. Insgesamt beziehen sich die Analysen auf 9431 in Messmodellen erfasste RTR-Verläufe, die aus 255595 einzelnen RTR-Messungen bestehen.

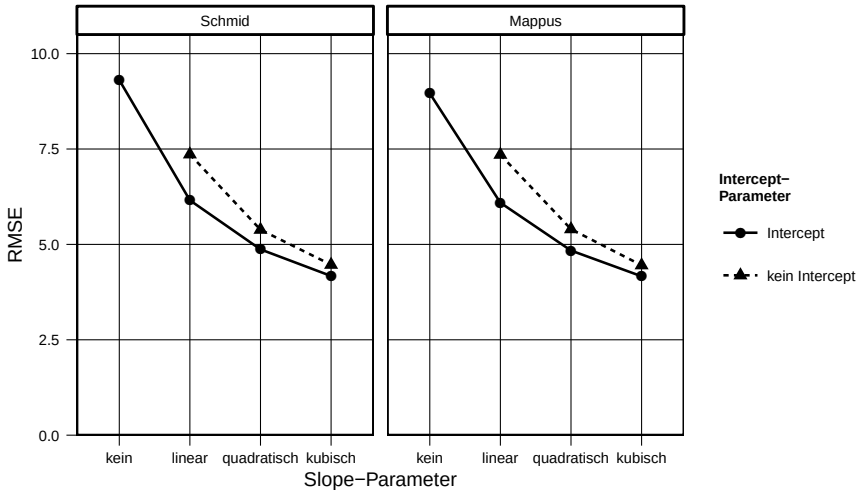
Datenpassung der gewählten L1-Spezifikation In den folgenden Analysen erfassen wir die dynamische Veränderung der unmittelbaren Kandidatenbewertungen innerhalb einer Antwort mit einem Messmodell, das aus einem latenten Intercept und einem latenten linearen Slope besteht. Für diese Auswahl gibt es drei wesentliche Gründe: Zuvorderst ist die sehr gute Interpretierbarkeit der Spezifikation zu nennen, die differenzierte inhaltliche Schlüsse zulässt.

Wie bereits anhand der Analyse der Bewertungen während einer einzelnen Antwort deutlich wurde (Kapitel 6.1), fängt der latente Intercept vor allem sehr schnelle und heftige Reaktionen auf, die für eine durch Schlüsselreize ausgelöste Veränderung der Bewertung sprechen. Im latenten Slope werden dagegen spätere und weniger starke Reaktionen sichtbar, die für eine Bewertung nach einer kognitiven Auseinandersetzung der Rezipienten mit den Inhalten des Stimulus sprechen. Weiter spricht für diese Spezifikation, dass sie die Dynamik der beobachteten RTR-Messungen wie auch der interindividuellen Varianz zwischen den Messungen sowohl zu Beginn der Antwort als auch in ihrem weiteren Verlauf angemessen repräsentiert. Schließlich erweist sich die numerische Evaluation der Modelle mit dieser L1-Spezifikation auch nach der Erweiterung um mehrere Prädiktoren als ausreichend stabil.

Empirisch können wir zeigen, dass die L1-Spezifikation mit Intercept und linearem Slope auch im Vergleich mit weiteren möglichen Spezifikationen eine akzeptable Datenpassung ausweist. Abbildung 6.29 zeigt zum Vergleich der Modellpassungen die Wurzel der mittleren quadrierten Abweichung der vorhergesagten von den beobachteten RTR-Messungen für die Bewertungen von Schmid und Mappus. Eine ausführliche Übersicht der geprüften Messmodelle findet sich in Tabelle 6.1 (S. 184) und Abbildung 6.4 (S. 183).

Die Vorhersagefehler der jeweils gleich spezifizierten Messmodelle zu den Bewertungen von Schmid und Mappus sind einander sehr ähnlich. Bei aller Vorsicht, die eine Verallgemeinerung auf der Basis von zwei Fällen verlangt, spricht dies für eine grundsätzlich sehr ähnlich funktionale Form der RTR-Verläufe von Bewertungen, die sich ausgehend von einem neutralen Skalenmittelpunkt verändern. Der Vorhersagefehler des Intercept-Only-Modells liegt deutlich über den Messmodellen, die die zeitliche Dynamik durch einen oder mehrere Slope-Parameter berücksichtigen. Wie bereits bei der Modellierung der Bewertung während einer einzelnen Antwort bringt die Aufnahme eines zweiten Parameters die relativ größte Verbesserung der Datenpassung. Besonders deutlich wird dies bei der Erweiterung des Intercept-Only-Modells durch einen Slope-Parameter zur von uns gewählten L1-Spezifikation. Die Berücksichtigung der Veränderung über die Zeit in einem Wachstumskurvenmodell ist damit gerechtfertigt.

Nur mit Blick auf die Datenpassung schneiden Messmodelle, die durch polynomische Slope-Terme nicht-lineare Verläufe realisieren, noch besser ab als das hier bevorzugte lineare Modell. Diese Modelle lassen sich jedoch weniger gut interpretieren. Zudem wird die numerische Evaluation der Modelle mit Polynomen höherer Ordnung der Zeitvariable zunehmend schwieriger, da die Terme des Polynoms innerhalb der Random Effects stark multikollinear sind. Insgesamt lässt sich die Wahl einer L1-Spezifikation aus latentem Intercept



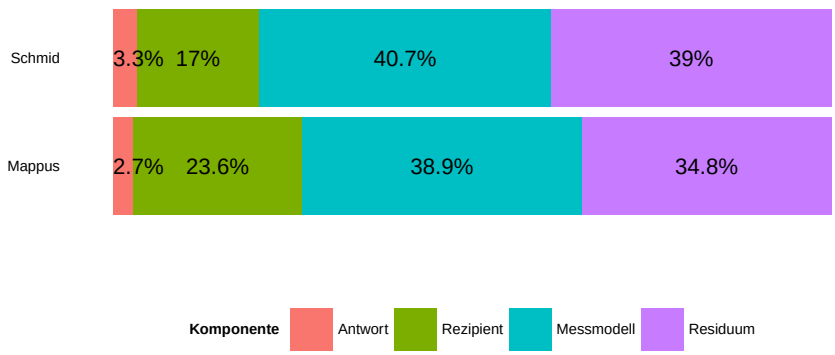
Anmerkungen
RMSE: Wurzel der mittleren quadrierten Abweichung der vorhergesagten von den beobachteten RTR-Messungen.

Abbildung 6.29: Vorhersagefehler der kreuzklassifizierten Wachstumskurvenmodelle

und latentem linearen Slope in Abwägung von inhaltlicher Interpretierbarkeit, Sparsamkeit und Datenpassung gut rechtfertigen. Allerdings sollten wir uns immer bewusst machen, dass diese lineare Approximation der Veränderung im Vergleich zu den beobachteten individuellen RTR-Verläufen eine starke Vereinfachung darstellt.[77]

Varianzdekomposition Um zuerst das Verhältnis der Varianzkomponenten, die sich auf Rezipienten, Antworten, ihre Interaktionen (Messmodelle) und L1-Residuen beziehen, bestimmen zu können, nehmen wir eine Varianzdekomposition des einfachen kreuzklassifizierten Modells der Bewertung während der Antworten vor. Es sei daran erinnert, dass dieses Modell die Veränderung

77 Um die Eignung einfacherer wie komplexerer Modelle im Vergleich zu der gewählten Spezifikation weiter zu prüfen, haben wir die folgenden Analysen auch für die in Kapitel 6.1 weiter verfolgten L1-Spezifikationen durchgeführt. Dabei finden wir keine Anzeichen, dass die gewählte Spezifikation im Vergleich zu anderen schlechter geeignet ist. Die Darstellung der Befunde aus diesen Modellen würde an dieser Stelle jedoch deutlich zu weit gehen.

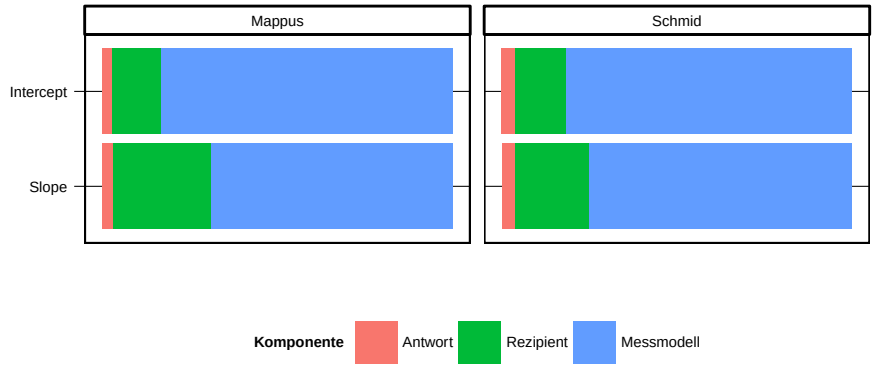


Anmerkungen
Varianzdekomposition von unkonditionalen kreuzklassifizierten Modellen mit Random Intercepts; Abweichungen von 100% sind Rundungsfehler.

Abbildung 6.30: Varianzdekomposition der Bewertungen von Schmid und Mappus während ihrer Antworten im kreuzklassifizierten Intercept-Only-Modell

der RTR-Messungen während der Antworten vernachlässigt und daher den Varianzanteil der L1-Residuen überschätzt. Da in Anwesenheit von Slopes in den Random Effects keine Dekomposition der gesamten Varianz möglich ist, dient uns dieses Modell als die bestmögliche Annäherung. Abbildung 6.30 zeigt die Anteile der Varianzkomponenten für die Bewertungen von Schmid und Mappus während ihrer Antworten.

Auch wenn wir die Antworten als Analyseeinheit für den Debatteninhalt heranziehen, zeigt sich das aus den Turn-Analysen vertraute Bild. Die Merkmale des Inhalts sind nur für einen sehr geringen Teil in der Streuung verantwortlich. Ein bedeutsamer Anteil der Varianz kann dagegen potenziell durch die Charakteristika der Rezipienten erklärt werden. Dieser Anteil ist für die Bewertungen von Mappus etwas größer, was darauf hindeutet, dass die Voreinstellungen der Rezipienten einen größeren Einfluss auf die Bewertungen von Mappus während des Duells hatten. Der größte Teil der prinzipiell durch diese Modellspezifikation erklärbaren Varianz entfällt auf die Messmodelle – sie kann nur durch Kombinationen von Rezipienten- und Antwortmerkmalen reduziert werden. Die modellbedingt nicht erklärbare L1-Residualvarianz macht im Intercept-Only-Modell etwas über ein Drittel der gesamten Varianz aus. Dieser



Anmerkungen
Varianzdekomposition von unkonditionalen kreuzklassifizierten Wachstumskurvenmodellen mit Random Intercepts und Random Slopes. Die Anteile beziehen sich nur auf die erklärbare Varianz ohne die L1-Residualvarianz.

Abbildung 6.31: Varianzdekomposition der Bewertungen von Schmid und Mappus während ihrer Antworten im kreuzklassifizierten Wachstumskurvenmodell

Bestandteil wird verringert, wenn wir im Folgenden das Messmodell durch einen Slope-Term erweitern. Der exakte Anteil lässt sich in diesen Modellen jedoch nicht mehr bestimmen.

Allerdings ist es zumindest möglich, die relative Bedeutung der den Antworten, Rezipienten und Messmodellen zugeordneten Varianz für die einzelnen Random Effects zu bestimmen. So haben im oben formal definierten Beispiel des unkonditionalen kreuzklassifizierten Wachstumskurvenmodells (vgl. Gleichung 6.21, S. 270) die Varianzen der Random Intercepts und die Varianzen der Random Slopes jeweils dieselbe Skala und können miteinander verglichen werden. Da die L1-Residualvarianz hier nicht berücksichtigt wird, beziehen sich die so ermittelten Anteile jeweils nur auf die durch das Modell prinzipiell erklärbare Varianz. Abbildung 6.31 zeigt die Dekomposition der erklärbaren Varianz in den unkonditionalen kreuzklassifizierten Wachstumskurvenmodellen der Bewertung von Mappus und Schmid.

Die Varianzdekomposition ergibt für die Bewertungen beider Kandidaten ein ähnliches Bild. Der mit Abstand größte Anteil entfällt jeweils auf die messmodellspezifische Komponente. Relativ bedeutsam ist zudem der Varianzanteil,

der durch die Merkmale der Rezipienten erklärt werden kann. Dabei fällt auf, dass der rezipientenbezogene Anteil der beiden Random-Effects-Terme sich recht deutlich unterscheidet. Ein größerer Teil der Varianz im Slope, der die Veränderung der RTR-Verläufe über den Zeitverlauf der Antwort angibt, ist den Rezipienten zuzuschreiben. Im Random Intercept des Modells fällt dieser Anteil geringer aus. Dies deutet darauf hin, dass vor allem in der Veränderung der Bewertungen über die Zeit Unterschiede zu finden sind, die auf die Voreinstellungen der Rezipienten zurückgehen. Für die Erklärung der Unterschiede im Intercept, in dem sich vor allem sehr schnelle und/oder sehr extreme Veränderungen der unmittelbaren Kandidatenbewertung niederschlagen, sind die Eigenschaften der Rezipienten etwas weniger wichtig. Dafür haben die spezifischen Kombinationen eine etwas größere Bedeutung. Hierunter fallen neben sämtlichen denkbaren Interaktionen von Antwort- und Rezipientenmerkmalen auch situationsbedingte Messfehler, wenn z.B. ein Rezipient zu Beginn einer Antwort weniger aufmerksam war und erst einige Zeit benötigte, bis er den Inhalt der Antwort so weit „aufgeholt" hatte, dass er ein Urteil abgeben konnte. Ähnlich könnte sich hier auch stärker auswirken, ob ein Rezipient den RTR-Regler gerade schon in der Hand hatte und daher sehr schnell regeln kann, oder ob er erst zum Dial greifen muss. Alle diese Messfehler sind unabhängig von den Inhalten der Debatte und den Einstellungen der Rezipienten. Sie ließen sich nur erklären, wenn wir Informationen darüber besäßen, in welcher Verfassung sich jeder Rezipient während jeder Antwort befunden hat. Daher ist es nicht verwunderlich, dass diese messmodellspezifischen Komponenten sehr groß sind und dass wir in unseren Modellen nur einen recht geringen Anteil davon erklären können.[78] Viele der potenziellen situationsbedingten Fehler beeinflussen unter anderem die Reaktionszeit der Rezipienten. Daher ist es plausibel, dass ihr relativer Anteil im Random Intercept größer ist.

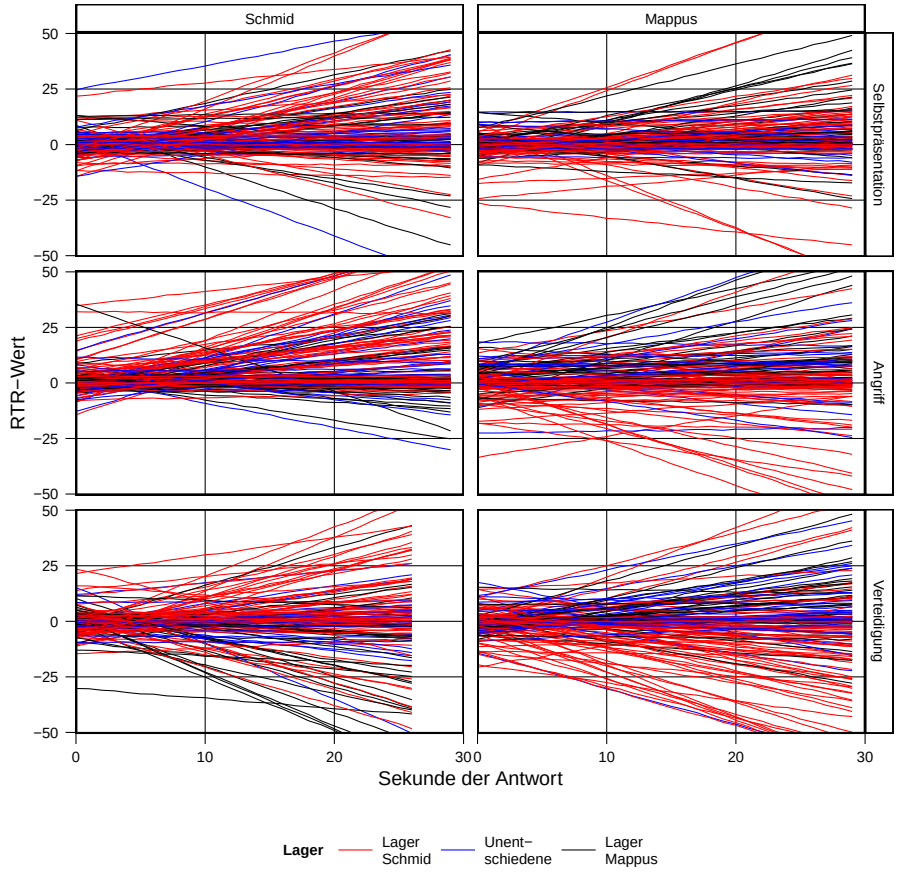
Über beide Kandidaten und beide Random-Effects-Terme hinweg ist der Varianzanteil, der von einfachen Effekten der Merkmale der Antworten erklärt werden kann, sehr klein. Wir können nicht erwarten, dass die Relationen (oder andere Merkmale des Debatteninhalts) auf dieser Ebene einen klaren einfachen Effekt wie „Angriffe werden von allen Rezipienten besser bewertet als Selbstpräsentationen" haben. Stattdessen wird es – wie schon bei den Effekten von Voreinstellungen und Thema in Kapitel 6.2 – vor allem auf konditionale Effekte der Relationen in Abhängigkeit von den Voreinstellungen der Rezipienten ankommen.

78 Vergleiche hierzu ausführlich die Erläuterung zu den Varianzkomponenten des kreuzklassifizierten Modells auf S. 222ff. und die Ergebnisse der Varianzreduzierungen in Kapitel 6.2.

Die Logik der Varianzdekomposition und der darauf folgenden Reduzierung der Varianzkomponeten in kreuzklassifizierten Wachstumskurvenmodellen kann anhand einer grafischen Darstellung der beiden Analyserichtungen einfacher nachvollzogen werden. Dazu zeigt Abbildung 6.32 die vorhergesagten RTR-Verläufe gruppiert nach Antworten für eine Auswahl von sechs Antworten (je eine Selbstpräsentation, ein Angriff und eine Verteidigung für jeden Sprecher). Die einzelnen Verläufe sind nach ihrer Lagerzugehörigkeit farblich codiert. Umgekehrt sind in Abbildung 6.33 die RTR-Verläufe nach Rezipienten gruppiert für drei aus den Lagern ausgewählte Rezipienten dargestellt. Farblich zugeordnet sind die Relationen der Antworten.

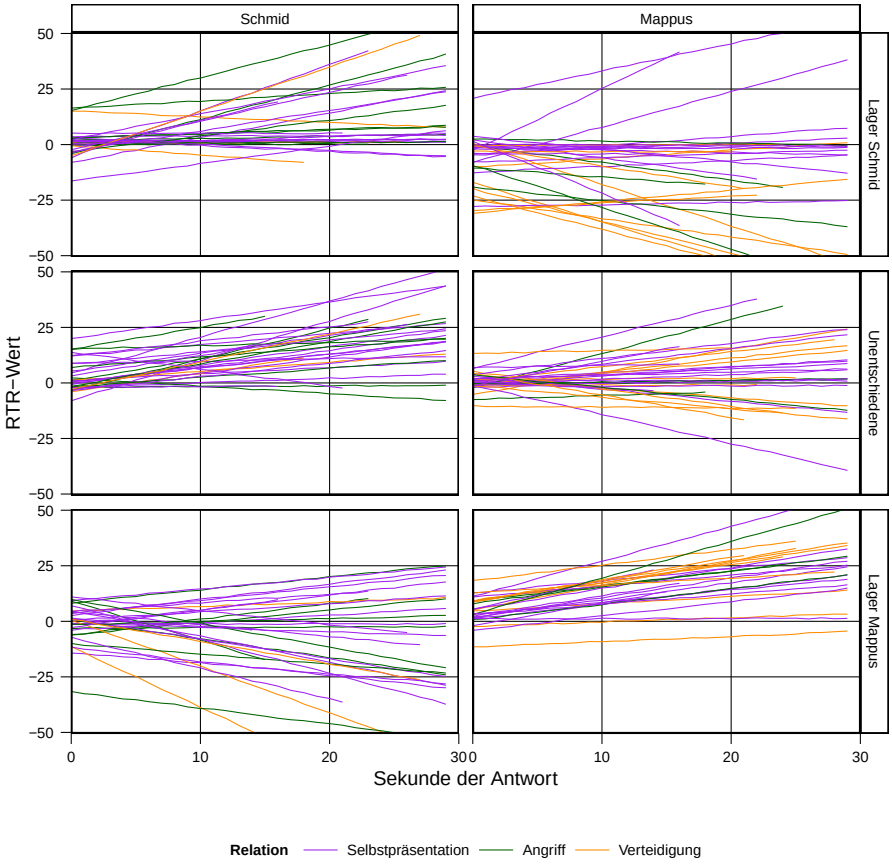
Die Abbildungen veranschaulichen die abstrakten Ergebnisse der Dekomposition der erklärbaren Varianz. Die einzelnen Antworten werden von den individuellen Rezipienten sehr unterschiedlich bewertet (Abbildung 6.32). Sowohl in den geschätzten Intercepts zu Beginn der Antworten als auch in den Veränderungen über die Zeit offenbart sich eine substantielle Streuung. Die Voreinstellungen der Rezipienten, hier dargestellt durch ihre Lagerzugehörigkeit, haben einen erkennbaren Einfluss auf die Bewertungen. Die Anhänger eines Kandidaten tendieren dazu, diesen als Sprecher bereits früher positiv zu bewerten (zu erkennen an den positiven Intercepts) und die Bewertung im Laufe der Antwort stärker zum Positiven zu verändern (steilere positive Slopes). Negative Bewertungen des gegnerischen Kandidaten sind dagegen im Aggregat schwächer ausgeprägt, sie kommen bei weniger Rezipienten vor. Eine Ausnahme ist die Bewertung der Verteidigung von Mappus. Hier sind viele negative RTR-Verläufe durch Angehörige des Lagers Schmid zu erkennen. Auch wenn dies nur eine einzelne exemplarische Verteidigung ist, so weist dies doch auf die Bedeutung der Interaktionen von Merkmalen der Antworten (hier: Verteidigung) und Voreinstellungen (hier: Zugehörigkeit zum gegnerischen Lager) hin. Eine ähnliche, wenn auch schwächere Tendenz ist bei der Bewertung der Verteidigung von Schmid zu erkennen. Schließlich gibt es auch unter den Unentschiedenen mehr individuelle Rezipienten, die jeweils den aktuellen Sprecher positiv bewerten, negative individuelle RTR-Verläufe sind in dieser Gruppe eher selten. Häufig geben die Angehörigen dieser Gruppe – wie auch viele andere Rezipienten – gar keine Wertung ab und sind daher in den Grafiken nur schwer auszumachen.

Abbildung 6.33 zeigt (wenn auch nur beispielhaft für drei Rezipienten), dass zwischen den Bewertungen der Antworten durch dieselben Rezipienten nur relativ wenig Varianz besteht. Die Rezipienten bewerten, wenn sie ihr RTR-Dial überhaupt weiter vom neutralen Skalenmittelpunkt entfernen, fast alle Antworten der Kandidaten mit derselben Grundtendenz. Die Interaktionen zwischen Rezipienten- und Antwortmerkmalen im Allgemeinen können mit



Anmerkungen
Bewertungen der Kandidaten auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher) für sechs exemplarische Antworten. Vorhersage durch das unkonditionale kreuzklassifizierte Wachstumskurvenmodell. Die farbliche Codierung der Verläufe entspricht der Lagerzugehörigkeit der Rezipienten.

Abbildung 6.32: Vorhergesagte Kandidatenbewertungen während sechs Antworten



Anmerkungen
Bewertungen der Kandidaten auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher) für drei exemplarische Rezipienten. Vorhersage durch das unkonditionale kreuzklassifizierte Wachstumskurvenmodell. Die farbliche Codierung der Verläufe entspricht der Relation der Antworten.

Abbildung 6.33: Vorhergesagte Kandidatenbewertungen durch drei Rezipienten

einer visuellen Inspektion nur schwer herausgearbeitet werden. Einige Hinweise auf eine Interaktion der spezifischen Merkmale Lagerzugehörigkeit und Relation sind bei genauer Betrachtung zu erkennen. In Abbildung 6.32 scheint die exemplarische Verteidigung von Mappus weniger gut bewertet zu werden als die Selbstpräsentationen – zumindest von den Anhängern der Oppositionsparteien. Auch in Abbildung 6.33 gehören die Verteidigungen meist zu den von beiden dem jeweils gegnerischen Lager zugeordneten Rezipienten negativer bewerteten Aussagen. Insgesamt können wir festhalten, dass offenbar die Voreinstellungen der Rezipienten wesentlich dazu beitragen, die Grundtendenz der Bewertungen (vor allem des eigenen Kandidaten) zu erklären. Ein einfacher Zusammenhang zwischen Relation und Bewertungen kann auf Basis der visualisierten Beispiele nicht festgestellt werden. Es gibt jedoch einige Hinweise auf konditionale Effekte der Relationen in Abhängigkeit von den Voreinstellungen. Im Folgenden werden wir diese ersten qualitativen Eindrücke anhand der systematischen Analyse der Bewertung aller Antworten durch alle Rezipienten überprüfen.

Effekte der Relation und der Lagerzugehörigkeit: Modellprüfung Im folgenden Abschnitt prüfen wir, ob die Lagerzugehörigkeit der Rezipienten, die Relation der Antworten und die Interaktion dieser Rezipienten- und Antwortmerkmale Einfluss auf die Bewertung der Kandidaten während ihrer Antworten haben. Die Analysen werden separat für die Kandidaten durchgeführt. Um die Vergleichbarkeit der Resultate zu verbessern, nutzen wir für die Bewertung von Schmid wieder die umcodierte RTR-Skala. Die unmittelbare Bewertung der Kandidaten als abhängige Variable reicht damit von -50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Die Ergebnisdarstellung ist in mehrere Schritte gegliedert, die parallel für Schmid und Mappus nachvollzogen werden.

Für beide Kandidaten müssen das unkonditionale Modell und das volle Modell mit allen L2-Prädiktoren und ihren Interaktionseffekten geschätzt werden. Um statistische Signifikanz und Beitrag zur Modellverbesserung der Fixed Effects zu bestimmen, muss für alle Fixed-Effect-Terme der höchsten Ordnung (d.h. für alle Interaktionseffekte sowie gegebenenfalls für alle einfachen Effekte der Terme, die nicht in den Interaktionstermen enthalten sind) jeweils ein Modell geschätzt werden, in dem die Fixed-Effects-Terme der jeweiligen Einflussgröße fehlen. Durch den Vergleich dieser reduzierten Modelle mit dem vollen Modell ergibt sich, ob die jeweiligen Terme zur Erklärung der Kandidatenbewertung beitragen. Dieses aufwändige Vorgehen ist notwendig, da es nicht nur von Interesse ist, die Signifikanz eines einzelnen Koeffizienten

zu bestimmen. So ist es nicht nur relevant, ob sich z.B. die Bewertungen von Angriffen im Intercept *oder* im Slope unterscheiden. Wir wollen auch wissen, ob die Relation Angriff insgesamt einen signifikanten Einfluss auf die durch beide Koeffizienten bestimmten Bewertungen eines Kandidaten hat. Den simultanen Test mehrerer Koeffizienten können wir nur durch solche Modellvergleiche vornehmen. Im zweiten Schritt betrachten wir die Varianzreduzierung im Vergleich zu den unkonditionalen Modellen. Auch diese Analyse fällt etwas umfangreicher aus, da wir, wie im vorangegangenen Abschnitt dargestellt, zwei Parameter (Intercept und Slope) in jeder Varianzkomponente beachten müssen. Im dritten Schritt wenden wir uns schließlich der Richtung der Effekte zu. Bei diesen Analysen stützen wir uns vor allem auf die Plots der durch die Modelle vorhergesagten RTR-Bewertungen im Verlauf der Antworten.

Tabelle 6.20 weist für die Modelle zur Erklärung der unmittelbaren Bewertung von Schmid bzw. Mappus durch die Lagerzugehörigkeit der Rezipienten und die Relation der Antworten die Kennzahlen zweier Tests aus, um den Beitrag der Interaktionen der Lagerzugehörigkeit mit den Relationen Angriff bzw. Verteidigung festzustellen. Beide Tests basieren auf dem Vergleich der Güte von Modellen mit und ohne Interaktionsterm. Durch den Vergleich des Informationskriteriums $AICc_m$ wird geprüft, ob die Aufnahme der Interaktion das Modell im Verhältnis zur dadurch entstehenden Komplexitätssteigerung verbessert. Dargestellt ist in den Tabellen die Differenz ΔAIC der Informationskriterien der jeweiligen Modelle. Ein niedrigerer Wert für das Modell mit Interaktionsterm und damit ein negatives ΔAIC spricht für die Annahme des Modells. Die Modelle mit und ohne Interaktionsterm werden zudem mit einem LR-Test verglichen. Der χ^2-basierte Test ist signifikant, wenn das Weglassen des Interaktionsterms zu einer Verschlechterung der Modellgüte führt. An dieser Stelle sei nochmals darauf hingewiesen, dass der LR-Test von Fixed Effects zu einer erhöhten α-Fehlerrate neigt (z.B. Manor & Zucker, 2004; Pinheiro & Bates, 2000). Für den gleichzeitigen Test mehrerer Fixed-Effect-Terme stehen jedoch keine Alternativen zur Verfügung. Im Zweifelsfall ist die Orientierung am Vergleich der $AICc_m$-Werte ratsam, auch wenn dieser kein vertrautes Maß der statistischen Signifikanz bereitstellt.

Formal werden in Tabelle 6.20 die folgenden nicht gerichteten Hypothesen zur Interaktion der L2-Prädiktoren geprüft:

> **H1**: Der Effekt von Verteidigungen auf die unmittelbare Bewertung der Kandidaten wird durch die Lagerzugehörigkeit der Rezipienten moderiert.

Tabelle 6.20: Vergleich der Modelle zur Erklärung der Bewertung von Schmid und Mappus während aller Antworten durch Relation und Lagerzugehörigkeit

	Schmid				Mappus			
	ΔAIC	χ^2	df	p	ΔAIC	χ^2	df	p
Lager X Angriff	1	10	4	.039	−12	23	4	<.001
Lager X Verteidigung	−10	21	4	<.001	−30	41	4	<.001

Anmerkungen
Differenz im Informationskriterium $AICc_m$ und Kenngrößen der LR-Tests im Vergleich eines Modells mit und ohne die in der Zeile genannten Fixed-Effect-Terme.
Schmid: $n_{Antworten}$ = 34, $n_{Rezipienten}$ = 172, $n_{Messmodelle}$ = 4726, $n_{RTR\text{-}Messungen}$ = 130418.
Mappus: $n_{Antworten}$ = 34, $n_{Rezipienten}$ = 172, $n_{Messmodelle}$ = 4705, $n_{RTR\text{-}Messungen}$ = 125177.

> **H2**: Der Effekt von Angriffen auf die unmittelbare Bewertung der Kandidaten wird durch die Lagerzugehörigkeit der Rezipienten moderiert.

Die linke Hälfte von Tabelle 6.20 berichtet die Tests für die Erklärung der unmittelbaren Bewertung von Schmid. Der Effekt der Verteidigungen wird signifikant durch die Lagerzugehörigkeit moderiert. Wir können damit davon ausgehen, dass die Bewertung von Verteidigungen in Abhängigkeit von der Lagerzugehörigkeit der Rezipienten unterschiedlich von der Bewertung der Referenzausprägung der Selbstpräsentationen abweicht. Schwieriger ist der Befund zum von der Lagerzugehörigkeit moderierten Effekt der Angriffe einzuordnen. Hier zeigt sich die größere Sensibilität des LR-Tests für Fixed Effects: Er ist auf $p < .05$ signifikant, das Informationskriterium $AICc_m$ steigt allerdings leicht an. Betrachten wir die einzelnen Koeffizienten des Terms Angriff X Lager (vgl. Tabelle A.8), so fällt auf, dass eine gewisse Differenzierung zwischen den Lagern stattfindet. Sowohl im Intercept als auch im Slope weichen die Bewertungen durch die Anhänger der beiden Kandidaten leicht in der erwarteten Richtung von den Unentschiedenen ab. Die einzelnen Koeffizienten sind jedoch ebenfalls nicht statistisch signifikant. Dies ist auch deswegen der Fall, da sich der Effekt recht gleichmäßig auf Intercept- und Slope-bezogene Koeffizienten verteilt, diese jedoch dann jeweils recht unpräzise geschätzt werden. Insgesamt können wir festhalten, dass die Tests übereinstimmend für einen von der Lagerzugehörigkeit moderierten Effekt der Verteidigungen sprechen (H1 für Schmid gestützt). Der von der Lagerzugehörigkeit moderierte Effekt der Angriffe ist – falls er existiert – offenbar recht schwach und/oder uneinheitlich.

Der zweite Teil von Tabelle 6.20 präsentiert die entsprechenden Tests für das Modell zur Erklärung der unmittelbaren Bewertungen von Mappus. Auch hier haben die Verteidigungen einen konditionalen Effekt in Abhängigkeit von der Lagerzugehörigkeit der Rezipienten. Zudem zeigt sich ein konditionaler Effekt der Angriffe, der im Vergleich zum konditionalen Effekt der Verteidigungen etwas weniger zur Modellverbesserung beiträgt. Gestützt werden damit beide Hypothesen zu den konditionalen Effekten der Relationen in Abhängigkeit der Lagerzugehörigkeit (H1 und H2 für Mappus gestützt). Im Vergleich der Kandidaten scheinen die Effekte für Mappus deutlicher zu sein als für Schmid. Dies ist ein weiterer Hinweis darauf, dass die Voreinstellungen die Bewertung von Mappus während des Duells prägen und dass sie auch die Verarbeitung der Inhalte (hier: der Relationen) beeinflussen.

Um den Erklärungsbeitrag der L2-Prädiktoren Lagerzugehörigkeit und Relation quantifizieren zu können, sind in Tabelle 6.21 die relativen Reduzierungen der Varianzkomponenten (R_1^2 bzw. R_2^2) im Vergleich zu den entsprechenden unkonditionalen Modellen ohne L2-Prädiktoren dargestellt. Die Kennzahlen sind vor dem Hintergrund der Befunde der Varianzdekomposition (vgl. Abbildung 6.31) zu interpretieren.

Tabelle 6.21: Varianzerklärung der Modelle zur Erklärung der Bewertung von Schmid und Mappus während aller Antworten durch Relation und Lagerzugehörigkeit

	Schmid	Mappus
$R^2_{2\,\text{Intercept} \mid \text{Antworten}}$	−.07	.17
$R^2_{2\,\text{Slope} \mid \text{Antworten}}$	−.02	.19
$R^2_{2\,\text{Intercept} \mid \text{Rezipienten}}$	.06	.12
$R^2_{2\,\text{Slope} \mid \text{Rezipienten}}$	.24	.28
$R^2_{1\,\text{Intercept} \mid \text{Messmodelle}}$	.00	.01
$R^2_{1\,\text{Slope} \mid \text{Messmodelle}}$	.00	.00

Anmerkung
Reduzierung der Varianzkomponenten (R_1^2 bzw. R_2^2) im Vergleich zu den unkonditionalen Modellen.

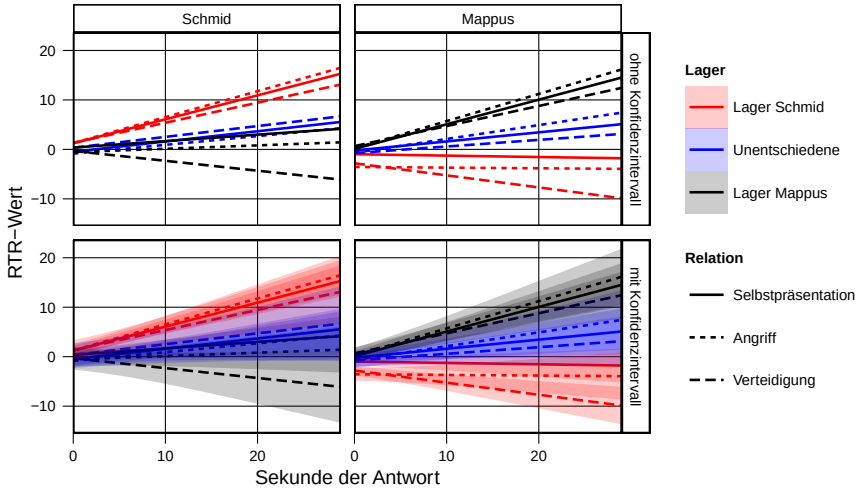
Das Modell zur Erklärung der Bewertung von Schmid leistet bereits durch die einfache Operationalisierung der Voreinstellungen durch die Lagerzugehörigkeit gute Dienste. Ein bedeutsamer Anteil der rezipientenbezogenen Varianzkomponenten wird durch diesen L2-Prädiktor erklärt. In Bezug auf die

Messmodelle, in denen die größten Varianzanteile liegen, zeigt sich die bereits aus der kreuzklassifizierten Modellierung vertraute geringe Varianzreduzierung. Keinen Beitrag leisten die Modelle zur Erklärung der antwortbezogenen Varianz. Hier finden sich negative Werte, die für eine R^2-artiges Maß nicht intuitiv zu interpretieren sind. Diese negativen Werte können auftreten, wenn nur ein sehr geringer Anteil einer kleinen Varianzkomponente erklärt wird und eine nicht balancierte Datenstruktur vorliegt (Snijders & Bosker, 2011, S. 110). Genau dies ist hier der Fall: Die antwortbezogene Varianzkomponente macht nach den Befunden der Varianzdekomposition nur einen sehr geringen Teil der Varianz in den unmittelbaren Urteilen aus. Durch den Ausschluss von RTR-Verläufen, die fehlende Werte aufweisen oder deren Start zu weit vom neutralen Nullpunkt der Skala beginnt, ist die Datenstruktur nicht balanciert. Die in ihrem Betrag geringen negativen Werte liefern hier keinen generellen Anlass zur Besorgnis über die Qualität der Modellschätzung. Sie weist aber deutlich darauf hin, dass einfache, für alle Rezipienten gleichgerichtete Effekte der Relationen Angriff und Verteidigung nicht zur Erklärung der unmittelbaren Bewertung von Schmid beitragen.

Die Varianzreduzierung des Modells für Mappus zeigt hinsichtlich der rezipienten- und messmodellbezogenen Komponenten ähnliche Muster. Die Erklärungskraft erscheint im Vergleich zu Schmid in (fast) allen Komponenten etwas größer – ein weiterer Hinweis auf die prägende Wirkung der Voreinstellung für die Bewertung des Kandidaten während des Duells. Auch die Schätzung der Varianzreduzierung in der antwortbezogenen Komponente ist in diesem Modell stabil. Etwa ein knappes Fünftel der Varianz kann durch einfache Effekte der Relationen erklärt werden. Bei der Einordnung dieses Befunds ist jedoch zu beachten, dass diese Komponente auch in den Bewertungen von Mappus die mit Abstand unbedeutendste ist. Wir können damit festhalten, dass die Relation der Antworten einen substantiellen Teil der insgesamt aber wenig bedeutsamen antwortbezogenen Varianz erklärt.

Insgesamt offenbaren die Befunde über die Kandidaten hinweg ein einheitliches Muster. Die Größenordnungen der Varianzreduktion ähneln sich mit Ausnahme des erläuterten Unterschieds bei den antwortbezogenen Komponenten weitgehend. Übereinstimmend ergibt sich die jeweils (etwas) stärkere Reduzierung im Slope-Parameter. Hier liegt offenbar diejenige Varianz, auf die mit der Modellierung der Rezipienten- und Antworteigenschaften zugegriffen werden kann.

Effekte der Relation und der Lagerzugehörigkeit: Richtung der Effekte Bis zu diesem Punkt haben wir festgestellt, dass von der Lagerzugehörigkeit mo-



Anmerkungen
Bewertung von Schmid und Mappus auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Vorhersage durch die Modelle in den Tabellen A.8 (Schmid) bzw. A.9 (Mappus). Die Flächen in den Abbildungen der zweiten Zeile zeigen 95%-Konfidenzintervalle.

Abbildung 6.34: Effekte der Relation und der Lagerzugehörigkeit auf die Bewertung von Schmid und Mappus während aller Antworten

derierte Effekte der Verteidigungen und (bei Schmid mit Einschränkungen) Angriffe auf die Bewertung der Kandidaten vorliegen. Die Befunde der Varianzreduzierung weisen zudem darauf hin, dass die Lagerzugehörigkeit als Rezipientenmerkmal einen deutlichen einfachen Effekt auf die unmittelbaren Bewertungen hat. Nun soll es darum gehen, die Richtung dieser Effekte zu beschreiben. Dazu ziehen wir die auf Basis der Modelle vorhergesagten RTR-Verläufe in Abhängigkeit von Lagerzugehörigkeit und Relation heran. In der ersten Zeile von Abbildung 6.34 verzichten wir auf die Darstellung der statistischen Unsicherheit, um die Übersichtlichkeit zu erhöhen und damit die Interpretation zu erleichtern. Die zweite Zeile von Abbildung 6.34 stellt zusätzlich die 95%-Konfidenzintervalle um die vorhergesagten Verläufe herum dar. So wird verdeutlicht, welches Ausmaß die einzelnen Effekte im Verhältnis zur Unsicherheit der Schätzung haben.

Auf den ersten Blick wird deutlich, dass die Lagerzugehörigkeit den wesentlichen Einfluss auf die Bewertung beider Kandidaten hat. Die vorhergesagten RTR-Verläufe sind nach den Lagern geordnet: Beide Kandidaten werden von ihrem Lager am besten bewertet, es folgen die Wertungen der Unentschiedenen und der Anhänger des Gegenkandidaten. Dieser Befund ist angesichts des Forschungsstands und der vorangegangenen Analysen kaum überraschend. Bemerkenswert ist dagegen, wie sich der konditionale Effekt der beiden Relationen in Abhängigkeit der Lagerzugehörigkeit der Rezipienten zeigt. Die Anhänger eines Kandidaten bewerten diesen positiv, mehr oder minder unabhängig davon, ob er sich selbst präsentiert, den Gegenkandidaten angreift oder sich verteidigt. Die Effekte der Relationen zeigen sich vor allem bei der Bewertung durch die Anhänger des Gegenkandidaten.

Auffällig ist dabei vor allem der konditionale Effekt der Verteidigung. Wenn ein Kandidat in der Defensive ist und seine Position rechtfertigt, wird er von den Anhängern des Gegners zunehmend negativ bewertet. Besonders gut zu erkennen ist dies bei der Bewertung von Mappus. Der geschätzte Intercept der Bewertungen durch die Anhänger Schmids liegt bei den Verteidigungen bereits um einiges niedriger als bei den Selbstpräsentationen. Dies spricht dafür, dass in diesem Lager bereits relativ früh während der Antworten negative Bewertungen vorkommen. Angesichts des Auslösers einer Verteidigung ist dies plausibel. Die meisten Verteidigungen von Mappus folgen auf kritische Fragen der Moderatoren. Die Rezipienten wissen also bereits, dass sich Mappus nun gleich zu einem Punkt äußern muss, bei dem er sich wird rechtfertigen müssen. Da die Anhänger des rot-grünen Lagers Mappus gegenüber ohnehin negativ eingestellt sind, liegt es nahe, dass sie auch seine Position zu diesem offenbar kritischen Thema ablehnen. Wenn Mappus dann beginnt, sich zu diesem Thema zu verteidigen und damit die Aufmerksamkeit weiter auf diese Position lenkt, wird er sehr schnell negativ bewertet. Auch im weiteren Verlauf der Verteidigungen nehmen die negativen Bewertungen zu. Es gelingt Mappus nicht, die Anhänger der Opposition mit den Verteidigungen zu überzeugen. Unter den Unentschiedenen und den eigenen Anhängern ist dieser negative Effekt der Verteidigungen kaum sichtbar. Die Richtung der geschätzten RTR-Verläufe ist auch bei den Verteidigungen positiv, die Steigung nur wenig geringer als bei den anderen Relationen. Im Muster ähnlich fallen auch die Reaktionen der Regierungsanhänger auf die Verteidigungen von Schmid aus. Hier finden sich die einzigen negativen Veränderungen für die Bewertungen des Oppostionskandidaten. Allerdings ist der Effekt seiner Verteidigungen weniger stark ausgeprägt, und die Intercept-Schätzung weist nicht auf besonders heftige oder schnelle Reaktionen hin.

Die Angriffe von Mappus werden von den Anhängern der Opposition ebenfalls etwas negativer bewertet als seine Selbstpräsentationen. Der negative geschätzte Intercept in dieser Gruppe deutet darauf hin, dass diese Reaktionen bei einigen Rezipienten in dieser Gruppe vor allem relativ schnell und stark erfolgen, dann aber kein weiteres Absinken der durchschnittlichen Bewertung mehr festzustellen ist. Auch wenn die Bewertungen der Relationen innerhalb der Gruppen der Unentschiedenen und der eigenen Anhänger sich kaum unterscheiden, so ist es doch bemerkenswert, dass die Reihenfolge von Selbstpräsentationen und Angriffen bei diesen Gruppen umgekehrt ist wie bei den Oppositionsanhängern: Erste bewerten Angriffe etwas besser, zweite Angriffe schlechter als Selbstpräsentationen. Das Ausmaß der negativen Polarisierung – also die Ablehnung durch die Gegner – ist stärker als die positive Polarisierung – also die Zustimmung und damit mögliche Mobilisierung durch die eigenen Anhänger. Jedoch führen Angriffe auch nicht zur Ablehnung durch die Unentschiedenen, wie es manchmal nach dem aus der Forschung zu Negative Campaigning bekannten „Backlash-Effekt“ (Spieker, 2011) vermutet wird. Der konditionale Effekt der Angriffe Schmids ist nur sehr schwach ausgeprägt. Seine Angriffe werden in allen Gruppen – selbst von den Regierungsanhängern – nur unwesentlich anders bewertet als die Selbstpräsentationen. Keinerlei Indiz findet sich in dieser Analyse des Duells Mappus gegen Schmid für die an anderer Stelle vermuteten positiven einfachen Effekte der Angriffe auf alle Rezipienten (z.B. Nagel et al., 2012; Strömbäck et al., 2009).

Alles in allem fallen die konditionalen Effekte der Relationen nicht nur im Vergleich mit dem Effekt der Lagerzugehörigkeit, sondern auch relativ zur statistischen Unsicherheit, die durch die Variation in den RTR-Messungen entsteht, schwach aus. Dies geht deutlich aus der zweiten Zeile von Abbildung 6.34 hervor. Die Unübersichtlichkeit dieser Darstellungsform ist bewusst gewählt. Wenn wir die Unsicherheit durch Konfidenzintervalle um die vorhergesagten mittleren Bewertungen visualisieren, so lässt sich fast ausschließlich die prägende Wirkung der Lagerzugehörigkeit erkennen. In dieser Hinsicht können wir festhalten, dass die konditionalen Effekte der Relationen in Abhängigkeit der Lagerzugehörigkeit zwar erkennbar herausgearbeitet werden können. Damit bietet sich ein wichtiger Ansatzpunkt, um die Verarbeitung und unmittelbare Bewertung bestimmter Merkmale von politischen Botschaften besser zu verstehen. Viel relevanter zur Erklärung der unmittelbaren Bewertung der Kandidaten im TV-Duell sind jedoch die politischen Voreinstellungen, mit denen die Rezipienten in das Duell gehen. Selbst wenn sie wie in dieser Analyse lediglich durch die Zuordnung der Wahlabsicht zum Lager eines Kandidaten operationalisiert werden, tragen sie ganz wesentlich zur Erklärung der Kandidatenbewertung während der Debatte bei.

Zuletzt können die Korrelationen zwischen latentem Intercept und latentem linearem Slope innerhalb der gruppenbildenden Faktoren betrachtet werden, um unsere Interpretation von auf die beiden Parameter bezogenen Effekten abzusichern. Innerhalb der Messmodelle bestehen leicht negative Korrelationen zwischen dem latenten Intercept und dem latenten Slope (Mappus: $-.24$; Schmid: $-.25$). Dies bedeutet, dass bei der einzelnen Bewertung einer Antwort durch einen Rezipienten entweder der Intercept oder der Slope sich stark in eine Richtung verändert und dann durch die Restriktionen des zweiparametrischen linearen Modells erzwungen der andere Koeffizient in der anderen Richtung ausfällt. Weniger abstrakt ausgedrückt: Wenn ein Rezipient einen Kandidaten bereits sehr früh während einer Antwort sehr positiv bewertet, so kann dies im Modell am besten durch einen höheren Intercept abgebildet werden. Da die RTR-Skala aber beschränkt ist, kann von diesem geschätzten Ausgangswert nur noch eine schwache weitere Steigung im Verlauf der Antwort folgen – der Slope-Koeffizient ist also recht klein. Betrachten wir die Korrelationen zwischen Intercept und Slope innerhalb der individuellen Rezipienten, so stellen wir einen positiven Zusammenhang fest (Mappus: .44; Schmid: .19). Wenn ein Kandidat durch einen Rezipienten im Durchschnitt positiv bewertet wird, so zeigt sich dies manchmal (bei sehr frühen und extremen Veränderungen der Bewertung) in einem positiven Intercept-Koeffizienten, manchmal (bei späteren oder weniger extremen Veränderungen der Bewertung) in einem positiven Slope-Koeffizienten. Beide Parameter tragen im Messmodell dieser L1-Spezifikation dazu bei, die durchschnittliche Bewertungen des Kandidaten während des Duells durch die individuellen Rezipienten konsistent zu erfassen.

Effekte der Relation und der Voreinstellungen Die Varianzdekomposition hat ergeben, dass die rezipientenbezogene Varianzkomponente das im Vergleich zur antwortbezogenen Komponente größere Potenzial zur Modellerweiterung bietet. Im nächsten Schritt nehmen wir daher die Voreinstellungen zu den Kandidaten und ihren Parteien zusätzlich zur Lagerzugehörigkeit in die Modelle auf, um weitere Teile der rezipientenbezogenen Varianz zu erklären. Wie in den vorangegangenen Analysen schätzen wir hierzu erst die vollständigen Modelle, die alle L2-Prädiktoren und ihre Cross-Level-Interaktionen enthalten. Um die Kommunizierbarkeit der Resultate zu erhöhen, reduzieren wir die vollen Modelle um die Terme, die nicht signifikant zur Erklärung der unmittelbaren Kandidatenbewertung beitragen. Die Abbildungen auf den folgenden Seiten zeigen die durch diese „optimalen“ Modelle vorhergesagten RTR-Verläufe. Dabei werden die Effekte der einzelnen Prädiktoren wieder-

um unter Annahme neutraler und typischer Ausprägungen in den übrigen L2-Prädiktoren dargestellt.

Die linke Hälfte von Tabelle 6.22 präsentiert die $AICc_m$-Vergleiche und LR-Tests der Effekte höchster Ordnung für die Modelle zur Erklärung der unmittelbaren Bewertung von Schmid. Der Faktor Lager, repräsentiert durch die beiden Dummy-Variablen Lager Schmid und Lager Mappus (Unentschiedene als Referenzausprägung), bleibt trotz der Aufnahme weiterer Variablen zur Operationalisierung der politischen Voreinstellungen im Modell erhalten. Zusätzlich tragen die mit Skalometerfragen gemessenen Einstellungen zu beiden Kandidaten zur Erklärung der Bewertung von Schmid während der Debatte bei. Dies spricht dafür, dass die über die Wahlabsicht vor der Debatte operationalisierte Lagerzuordnung für die Erklärung der unmittelbaren Bewertung von Schmid wichtige Informationen enthält, die nicht durch die Einstellungen zu den Kandidaten und ihren Parteien abgebildet werden. Die Voreinstellungen zu den Kandidaten leisten über die Lagerzugehörigkeit hinaus einen zusätzlichen Erklärungsbeitrag, verdrängen aber den einfachen Faktor nicht aus dem Modell. Die Einstellungen zu den Parteien, die in einer Modellierung ohne den Faktor Lager zur Erklärung der Bewertung von Schmid während des Duells beitragen, werden aber offenbar besser durch die einfache Lagerzuordnung abgebildet.

Tabelle 6.22: Vergleich der Modelle zur Erklärung der Bewertung von Schmid und Mappus während aller Antworten durch Relation und Voreinstellungen

	Schmid					Mappus			
Fixed-Effect-Term	ΔAIC	χ^2	df	p	Fixed-Effect-Term	ΔAIC	χ^2	df	p
Lg. Schmid	−4	10	2	.007	Sk. CDU	1	5	2	.089
Sk. Schmid	−13	18	2	<.001	Lg. Sch. X Angr.	−17	23	2	<.001
Sk. Map. X Angr.	−11	16	2	<.001	Sk. Sch. X Vert.	−1	7	2	.029
Lg. Map. X Vert.	−15	21	2	<.001	Sk. Map. X Vert.	−82	88	2	<.001

Anmerkungen
Differenz im Informationskriterium $AICc_m$ und Kenngrößen der LR-Tests im Vergleich eines Modells mit und ohne die in der Zeile genannten Fixed-Effect-Terme.
Schmid: $n_{Antworten}$ = 34, $n_{Rezipienten}$ = 172, $n_{Messmodelle}$ = 4726, $n_{RTR\text{-}Messungen}$ = 130418.
Mappus $n_{Antworten}$ = 34, $n_{Rezipienten}$ = 172, $n_{Messmodelle}$ = 4705, $n_{RTR\text{-}Messungen}$ = 125177.

Interessanterweise findet sich für die beiden Variablen, die sich auf den Sprecher beziehen – die Zugehörigkeit zum Lager Schmids und die Einstellung zu Schmid – einfache Effekte, jedoch keine Interaktionen mit den Relationen. Deren moderierte Effekte zeigen sich in Abhängigkeit von den Variablen, die sich auf den Gegenkandidaten Mappus beziehen. Der Effekt der Angriffe wird von

der Voreinstellung zu Mappus moderiert. Der Effekt der Verteidigungen fällt für die Angehörigen des schwarz-gelben Lagers anders aus als für die übrigen Rezipienten. Im Vergleich zur einfacheren Analyse zum Einfluss der Lagerzugehörigkeit fällt auf, dass durch das zusätzliche Berücksichtigen der Einstellung zu Mappus nun auch ein konditionaler Effekt der Angriffe deutlicher gezeigt werden kann. Auch der Vergleich der $AICc_m$, der in der einfachen Analyse zu widersprüchlichen Ergebnissen kommt, weist hier übereinstimmend mit dem LR-Test auf die Existenz einer solchen Interaktion hin.

Die äquivalenten Tests für die Erklärung der Bewertung von Mappus sind in der rechten Hälfte von Tabelle 6.22 dargestellt. Der Effekt der Angriffe wird von der Zugehörigkeit zum rot-grünen Lager moderiert. Verteidigungen haben konditionale Effekte in Abhängigkeit von den Voreinstellungen zu Mappus und zu Schmid. Schließlich ist die Aufnahme des einfachen Effekts der Einstellung zur CDU in das Modell notwendig, obwohl der kombinierte Effekt des Prädiktors auf Intercept und Slope nicht signifikant zur Modellverbesserung beiträgt. Dies erklärt sich durch die Logik des L1-Messmodells mit zwei Parametern: In diesem Fall ist der auf den Slope bezogene Koeffizient mit $t(165) = 1.95$ ($p = .053$) fast signifikant, der auf den Intercept bezogene Koeffizient trägt nicht signifikant zur Erklärung bei ($t(165) = 1.38$, $p = .169$) (vgl. Tabelle A.11). Würden wir den auf den Intercept bezogenen Koeffizienten aus dem Modell entfernen und somit den gesamten Effekt des Prädiktors auf den Slope richten, so würde dieser die konventionelle Signifikanzgrenze unterschreiten. Hiervon sehen wir jedoch ab, da es keine theoretisch plausible Begründung dafür gibt, warum die Einstellung zur CDU lediglich die Veränderung über die Zeit, nicht aber das geschätzte Ausgangsniveau der Bewertung beeinflussen sollte. Der Prädiktor kann jedoch auch nicht völlig aus dem Modell entfernt werden, da so der bestehende Einfluss auf den Slope, der zur Verbesserung der Modellgüte beiträgt, vernachlässigt würde. Die Zugehörigkeit zum Lager des Sprechers ist in den Modellen zur Erklärung der Bewertung von Mappus nicht mehr enthalten. Offenbar lässt sich diese Information besser im Model berücksichtigen, indem die Einstellung zum Kandidaten und in geringerem Ausmaß die Einstellung zu seiner Partei als Prädiktoren herangezogen werden.

Ziel der Aufnahme weiterer L2-Prädiktoren zu den Eigenschaften der Rezipienten ist es, größere Anteile der rezipientenbezogenen Varianzkomponente aufzuklären. Tabelle 6.23 fasst die Reduzierung der Varianzkomponenten für die Modelle beider Kandidaten zusammen. Angegeben ist die Varianzaufklärung gegenüber dem unkonditionalen Modell ohne Prädiktoren sowie die zusätzliche Varianzaufklärung gegenüber den Modellen, die nur den Faktor Lagerzugehörigkeit als Merkmal der Rezipienten berücksichtigen. Die Aufklärung der rezipientenbezogenen Komponente wird durch die Aufnahme der

Tabelle 6.23: Varianzerklärung der Modelle zur Erklärung der Bewertung von Schmid und Mappus während aller Antworten durch Relation und Voreinstellungen

	Schmid		Mappus	
	vs. M0	vs. M1	vs. M0	vs. M1
$R^2_{2\,\text{Intercept}\,|\,\text{Antworten}}$	−.07	.00	.15	−.03
$R^2_{2\,\text{Slope}\,|\,\text{Antworten}}$	−.02	.00	.19	−.01
$R^2_{2\,\text{Intercept}\,|\,\text{Rezipienten}}$	.17	.12	.31	.21
$R^2_{2\,\text{Slope}\,|\,\text{Rezipienten}}$	.29	.07	.37	.12
$R^2_{1\,\text{Intercept}\,|\,\text{Messmodelle}}$	.00	.00	.02	.01
$R^2_{1\,\text{Slope}\,|\,\text{Messmodelle}}$	.01	.00	.00	.00

Anmerkungen
vs. M0: Reduzierung der Varianzkomponenten im Vergleich zu den unkonditionalen Modellen (R^2_1 bzw. R^2_2).
vs. M1: Reduzierung der Varianzkomponenten im Vergleich zu den Modellen nur mit Lagerzugehörigkeit als Rezipientenmerkmal (ΔR^2_1 bzw. ΔR^2_2).

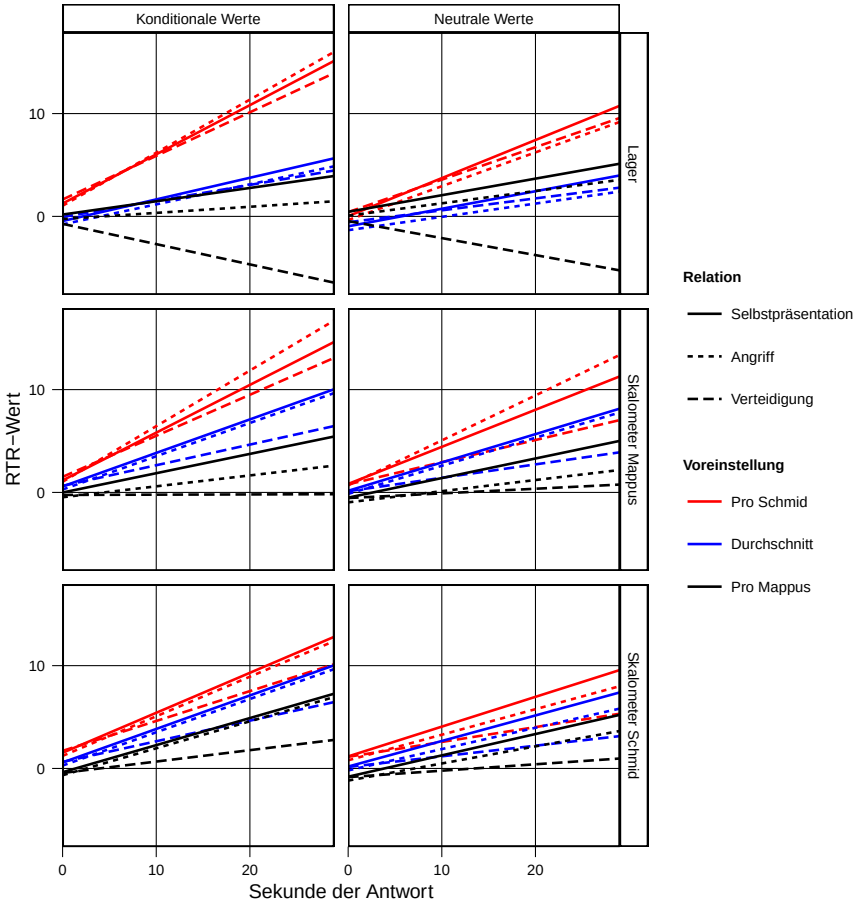
Voreinstellungen wesentlich verbessert. Im Modell zur Erklärung der unmittelbaren Urteile über Schmid erhöht sich der Anteil der erklärten Varianz in den rezipientenbezogenen Komponenten um zwölf bzw. sieben Prozentpunkte. Noch wesentlich hilfreicher sind die zusätzlichen Prädiktoren zur Erklärung der Bewertung von Mappus. Sie reduzieren die rezipientenbezogenen Varianzkomponenten um 21 bzw. zwölf weitere Prozentpunkte. Verantwortlich dafür ist in erster Linie die Skalometer-Variable zur Erfassung der Einstellung gegenüber Mappus. Für diesen Befund liegt eine fallspezifische Erklärung nahe: Die Person Mappus polarisierte sehr stark und war auch im eigenen Lager nicht unumstritten (z.B. Bachl, 2013b; Gabriel & Kornelius, 2011; Wehner, 2013). Durch die Aufnahme dieser Variable wird die weniger informative Dummy-Variable Zugehörigkeit zum schwarz-gelben Lager ersetzt, da die tatsächliche Einstellung zum Kandidaten für dessen Bewertung in der Debatte aussagekräftiger ist als die einfache Zuordnung zu seinem Lager.

Die Reduzierung der messmodellbezogenen Varianzkomponente fällt nicht bzw. nur in sehr geringem Maße besser aus. Die Schätzer der sehr kleinen antwortbezogenen Varianzanteile weisen wiederum teilweise kleine negative Beträge auf. Dies ist, wie oben erläutert, kein Anlass für Bedenken bezüglich der grundsätzlichen Qualität der Modelle. Dass bezüglich dieser Komponente keine

Modellverbesserungen festzustellen sind, ist zu erwarten, da mit den beiden Relationen dieselben Antwortmerkmale enthalten sind. Insgesamt können wir festhalten, dass die Aufnahme weiterer, auch quasi-metrisch skalierter Prädiktoren aus Sicht der erklärten Varianz lohnenswert ist. Wir erkaufen uns diese zusätzlich erklärte Varianz allerdings durch eine höhere Komplexität der Modelle, die sich bei Interpretation und Darstellung der Richtung der Effekte bemerkbar macht.

Für diese Interpretation nutzen wir ein weiteres Mal die durch die Modelle vorhergesagten Bewertungen der Kandidaten. In den folgenden Abbildungen sind diese Werte für die Ausprägungen jedes L2-Prädiktors unter der Annahme neutraler und typischer Werte für die Ausprägungen der anderen L2-Prädiktoren dargestellt. Zur Vereinheitlichung der Bezeichnungen werden die Ausprägungen der L2-Prädiktoren von Interesse in Bezug auf ihre Bedeutung für die Einstellung zu den beiden Kandidaten benannt. So bezeichnen wir die Ausprägung des Skalometers Schmid von einer Standardabweichung über dem Mittelwert als *Pro Schmid* und die Ausprägung von einer Standardabweichung unter dem Mittelwert als *Pro Mappus*. Umgekehrt heißt die um eine Standardabweichung überdurchschnittliche Ausprägung des Skalometer Mappus *Pro Mappus*, die um eine Standardabweichung unterdurchschnittliche Ausprägung *Pro Schmid*. Für den Faktor Lager entsprechen die Bezeichnungen der Zugehörigkeit zum Lager des Kandidaten, die Bezeichnung *Durchschnitt* entspricht den Unentschiedenen. In Abbildung 6.35 sind die vorhergesagten RTR-Bewertungen von Schmid in Abhängigkeit von den Voreinstellungen der Rezipienten und den Relationen der Antworten dargestellt.

Gut zu erkennen sind die Effekte der Rezipientenmerkmale. Wie erwartet ist die Voreinstellung zu Mappus negativ und die Voreinstellung zu Schmid positiv mit der Bewertung von Schmid während der Debatte assoziiert. Entsprechend führt eine Zugehörigkeit zum Lager des Sprechers zu einer besseren, die Zugehörigkeit zum Lager des Gegenkandidaten zu einer weniger guten Bewertung. Die unmittelbaren Urteile der Unentschiedenen sortieren sich zwischen den Lagern ein. Die gemeinsame Wirkung der Voreinstellungen lässt sich unter Annahme der konditionalen Werte gut nachvollziehen. In dieser Darstellung wird berücksichtigt, dass beispielsweise die Anhänger des rot-grünen Lagers im Mittel eine positivere Einstellung gegenüber Schmid und eine negativere Einstellung gegenüber Mappus haben. Bemerkenswert ist, dass selbst in dieser Betrachtungsweise die Bewertungen von Schmid auch unter der Annahme von für ihn negativen Ausprägungen in den L2-Prädiktoren nur in einer Konstellation eine negative Tendenz aufweisen. Während Schmids Antworten kommt es also kaum zu negativen Reaktionen des Publikums, sondern lediglich zu einer neutralen bis leicht positiven Bewertung.



Anmerkungen
Bewertung von Schmid auf einer Skala von -50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch das Modell in Tabelle A.10.
Vorhergesagte RTR-Werte beim Mittelwert (M) und ± 1 Standardabweichung (SD) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf einen neutralen Wert gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt.

Abbildung 6.35: Effekte der Relation und der Voreinstellungen auf die Bewertung von Schmid während seiner Antworten

Der konditionale Effekt der Verteidigungen in Abhängigkeit von der Zugehörigkeit zum Regierungslager ist in der ersten Zeile der Abbildung deutlich sichtbar. Für die Bewertung von Schmid durch seine eigenen Anhänger und die Unentschiedenen ist die Relation der Antwort unerheblich. Seine Verteidigungen haben jedoch einen deutlichen negativen Effekt auf die Anhänger des Regierungslagers. Die Bewertungen der Verteidigungen durch diese Rezipienten ist die einzige Konstellation der L2-Prädiktoren, die zu einer negativen Bewertung von Schmid führt. Wenn Schmid in die Defensive gerät und Positionen rechtfertigt, die durch kritische Fragen der Moderatoren bzw. Angriffe von Mappus angesprochen werden, reagieren die Rezipienten, die eine Wahl von CDU oder FDP beabsichtigen, mit negativen Bewertungen. Sie teilen die aufgebrachte Kritik und lassen sich auch von den Argumenten Schmids nicht überzeugen.

Der von der Voreinstellung zu Mappus moderierte Effekte der Angriffe ist im Vergleich dazu schwächer, lässt sich in der zweiten Zeile der Abbildung jedoch zumindest für die über- und unterdurchschnittlichen Ausprägungen der Rezipientenvariable recht gut erkennen. Für die Angriffe ist die Beziehung zwischen der Voreinstellung zu Mappus und der Bewertung von Schmid während des Duells noch enger als für die Selbstpräsentationen. Je schlechter ein Rezipient Mappus vor der Debatte beurteilt, desto besser werden die Angriffe Schmids auf den Gegenkandidaten (und sein Lager) während des Duells bewertet. Entsprechend führt eine positive Meinung von Mappus vor dem Duell zu einer schlechteren Bewertung der Angriffe. Dies zeigt sich in der Abbildung darin, dass Rezipienten mit einer unterdurchschnittlichen Einstellung zu Mappus den Sprecher Schmid während seiner Angriffe noch etwas besser bewerten als während seiner Selbstpräsentationen. Umgekehrt werden die Angriffe von Rezipienten mit einer überdurchschnittlichen Meinung von Mappus schlechter bewertet als die Selbstpräsentationen. Selbst für diese Ausprägung der Rezipientenvariable findet sich allerdings ein leicht ansteigender Slope der vorhergesagten RTR-Werte. Von einer beidseitigen Polarisierung durch die Angriffe, die eine Mobilisierung der Mappus-Gegner bei einer gleichzeitigen Reaktanz der Mappus-Anhänger verursacht, kann im hier untersuchten Ausschnitt der Antworten von Schmid also nicht die Rede sein.

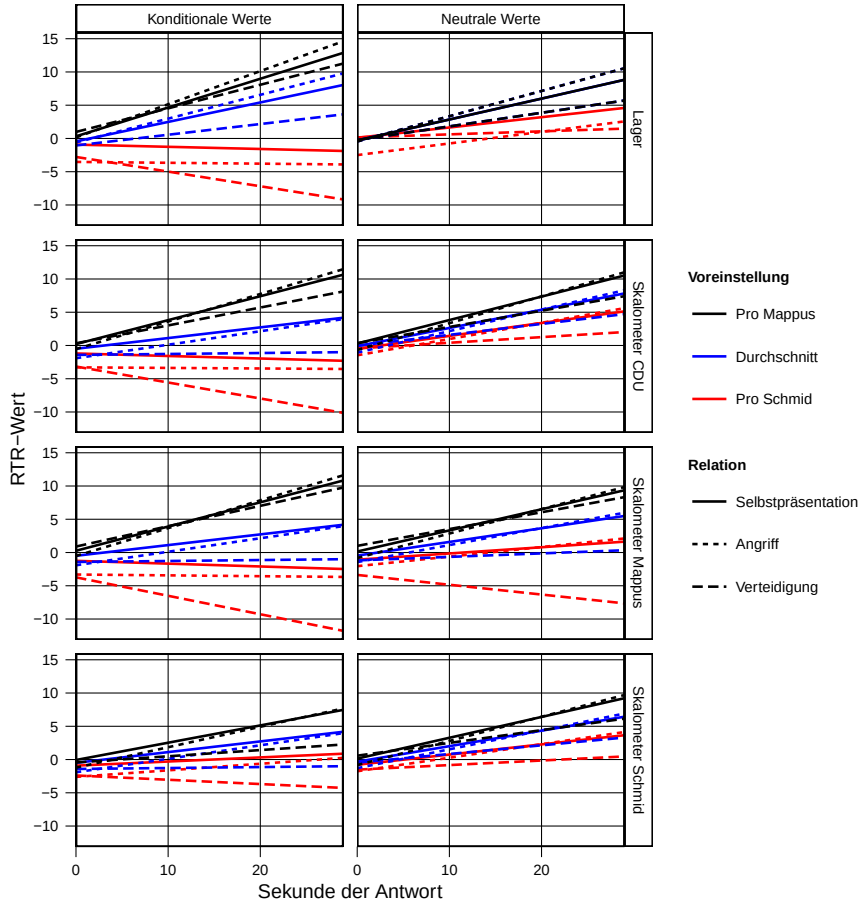
Der Befund, dass die Effekte der Relationen in Interaktion mit der Zugehörigkeit zum Lager Mappus und der Einstellung zu Mappus auftreten, lässt sich *ex post* plausibel mit den spezifischen Inhalten der Debatte erklären. Schmid muss sich vor allem in Bezug auf Themen verteidigen, in denen er Positionen vertritt, die im Kern den Ansichten des schwarz-gelben Lagers widersprechen. Herauszuheben sind hier seine Versuche, das Konzept des längeren gemeinsamen Lernens gegen die Angriffe von Mappus zu verteidigen (Bachl, Käfferlein &

Spieker, 2013b; Bachl & Vögele, 2013). Ein größerer Teil seiner Angriffe richtet sich direkt gegen die Rolle von Mappus bei der Verlängerung der Laufzeiten für Atomkraftwerke und beim Kauf der EnBW-Aktienmehrheit (Bachl, Käfferlein & Spieker, 2013b). Es liegt nahe, dass für die Bewertung dieser Angriffe vor allem die Einstellung der Rezipienten gegenüber dem Kandidaten Mappus und weniger die Einstellung gegenüber seiner Partei oder die Wahlabsicht für CDU oder FDP relevant ist.

Schließlich können wir weiter im Detail zwischen Effekten, die sich im latenten Intercept oder im geschätzten Slope niederschlagen, unterscheiden. In den Vorhersagen des Intercepts finden sich lediglich einfache Effekte der Voreinstellungen, jedoch keine Effekte der Interaktionen von Rezipienten- und Antwortmerkmalen. Die Verteilung der Effekte auf Intercept und Slope kann dabei helfen, den Urteilsprozess der Rezipienten besser zu verstehen. Sie deutet darauf hin, dass die Voreinstellungen bzw. die Lagerzugehörigkeit zunächst als einfache Cues dienen, die eine erste Bewertung von Schmid steuern, relativ unabhängig von dem, was er in der jeweiligen Antwort sagt (in dieser Analyse natürlich streng genommen nur auf die Relation seiner Antworten, nicht auf die anderen Inhalte bezogen). Entsprechend sind die vorhergesagten RTR-Werte zu Beginn der Antwort nach den Ausprägungen der jeweiligen Rezipientenmerkmale geordnet, ein größerer Unterschied nach den Relationen lässt sich nicht ausmachen. Im weiteren Zeitverlauf werden die Inhalte der Antworten deutlich, und die (ebenfalls von den Voreinstellungen beeinflusste) Verarbeitung der Inhalte verursacht unterschiedliche Bewertungen in Abhängigkeit von Selbstpräsentationen, Angriffen und Verteidigungen. Da die Veränderungen der unmittelbaren RTR-Bewertungen erst später erfolgen, zeigen sie sich in den auf den Slope bezogenen Koeffizienten (vgl. im Detail auch die Koeffizienten in Tabelle A.10).

Abbildung 6.36 zeigt die vorhergesagten RTR-Werte der Modelle zur Erklärung der Bewertung von Mappus. Der einfache Effekt der Voreinstellung zur CDU fällt sehr schwach aus und führt für die in Abbildung 6.36 gewählten typischen Ausprägungen des Prädiktors nur zu einer sehr geringen Differenzierung zwischen den vorhergesagten RTR-Verläufen. Berücksichtigen wir zusätzlich die konditionale Verteilung der Ausprägungen der anderen L2-Prädiktoren (linke Spalte der Abbildung), so treten die Unterschiede deutlicher hervor. Im Überblick über alle Facetten der Abbildung wird aber auch klar, dass nur ein geringer Teil dieser Unterschiede exklusiv auf den Prädiktor Skalometer CDU zurückgeführt werden kann.

Unter den konditionalen Effekten der Relationen sticht vor allem die Interaktion zwischen der Voreinstellung zum Sprecher und den Verteidigungen heraus. Sie hat einen deutlich an den vorhergesagten RTR-Verläufen ablesbaren



Anmerkungen
Bewertung von Mappus auf einer Skala von -50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle A.11.
Vorhergesagte RTR-Werte beim Mittelwert (M) und ± 1 Standardabweichung (SD) des L2-Prädiktors. Facetten: *Neutrale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf einen neutralen Wert gesetzt. *Konditionale Werte*: Die Werte aller anderen L2-Prädiktoren sind auf die für die Ausprägungen M und $M \pm 1SD$ des dargestellten L2-Prädiktors typischen Werte gesetzt.

Abbildung 6.36: Effekte der Relation und der Voreinstellungen auf die Bewertung von Mappus während seiner Antworten

Effekt. Für die Bewertung durch Rezipienten, die Mappus vor der Debatte überdurchschnittlich positiv gegenüberstehen, macht die Relation der Antwort keinen Unterschied. Verteidigungen führen wie auch Selbstpräsentationen oder Angriffe zu einem Anstieg der unmittelbaren RTR-Bewertungen im Zeitverlauf. Unter der Annahme einer durchschnittlichen Bewertung von Mappus vor dem Duell zeigt sich im vorhergesagten Ausgangswert des RTR-Verlaufs kein wesentlicher Unterschied. Der Slope der Verteidigungen, der die Veränderung der Bewertung über die Zeit quantifiziert, weist jedoch im Unterschied zu den Selbstpräsentationen und Angriffen keine Steigung auf. Die Bewertung von Verteidigungen wird also im Gegensatz zu den übrigen Relationen nicht besser. Für Rezipienten, die Mappus gegenüber negativ eingestellt sind, ergeben sich im Vergleich zur Referenz-Relation der Selbstpräsentationen in beiden Parametern der Wachstumskurve signifikante Unterschiede: Die vorhergesagten RTR-Werte starten im Gegensatz zu den Selbstpräsentationen bereits im negativen Bereich, und die Bewertungen werden im Zeitverlauf negativer. Inhaltlich lässt sich dieses Effektmuster gut interpretieren: Verteidigungen werden häufig bereits durch die vorangehenden Inhalte – eine kritische Frage der Moderatoren oder einen Angriff des Gegenkandidaten – eingeleitet. Es ist also zu Beginn der Antwort bereits klar, dass nun ein Punkt besprochen wird, der Anlass für Kritik am Sprecher bietet. Greift Mappus dieses Thema nun auf, um sich zu verteidigen, reagieren die Personen, die Mappus ohnehin (sehr) kritisch sehen, schnell und stark negativ, was zu negativeren Intercept-Schätzungen führt. Auch im weiteren Verlauf der Verteidigungen werden die geschätzten unmittelbaren Bewertungen negativer, je schlechter die Voreinstellung zum Kandidaten ausgeprägt ist.

Der Effekt der Angriffe wird von der Zugehörigkeit zum rot-grünen Lager moderiert. Rezipienten, die vor dem Duell nicht beabsichtigen, SPD oder Grüne zu wählen, bewerten die Angriffe ähnlich wie die Selbstpräsentationen. Auch für die Angriffe findet sich eine positive Veränderung der RTR-Verläufe. Die Anhänger des Oppositionslagers reagieren dagegen, wenn ihre Parteien bzw. ihr Kandidat von Mappus angegriffen wird, negativ. Die negativen Reaktionen auf die Angriffe treten offenbar relativ früh und heftig während der Antworten auf. Der geschätzte Intercept der Verläufe liegt für die Angriffe deutlich unter dem der Selbstpräsentationen. Ein weiteres Absinken kann dann nicht festgestellt werden, die Verläufe zu den Angriffen bleiben wie auch die der Selbstpräsentationen konstant. Dass der Effekt der Angriffe in Interaktion mit der Zugehörigkeit zum Oppositionslager auftritt, können wir im Nachhinein gut erklären. Die meisten hier untersuchten Angriffe von Mappus richten sich nicht direkt gegen den Kandidaten Schmid, sondern gegen politische Vorschläge, die Mappus verallgemeinernd dem rot-grünen Lager unterstellt. In-

sofern erscheint es plausibel, dass die Angehörigen dieses Lagers eine negative Reaktion auf die Angriffe zeigen.

Ein weiterer konditionaler Effekt der Verteidigungen besteht in Abhängigkeit von der Einstellung zu Schmid. Er zeigt sich in der erwarteten Richtung: Während der Verteidigungen wird Mappus von den Anhängern der Opposition bzw. von Rezipienten, die ihm gegenüber negativ eingestellt sind, negativer bewertet als während der Selbstpräsentationen. Der Effekt ist statistisch signifikant und auch in den Darstellungen der vorhergesagten RTR-Verläufe leicht zu erkennen. Er fällt allerdings um einiges schwächer aus als der konditionale Effekt der Verteidigungen in Abhängigkeit der Einstellung zu Mappus. Der Befund weist darauf hin, dass für die Bewertungen der untersuchten Verteidigungen die Einstellung zum sich verteidigenden Kandidaten Mappus wichtiger ist als die Zugehörigkeit zum die Kritik teilenden Lager bzw. dessen Repräsentanten in der Debatte.

Schließlich fällt beim Vergleich der Facetten der vorhergesagten RTR-Verläufe auf Basis von konditionalen Ausprägungen der jeweils anderen L2-Prädiktoren eine große Ähnlichkeit auf. Sie weist auf die Kumulation der gleichgerichteten Einzeleffekte bei der Erklärung der beobachteten RTR-Verläufe hin. Statistisch zeigt sich dies als Problem der Multikollinearität der Rezipientenmerkmale. Sie erschwert die Trennung der Effekte der einzelnen Prädiktoren und führt in Teilen auch zu Problemen bei der numerischen Evaluation der (Residual-) Likelihood bei der Modellschätzung. Hierbei handelt es sich im Grunde jedoch nicht um ein statistisches Problem, sondern um ein inhaltliches Phänomen. Es ist höchst plausibel, dass die politischen Voreinstellungen zu einem gewissen Maß miteinander korrelieren und dass sich ihre Effekte auf dieselben Varianzanteile der unmittelbaren Bewertung des Kandidaten beziehen. Dementsprechend ist die Erklärungsleistung dieser Modelle hoch (vgl. Tabelle 6.23). Die Effekte der einzelnen Prädiktoren sind jedoch nur schwer zu trennen, was nicht verwunderlich ist, da sie zu einem großen Teil gemeinsam den latenten Einfluss einer allgemeinen politischen Prädisposition der Rezipienten auf die Verarbeitung und unmittelbare Bewertung der Kandidatenaussagen abbilden. Die Präzision der einzelnen Schätzer sollte daher nicht überinterpretiert werden. Wichtiger ist die Erkenntnis, dass die Voreinstellungen und auch die Relationen in Interaktionen mit diesen einen wichtigen Beitrag zur Erklärung der unmittelbaren Bewertungen der Kandidaten während ihrer Antworten leisten.

6.3.3 Bewertung der Kandidaten nach Relationswechseln

Im ersten Anwendungsbeispiel der kreuzklassifizierten Wachstumskurvenmodelle haben wir uns auf die Erklärung der unmittelbaren Kandidaten-

bewertungen während der direkten Antworten der Kandidaten auf Fragen der Moderatoren beschränkt. Die Antworten als Analyseeinheit des Debatteninhalts haben den Vorteil, dass die Veränderungen in den individuellen RTR-Messungen gut untersucht werden können, da die meisten Verläufe nahe des neutralen Skalenmittelpunkts starten. Dies hat jedoch den Nachteil, dass wir einige Teile der Bewertungen, die sich auf Aussagen zu einem späteren Zeitpunkt des Turns oder Aussagen, die nicht auf die Frage eines Moderators folgen, außen vor lassen. Diesen Nachteil wollen wir nun beheben, indem wir die Veränderung der unmittelbaren Bewertungen als Folge eines Wechsels der Relation untersuchen.

Technisch ist diese Analyseeinheit für die Inhaltsebene recht einfach definiert: Immer, wenn sich die Relation während der Aussage eines Kandidaten ändert, beginnt eine neue Analyseeinheit. Vom Zeitpunkt, an dem ein Relationswechsel stattfindet, untersuchen wir, wie sich die Bewertungen in den nächsten zehn Sekunden ändern. Damit tragen wir dem Umstand Rechnung, dass die individuellen Reaktionen auf die Relationswechsel in einem gewissen Zeitfenster auftreten können und dass wir ohnehin nicht davon ausgehen, einen Relationswechsel in jedem Fall auf die exakte Sekunde genau in der Inhaltsanalyse bestimmen zu können. Um eine formale Standardisierung der Analyseeinheiten sicherzustellen, betrachten wir nur Relationswechsel, nach denen die neue Relation für mindestens zehn Sekunden beibehalten wird.[79]

Während es technisch einfach ist, die Informationen zu Rezipienten, Inhalten und RTR-Bewertungen in diese Datenstruktur zu transformieren und sie mit kreuzklassifizierten Wachstumskurvenmodellen zu analysieren, so formulieren wir durch diese Wahl der Analyseeinheit zumindest implizit sehr starke Prämissen. Wir setzen zum einen voraus, dass der Relationswechsel nicht nur ein konzeptionelles Konstrukt der Forscher bei der Erstellung des Codebuchs für die Inhaltsanalyse ist, sondern auch als Konstrukt der Rezipienten existiert und für die Abgabe ihrer Bewertungen potenziell relevant ist. Nur dann können wir davon ausgehen, dass ein Relationswechsel auch eine Veränderung in der unmittelbaren Bewertung der Kandidaten haben kann. Diese Prämisse ist nicht mit der Annahme zu verwechseln, dass die Verwendung einer bestimmten Relation generell die Bewertung der Kandidaten beeinflusst, wie wir es am Beispiel der Relationen der Antworten gezeigt haben. Da wir nun Veränderungen untersuchen, die auf den Wechsel der Relation folgen, müssen diese Veränderungen auch im definierten Zeitraum (hier: zehn Sekunden) nach dem Wechsel auftreten. Damit ist dieses Modell zwar etwas flexibler als die These der sekundengenauen Transferfunktion bei Nagel (2012), formuliert jedoch

79 Vergleiche dazu ausführlich die Erläuterung der Analyseeinheiten auf S. 173.

trotzdem recht hohe formale Erwartungen an das Erkennen der durch die Forscher definierten Merkmale und die zeitliche Struktur der Effekte. Zum anderen muss die Inhaltsanalyse das Rezipientenkonstrukt des Relationswechsels – sofern es überhaupt existiert – valide erfassen, das heißt, das Codebuch muss wirklich das beschreiben, was auch für die Rezipienten als Wechsel einer Relation erkennbar ist. Diese Definition muss dann ebenfalls valide von den Codierern auf den Debatteninhalt angewendet werden. Dabei ist nicht nur die inhaltliche Validität wichtig, sondern auch die zeitliche Präzision im Sinne des Nachvollziehens des Rezeptionsprozesses.

Im Folgenden führen wir die Analysen zum Einfluss der Rezipientenmerkmale und der Relation auf die Veränderung der Bewertung der Kandidaten bei einem Relationswechsel wieder getrennt für beide Kandidaten durch. Als Merkmal der Rezipienten ziehen wir den Faktor Lagerzugehörigkeit heran. Merkmal des Relationswechsels ist die Ausprägung der neuen Relation. Im kreuzklassifizierten Datensatz sind insgesamt 212325 RTR-Messungen enthalten, die in 21335 Messmodellen, also Kombinationen von Rezipienten und Relationswechseln, zusammengefasst sind. Insgesamt analysieren wir die Reaktionen auf 122 Relationswechsel. Von den 61 Relationswechsel von Schmid sind 34 Selbstpräsentationen, 13 Angriffe und sieben negative Lagebeschreibungen.[80] Die übrigen Ausprägungen (darunter drei Verteidigungen) der Kategorie kommen nur selten vor und werden gemeinsam mit der Ausprägung Selbstpräsentation zur Referenzausprägung zusammengefasst. Diese bezeichnen wir der Einfachheit halber weiterhin als Selbstpräsentation. Unter den 61 Relationswechseln von Mappus sind 30 Selbstpräsentationen, 13 Verteidigungen und 15 Angriffe. Wieder dienen die Selbstpräsentationen gemeinsam mit den drei weiteren Relationswechseln als Referenzausprägung.

Die RTR-Messungen zu den Relationswechseln stammen von insgesamt 176 Rezipienten. Von einem Rezipienten liegen im Mittel zu 121 Relationswechseln RTR-Messungen vor ($Min = 110$, $Max = 122$, $SD = 2$). Umgekehrt sind zu einem Relationswechsel durchschnittlich von 175 Rezipienten RTR-Messungen vorhanden ($Min = 171$, $Max = 176$, $SD = 1$). Damit ist der Datensatz fast balanciert, das heißt, von einem Großteil der Rezipienten liegen zu jedem Relationswechsel RTR-Messungen vor.

Als Modellspezifikation wählen wir wie im ersten Anwendungsbeispiel ein Messmodell mit Intercept und einem linearen Slope. Mit diesem Modell können wir im Intercept den (geschätzten) Ausgangswert der RTR-Verläufe zu Beginn des Relationswechsels und im Slope die (geschätzte) Veränderung

80 Negative Lagebeschreibungen sind Aussagen, in denen die Lage in einem Politikfeld negativ beschrieben wird, ohne dass hierfür direkt ein Verantwortlicher benannt wird (vgl. ausführlich Bachl, Käfferlein & Spieker, 2013a, 2013b).

infolge des Relationswechsels bestimmen. Inhaltlich trägt der Intercept dem Umstand Rechnung, dass zum Zeitpunkt, an dem ein Wechsel der Relation auftritt, bereits eine RTR-Bewertung vorliegt, die ihre Ursache in den zuvor geäußerten Inhalten der Debatte hat. Von Interesse für unsere Forschungsfrage, ob der Wechsel der Relation zu einer Veränderung der Bewertung führt, die dann durch die Lagerzugehörigkeit der Rezipienten und die neue Relation erklärt werden kann, ist der Slope. Wenn sich ein auf den Slope bezogener Koeffizient signifikant von Null unterscheidet, so liegt ein systematischer Einfluss dieses Merkmals bzw. dieser Merkmalskombination auf die durch den Relationswechsel bedingte Veränderung vor.

Varianzdekomposition Wie im ersten Anwendungsbeispiel können wir zuerst das Verhältnis der erklärbaren Varianzkomponenten eines unkonditionalen Modells ohne L2-Prädiktoren bestimmen. Dabei unterscheiden wir zwischen den Intercept-bezogenen Varianzkomponenten der geschätzten Ausgangswerte und den Slope-bezogenen Komponenten der geschätzten Veränderungen. Die Varianzdekomposition ergibt für die Modelle zur Erklärung der Veränderung der Bewertung beider Kandidaten ein sehr ähnliches Muster. Im geschätzten Ausgangswert zu Beginn der neuen Relation liegt der größte Varianzanteil in den Messmodellen, also in den Cross-Level-Interaktionen und Messfehlern (Schmid: 68%; Mappus: 70%). Auf die rezipientenbezogene Komponente entfallen in beiden Modellen 22 Prozent der Varianz, auf die Komponente der Relationswechsel zehn bzw. sieben Prozent. Um die RTR-Bewertung der Kandidaten zu Beginn jedes Relationswechsels zu erklären, sind die Eigenschaften der Rezipienten damit wiederum wichtiger als die Eigenschaften der Relationswechsel. Die Varianz der Veränderungen nach dem Wechsel der Relation lässt sich fast vollständig den Messmodellen zuordnen (Schmid: 91%; Mappus: 89%). Hier deutet sich bereits an, dass die Veränderungen in hohem Maße situationsspezifisch sind und sich nicht einfach durch die Merkmale der Rezipienten (Schmid: 4%; Mappus: 8%) oder der Relationswechsel (Schmid: 6%; Mappus: 3%) erklären lassen.[81] Die Varianzdekomposition der Veränderung der unmittelbaren Kandidatenbewertungen kündigt bereits an, dass die Identifikation von L2-Prädiktoren der Rezipienten und/oder Relationswechsel, die einen systematischen Einfluss auf die Veränderung der Bewertung infolge der Relationswechsel haben, wenig erfolgsversprechend ist.

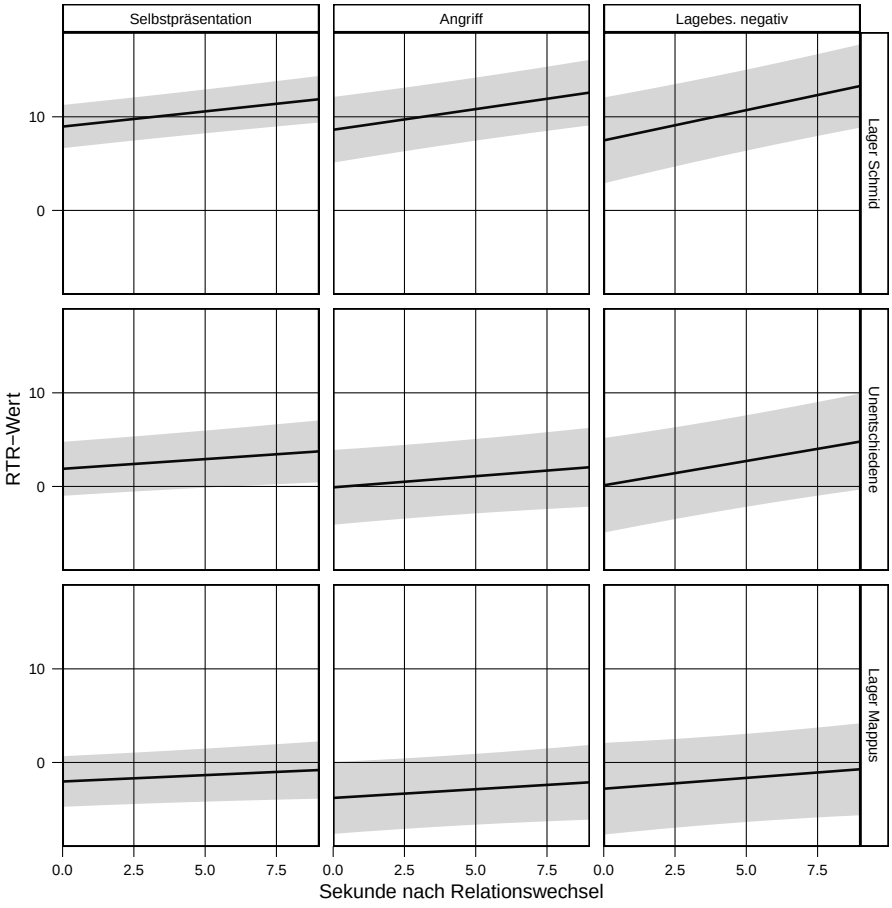
Effekte der Relation und der Lagerzugehörigkeit Dies bestätigt sich bei der Analyse der Fixed Effects der Modelle, in denen die RTR-Verläufe durch

81 Abweichungen von 100 Prozent ergeben sich durch Rundungsfehler.

die Lagerzugehörigkeit, die Relation und deren Interaktion erklärt werden. Abbildung 6.37 zeigt die durch das Modell vorhergesagten Verläufe der unmittelbaren Bewertungen von Schmid. Bezüglich des Ausgangsniveaus der Verläufe erkennen wir deutlich den typischen Effekt der Lagerzugehörigkeit. Beim Wechsel einer Relation ist die mittlere Bewertung Schmids durch die Anhänger des rot-grünen Lagers positiver als die Bewertung durch die Referenzausprägung der Unentschiedenen. Die unmittelbaren Bewertungen des schwarz-gelben Lagers liegen leicht unterhalb der Referenzgruppe. Für das geschätzte Ausgangsniveau der RTR-Verläufe macht die folgende Relation keinen Unterschied. Es ist also bei den vorliegenden Bewertungen Schmids nicht so, dass der Wechsel auf einen Angriff oder eine negative Lageschreibung von einem anderen Ausgangswert startet und dadurch die möglichen Reaktionen durch die Grenzen der RTR-Skala eingeschränkt wären. Unter den auf die Veränderung der Bewertungen bezogenen Slope-Koeffizienten ist nur die einfache Veränderung über die Zeit signifikant. Dieser Effekt ist auch in der Abbildung gut zu erkennen: Die RTR-Verläufe steigen in den zehn auf einen Relationswechsel folgenden Sekunden im Durchschnitt leicht an. Die Erwartung, dass diese Veränderung von der Lagerzugehörigkeit der Rezipienten oder der neuen Relation beeinflusst wird, erfüllt sich nicht. Alle Koeffizienten der Interaktionen der Zeit mit den L2-Prädiktoren sind nicht signifikant (vgl. Tabelle A.12).

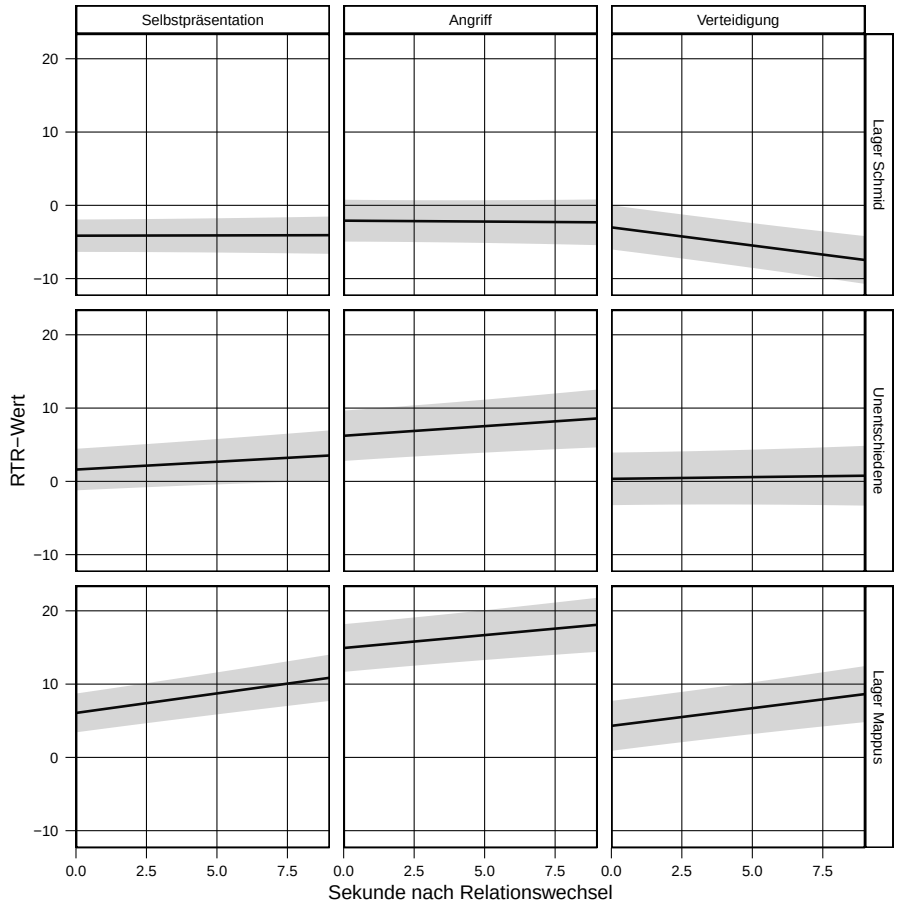
Sehr ähnlich ist das substantielle Ergebnis des Modells zur Erklärung der Veränderung der Bewertung von Mappus (vgl. Abbildung 6.38, Tabelle A.13). Zwar tragen einige Prädiktoren zur Erklärung der geschätzten Ausgangspunkte der Verläufe bei. So liegt dieser Punkt für die Selbstpräsentationen beim Regierungslager höher und beim Oppositionslager niedriger als bei den Unentschiedenen. Die Bewertung der Angriffe startet bei den Unentschiedenen auf einem höheren Niveau als die Bewertung der Selbstpräsentationen. Die Oppositionsanhänger liegen auch bei dieser Relation unter dem Niveau der Unentschiedenen, die Regierungsanhänger über diesem Niveau. Schließlich ist die Bewertung beim Wechsel auf eine Verteidigung im rot-grünen Lager positiver als bei den Unentschiedenen.

Die Veränderungen im Zeitverlauf werden von der Lagerzugehörigkeit beeinflusst. Die Bewertungen der Selbstpräsentationen und Verteidigungen durch die Angehörigen des schwarz-gelben Lagers steigen stärker an als die Bewertungen durch die Unentschiedenen. Ebenso ist der Steigungskoeffizient der Bewertungen der Selbstpräsentationen signifikant geringer als der Koeffizient der Unentschiedenen, was im Resultat zur Vorhersage eines konstanten Verlaufs führt. Nur zwei signifikante Effekte zeigen sich jedoch für die Interaktionen der Veränderung mit der Lagerzugehörigkeit und der Relation. Der



Anmerkungen
Bewertung von Schmid auf einer Skala von −50 (größter Nachteil Schmid) bis 50 (größter Vorteil Schmid). Vorhersage durch das Modell in Tabelle A.12. Die Flächen zeigen 95%-Konfidenzintervalle. Die Ausprägung Selbstpräsentation der Relation enthält auch die sonstigen Relationen.

Abbildung 6.37: Effekte der Relation und der Lagerzugehörigkeit auf die Veränderung der Bewertung von Schmid nach Relationswechseln



Anmerkungen
Bewertung von Mappus auf einer Skala von −50 (größter Nachteil Mappus) bis 50 (größter Vorteil Mappus). Vorhersage durch das Modell in Tabelle A.13. Die Flächen zeigen 95%-Konfidenzintervalle.
Die Ausprägung Selbstpräsentation der Relation enthält auch die sonstigen Relationen.

Abbildung 6.38: Effekte der Relation und der Lagerzugehörigkeit auf die Veränderung der Bewertung von Mappus nach Relationswechseln

Koeffizient der Veränderung der Bewertung der Angriffe durch die Anhänger von Mappus ist mit einem Betrag von $\beta = -0.23$ signifikant negativ. Dieses zunächst kontraintuitive Ergebnis klärt sich mit einem Blick auf Abbildung 6.38 auf. Bereits zu Beginn der Angriffe ist die Bewertung von Mappus durch diese Rezipienten sehr positiv. Von diesem positiven Ausgangswert folgt dann allerdings eine im Vergleich zu den Selbstpräsentationen geringere Steigung, die in der Verrechnung der Koeffizienten jedoch noch immer positiv ist. Der einzige erwartungskonforme Befund zum Zusammenspiel von Lagerzugehörigkeit und neuer Relation zeigt sich bei der Veränderung der Bewertung von den Verteidigungen von Mappus durch die Anhänger der Opposition. Wenn Mappus sich verteidigt, dann wird er durch diese Rezipienten negativer bewertet.

6.3.4 Zwischenfazit zu den kreuzklassifizierten Wachstumskurvenmodellen

In diesem Teilkapitel haben wir kreuzklassifizierte Wachstumskurvenmodelle auf die in zwei Analyseeinheiten des Debatteninhalts gruppierten unmittelbaren Kandidatenbewertungen angewendet. Während sich das analytische Vorgehen sich für die Einheit der direkten Antworten auf Fragen der Moderatoren bewährt, erweist es sich für die Einheit der Relationswechsel als weniger ertragreich. Im folgenden Abschnitt fassen wir die wichtigsten Befunde zu den kreuzklassifizierten Wachstumsmodellen zusammen, bevor wir in Kapitel 6.4 allgemeiner auf Limitationen und Potenziale der Mehrebenenmodelle für die Analyse unmittelbarer Kandidatenbewertungen eingehen.

Unmittelbare Kandidatenbewertungen nach Wechseln der Relation Insgesamt sind die Befunde, was ihre Erklärung der *Veränderung* der Bewertungen infolge von Relationswechseln angeht, beschränkt. Für Schmid findet sich ein schwacher, mehr oder minder unkonditionaler Anstieg der RTR-Verläufe nach den Relationswechseln. Die Ergebnisse für Mappus sind zwar etwas differenzierter, außer dem negativen Effekt des Wechsels zu einer Verteidigung auf die Veränderung der Bewertung durch die Anhänger des gegnerischen Lagers sprechen sie jedoch ebenfalls kaum für eine substantielle Bedeutung des Relationswechsels für die unmittelbaren Kandidatenbewertungen.

Zwei Erklärungen für das Ausbleiben der erwarteten Effekte auf die Veränderung der Bewertungen liegen nahe. Zunächst ist es möglich, dass die Relation für die Bewertung der Kandidaten auf das gesamte Duell hin betrachtet schlicht keine Bedeutung hat. Die Rezipienten erkennen zwar, dass ein Kandidat in seiner Aussage nun einen anderen Bezug herstellt. Dies ist für sie jedoch kein

relevanter Grund, ihre Bewertung des Kandidaten zu verändern. Für diese Erklärung spricht, dass auch in Studien zu anderen TV-Debatten keine oder nur sehr schwache Effekte der Relationen identifiziert wurden (z.B. J. Maier, 2007; Spieker, 2011; Strömbäck et al., 2009). Dagegen spricht allerdings, dass wir im ersten Anwendungsbeispiel durchaus konditionale Effekte der Relationen auf die Bewertung der Kandidaten während der Antworten nachweisen können. Dieses Argument wiegt unserer Ansicht schwerer, da sich die erste Analyse zwar nur auf bestimmte Ausschnitte der Debatte bezieht, diese sich jedoch zumindest teilweise mit den im zweiten Beispiel untersuchten Abschnitten decken. Zudem werden die Debatteninhalte von denselben Kandidaten präsentiert, und die Kandidaten werden von denselben Rezipienten bewertet. Es erscheint unwahrscheinlich, dass es möglich wäre, konditionale Effekte der Relationen auf die Bewertung der Kandidaten während der Antworten nachzuweisen, wenn die Relationen tatsächlich auf das gesamte Duell betrachtet keinerlei direkten oder konditionalen Effekt hätten.

Wenn wir also in Übereinstimmung mit den Befunden des ersten Anwendungsbeispiels die Hypothese aufrecht erhalten, dass die Relationen in Interaktion mit den Voreinstellungen der Rezipienten einen Einfluss auf die unmittelbare Kandidatenbewertung haben, dann bleibt die Erklärung, dass die Prämissen, die wir mit der vorliegenden Modellierung der Veränderungen gesetzt haben, so nicht zutreffen. Unsere Annahmen, dass erstens die Rezipienten einen Wechsel der Relation wahrnehmen und innerhalb von zehn Sekunden auf die neue Relation mit der Veränderung ihrer unmittelbaren Kandidatenbewertung reagieren und dass wir zweitens diese Relationswechsel inhaltsanalytisch hinreichend valide bezüglich der Verarbeitung der Debatteninhalte und hinreichend präzise bezüglich der zeitlichen Verortung des Relationswechsels erfassen können, sind demnach nicht haltbar.

Damit erweist sich der Analyseansatz, mit dem die deduktive Untersuchung der Effekte von Rezipientenmerkmalen und Relationen über die gesamte TV-Debatte hinweg erfolgen sollte, als wenig ertragreich. Voraussetzung für eine solche Analyse ist ein explizites Modell, das beschreibt, in welcher zeitlichen Struktur das Auftreten bestimmter Merkmale des Debatteninhalts zu einer messbaren Wirkung in den unmittelbaren Bewertungen der Kandidaten führt, und das eine Variabilität dieser Struktur über die Individuen und über den Verlauf des Duells hinweg erlaubt. Ein solches Modell bezieht sich nicht nur auf die Formulierung der statistischen Transferfunktion, sondern muss auch die valide und präzise inhaltsanalytische Identifikation der Merkmale leiten.

Es liegt außerhalb der Möglichkeiten dieser Arbeit, zu ergründen, ob die durchgeführte Inhaltsanalyse diesen Ansprüchen genügt. Zwar wurde der Versuch unternommen, die Codierung entlang des Rezeptionsprozesses zu

strukturieren (Bachl, Käfferlein & Spieker, 2013a). Ob dies vollständig gelungen ist, kann in Anbetracht der vorliegenden Ergebnisse jedoch bezweifelt werden. Weitere Entwicklungen zu einem rezeptionsgeleiteten Codiervorgehen erscheinen notwendig, wenn das Ziel ein einer vollumfänglichen Berücksichtigung der Debatteninhalte und ihrer Bewertungen Ziel der Analyse ist (vgl. auch Nagel, 2012).

Für die Entwicklung eines solchen Codierverfahrens, aber noch wichtiger für die Formulierung eines statistischen Modells zur Verknüpfung von Merkmalen des Debatteninhalts mit den RTR-Messungen auf Individualebene über den gesamten Duellverlauf hinweg, müssten präzisere explizite Vorstellungen von den Prozessen der Verarbeitung der Debatteninhalte und deren Übertragung in die beobachteten RTR-Messungen entwickelt werden. Das hier angewandte Modell eines Reaktionsfensters von zehn Sekunden nach dem inhaltsanalytisch (sicherlich nicht perfekt) erfassten Wechsel der Relation kann lediglich der Versuch einer ersten pragmatischen Annäherung sein. Immerhin erlaubt die Logik des kreuzklassifizierten Wachstumskurvenmodells innerhalb dieses Fensters intra- und interindividuelle Schwankungen und setzt damit keine statische Transferfunktion fest. Unsere Interpretation der vorliegenden Befunde, dass die gesetzten Prämissen in dieser Analyse nicht erfüllt werden, heißt natürlich nicht, dass die Formulierung eines derartigen Modells grundsätzlich nicht möglich ist. Eine Überarbeitung oder Neuformulierung eines solchen Modells macht jedoch in großem Umfang detaillierte theoretische und empirische Vorarbeiten nötig. Die Lösung kann es nicht sein, in einem empirischen Testprozess anhand eines Datensatzes so lange an den vielen möglichen „Stellschrauben" eines solchen Modells zu drehen, bis sich ein Nachweis der gewünschten Effekte einstellt. Sollte ein derart modellgenerierendes Vorgehen gewählt werden, befänden wir uns wieder im Bereich der induktiven, explorativen Datenanalyse. Es wäre dann anschließend die Replikation des Modells anhand weiterer Datensätze geboten, um ein an den einen Testdatensatz überangepasstes Modell zu vermeiden.

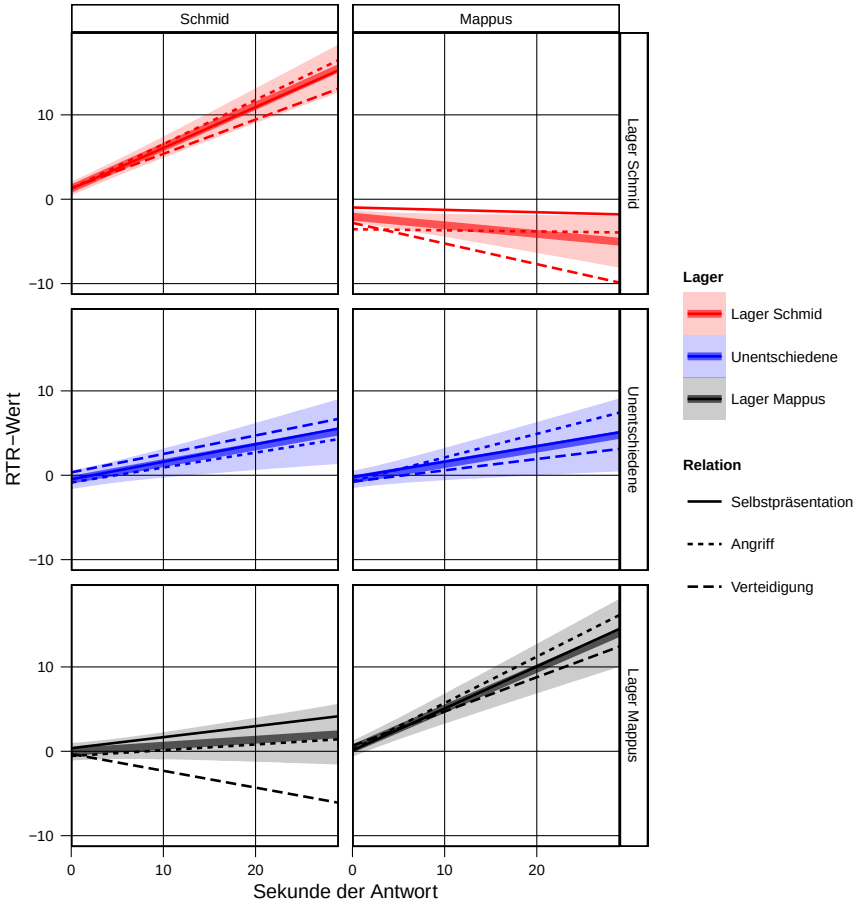
Nach unseren eingehenden Erfahrungen mit den komplexen Datenstrukturen, die sich bei der Analyse rezeptionsbegleitend erfasster Kandidatenbewertungen auf Individualebene über den gesamten Debattenverlauf hinweg ergeben, erscheint es uns zumindest zweifelhaft, ob ein solches konzeptionelles Modell überhaupt in einer Form formuliert werden kann, die sich in ein praktikables statistisches Modell übersetzten lässt. Zumindest nach dem bisherigen Kenntnisstand ist es hilfreicher, Ausschnitte aus den Debatten heranzuziehen, die untereinander eine strukturelle Ähnlichkeit und Standardisierung aufweisen und denen die Merkmale des Debatteninhalts auf einer höheren Abstraktionsebene – und damit eben nicht sekundengenau – zugeordnet werden

können. Die im ersten Anwendungsbeispiel untersuchte Einheit der direkten Antworten auf die Fragen der Moderatoren kann hierfür als geeignetes Beispiel betrachtet werden.

Unmittelbare Kandidatenbewertungen während der Antworten der Kandidaten Zu den substantiellen Befunden können wir an dieser Stelle festhalten, dass es signifikante Effekte der Voreinstellungen und auch signifikante konditionale Effekte der Relationen in Abhängigkeit von den Voreinstellungen auf die unmittelbare Bewertung von Schmid und Mappus gibt, wenn wir uns auf die Analyse der direkten Antworten beschränken. Der wesentliche Einfluss liegt dabei auf Seiten der Voreinstellungen der Rezipienten. Darauf deuten die Befunde der Varianzdekomposition und die Größe der Effekte in den Ergebnisdarstellungen hin.

Die wesentliche Erkenntnis lässt sich in einer Abbildung zusammenfassen: Abbildung 6.39 stellt zum einen die mittleren RTR-Verläufe für Schmid und Mappus nur in Abhängigkeit von der Lagerzugehörigkeit der Rezipienten dar (stärkere, leicht transparente Linien). Zu diesen Wachstumskurven gehören auch die 95%-Konfidenzintervalle als transparente Flächen. Zum anderen sind die Wachstumskurven in Abhängigkeit der Lagerzugehörigkeit der Rezipienten und der Relation der Antworten abgebildet (schwächere Linien). So wird der zentrale Mechanismus hinter der Bewertung der Kandidaten während der Debatte deutlich. Die Rezipienten bewerten ihren Kandidaten positiv, unabhängig davon, ob er seine eigenen Leistungen und Pläne präsentiert, das gegnerische Lager angreift oder sich verteidigt. Auch die Effekte der Relationen auf die unentschiedenen Rezipienten bleiben begrenzt. Diese Gruppe reagiert auf alle Relationen mit im Zeitverlauf gemäßigt zunehmender Zustimmung. Während der Antworten des gegnerischen Kandidaten erfolgt in der Regel eine neutrale (Schmid) bis leicht negative (Mappus) Veränderung der RTR-Verläufe. Bei dieser Gruppe findet sich auch die einzige bedeutendere Abweichung von dieser einfachen Regel. Verteidigungen des Kandidaten aus dem anderen politischen Lager werden negativer bewertet.

Durch die detaillierte Modellierung der Rezipientenmerkmale lässt sich dieser Mechanismus – wie oben gezeigt – noch feiner beschreiben, und wir können einige weitere konditionale Effekte der Relationen herausarbeiten. Im Wesentlichen bildet diese Darstellung das Grundprinzip, nach dem sich die Bewertung von Schmid und Mappus während ihrer Antworten erklären lässt, gut ab. Rezipienten stimmen dem ihnen nahestehenden Kandidaten zu, wenn dieser in der TV-Debatte das Wort erhält. Dieser Befunde ist in Anbetracht des Forschungsstands zur Bewertung von Kandidaten in TV-Duellen



Anmerkungen
Bewertung von Schmid und Mappus auf einer Skala von −50 (größter Nachteil Sprecher) bis 50 (größter Vorteil Sprecher). Vorhersage durch die Modelle in den Tabellen A.8 (Schmid) und A.9 (Mappus).
Die stärkeren, leicht transparenten Linien beschreiben die mittleren Wachstumskurven über alle Relationen. Die Flächen zeigen ihre 95%-Konfidenzintervalle. Die schwächeren Linien stellen die Wachstumskurven in Abhängigkeit der Relation dar.

Abbildung 6.39: Effekte auf die Bewertung von Schmid und Mappus während aller Antworten im Vergleich

wenig überraschend (vgl. Kapitel 3.4). Bei der Suche nach Effekten, die von bestimmten Merkmalen des Inhalts ausgehen, müssen wir uns jedoch immer vergegenwärtigen, dass diese Effekte vor dem Hintergrund dieses grundlegenden Mechanismus einzuordnen sind. In den vorliegenden Analysen tragen die Inhalte der Antworten, hier konkret die Relationen Selbstpräsentation, Angriff und Verteidigung, zur Erklärung der unmittelbaren Bewertung bei, allerdings immer in Interaktion mit den Voreinstellungen. So können wir zeigen, dass Verteidigungen schlechter bewertet werden, wenn die Rezipienten dem sich Verteidigendem kritisch gegenüberstehen. Angriffe führen zu negativen Reaktionen bei denen, die eine hohe Meinung vom Angegriffenen haben. Diese Effekte sind ohne Frage interessant und waren bereits Bestandteil ausführlicherer Forschungsbemühungen (vgl. Kapitel 3.4.2), die hiermit weitergeführt werden können. Wenn wir aber die grundsätzliche Frage beantworten sollen, was die Bewertung eines Kandidaten während einer Debatte im Wesentlichen beeinflusst, so ist die Antwort auf Basis der vorliegenden Analyse eindeutig: Die politischen Voreinstellungen der Rezipienten sind wesentlich wichtiger als die Relationen der Antworten.

Aus Perspektive des Modelltests können wir feststellen, dass sich die Klasse des kreuzklassifizierten Wachstumskurvenmodells insgesamt bewährt hat, auch wenn sich einige Einschränkungen und Probleme feststellen lassen. Wenn wir die dynamische Veränderung der individuellen RTR-Verläufe angemessen erfassen und auch einen Einblick in die Veränderung der Bewertungen innerhalb der Antworten gewinnen wollen, so hilft die Spezifikation eines Wachstumskurvenmodells als Messmodell weiter. Durch die Schätzung eines latenten Intercepts und eines latenten linearen Slopes ermöglicht die hier eingesetzte Modellspezifikation eine detaillierte Beschreibung und Interpretation der Befunde. Sie ist der Dynamik der Kandidatenbewertungen innerhalb der Antwort angemessen und lässt sowohl in der Schätzung zum Beginn der Antwort als auch in der Veränderung über die Zeit eine angemessene Repräsentation der Varianz zu. Nicht verschwiegen werden sollen hier aber auch die praktischen Probleme, die sich aus einer komplexeren Modellspezifikation ergeben. Ein Problem für die praktische Anwendung ist die große Zunahme in der Zahl der zu interpretierenden Koeffizienten. Eine Einordnung der Befunde anhand tabellarischer Darstellungen ist nur noch eingeschränkt möglich. Aber auch den Interpretationen mithilfe visueller Präsentationen sind ab einer gewissen Zahl von gleichzeitig zu berücksichtigenden Koeffizienten Grenzen gesetzt. Schließlich erweist sich auch die numerische Evaluation der komplexeren Modelle, wenn sie zusätzlich über eine größere Anzahl von L2-Prädiktoren und Interaktionstermen verfügen, als mühsam. Die numerische Evaluation eines Modells erfordert auf einem hochwertigen Rechner ca. 30 Minuten Rechenzeit,

und die Ergebnisse müssen intensiv auf fehlerhaft konvergierte Lösungen geprüft werden.

Durch die Verwendung einer kreuzklassifizierten Datenstruktur ermöglichten wir es, wie bereits in Kapitel 6.2 demonstriert, zusätzlich zu dem verbreiteten Aggregationsmerkmal der Lagerzugehörigkeit mehrere sowie (quasi-) metrisch skalierte Prädiktoren der Rezipientenmerkmale zu berücksichtigen. Aus Perspektive der Varianzaufklärung erweist sich dies als sinnvoll. Auch wenn bereits die einfache Lagerzuordnung auf Basis der Wahlabsicht vor dem Duell einen substantiellen Anteil der rezipientenbezogenen Varianz erklären kann, erhöht sich dieser Anteil durch die (noch immer sehr einfachen) Skalometervariablen zu den Kandidaten und ihren Parteien nochmals deutlich. Darüber hinaus können weitere konditionale Effekte der Relationen, z.B. der Effekt der Angriffe in Abhängigkeit von der Einstellung zu Mappus auf die unmittelbare Bewertung von Schmid, erst durch diese Modellerweiterung herausgearbeitet werden.

Nichtsdestotrotz bleibt fast die gesamte Information des Faktors Lagerzugehörigkeit in den Modellen vorhanden, lediglich die Zugehörigkeit zum Regierungslager wird in den Modellen zur Erklärung der Bewertung von Mappus ersetzt. Dies deutet darauf hin, dass die weiteren Variablen zur Voreinstellung vor allem zusätzliche Informationen liefern, die durch die Wahlabsicht vor dem Duell operationalisierte Lagerzugehörigkeit aber ebenfalls wichtige Informationen enthält. Der Vergleich der Befunde erhöht auch das Vertrauen in die einfacher zu kommunizierenden Modelle, die auf der Aufteilung der Rezipienten in Gruppen nach ihrer Wahlabsicht beruhen.

6.4 Limitationen und Potenziale der Analyse von unmittelbaren Kandidatenbewertungen mit Mehrebenenmodellen

In diesem Kapitel haben wir gezeigt, wie die unmittelbaren Kandidatenbewertungen während eines TV-Duells mit verschiedenen Klassen von Mehrebenenmodellen analysiert werden können. Der zentrale Vorteil dieser Modelle gegenüber den bisher verbreiteten aggregationsbasierten Vorgehen zur deduktiven Analyse ist, dass sie direkt an den individuellen RTR-Messungen ansetzen. Dies ermöglicht die angemessene Überprüfung von Annahmen über Effekte, die auf Überlegungen zur individuellen Wahrnehmung und Informationsverarbeitung beruhen. Neben den einfachen Wachstumskurvenmodellen, die in der Kommunikationswissenschaft schon recht verbreitet sind, wenn sie auch unseres Wissens noch nicht für publizierte Analysen von RTR-Messungen eingesetzt

wurden (vgl. aber Iyengar et al., 2010), haben wir zwei weitere Modellklassen eingeführt.

Kreuzklassifizierte Mehrebenenmodelle berücksichtigen die doppelte Schachtelung der unmittelbaren Kandidatenbewertungen in den Rezipienten, die sie abgegeben, und den Einheiten des Debatteninhalts, zu denen sie abgegeben werden. Der substantielle Nutzen dieser Modellklasse ist groß: Mit ihr können theoretische Annahmen zur Erklärung der individuellen unmittelbaren Kandidatenbewertung durch Rezipienten- und Inhaltsmerkmale sowie deren Interaktionen geprüft werden. Da im Gegensatz zu den aggregationsbasierten Verfahren keine Notwendigkeit besteht, die Zahl der Prädiktoren oder die Zahl ihrer Ausprägungen zu beschränken, können die Mechanismen, die theoretisch-konzeptionell hinter den Effekten liegen, präziser in statistische Modelle überführt werden. Schließlich deckt sich die Struktur eines kreuzklassifizierten Modells mit der Idee einer doppelten Stichprobenziehung, die zumindest implizit vielen Forschungsinteressen der TV-Duell-Studien zugrunde liegt. Die allgemeinen Annahmen zu Effekten auf die unmittelbaren Kandidatenbewertungen sollen vor dem Hintergrund überprüft werden, dass die untersuchten Rezipienten eine Stichprobe aus einer Grundgesamtheit von Rezipienten und die Kandidatenaussagen eine Stichprobe aus einer Grundgesamtheit von (politischen) Aussagen (in einer TV-Debatte) sind. In den Modellen sind die Varianzen zwischen den Rezipienten innerhalb der Aussagen ebenso enthalten wie die Varianzen zwischen den Aussagen innerhalb der Rezipienten. Auch die Stichprobengrößen von Rezipienten, Aussagen und ihren Kombinationen werden berücksichtigt. Damit wird zum einen der unstrittigen Tatsache Rechnung getragen, dass individuelle Rezipienten unterschiedlich auf verschiedene Aussagen reagieren. Zum anderen werden diese Informationen in den Modellen genutzt, um die inferenzstatistischen Tests unter den Unsicherheiten, die sich aus der doppelten Stichprobenziehung ergeben, angemessen durchzuführen. In Kapitel 6.2 bewährt sich das kreuzklassifizierte Modell bei der Untersuchung der konditionalen Effekte des Issue Ownership in Interaktion mit den Voreinstellungen der Rezipienten auf die unmittelbaren Kandidatenbewertungen.

Die Erweiterung des kreuzklassifizierten Modells um ein als Wachstumskurvenmodell spezifiziertes L1-Messmodell behält alle genannten Vorteile bei. Zusätzlich wird in den kreuzklassifizierten Wachstumskurvenmodellen die dynamische Veränderung der unmittelbaren Kandidatenbewertungen innerhalb der Einheiten des Debatteninhalts – hier innerhalb der direkten Antworten auf Fragen der Moderatoren bzw. innerhalb der zehn auf einen Relationswechsel folgenden Sekunden – durch einen oder mehrere Parameter erfasst. So wird es möglich, Annahmen über die dynamische Veränderung der individuellen Kandidatenbewertungen in Abhängigkeit von Rezipienten- und Aussagenmerk-

malen sowie deren Interaktion zu prüfen. Wenn die empirisch beobachtete Kandidatenbewertung wie in Kapitel 6.3.2 durch einen latenten Intercept und einen latenten linearen Slope angenähert wird, können Effekte unterschieden werden, die sich auf den geschätzten Ausgangswert der Bewertung und auf die geschätzte Veränderung der Bewertung im Verlauf der Antwort beziehen. Vor dem Hintergrund der „On-Line Versus Memory-Based Process Models of Political Evaluation" (Lavine, 2002, S. 225) können erstere vermutlich als direkte Reaktionen auf Schlüsselreize in den Aussagen interpretiert werden („on-line"). Zweite sind dagegen mit höherer Wahrscheinlichkeit als Folge einer kognitiven Auseinandersetzung mit der Aussage zu verstehen, bei denen bereits vorhandene Informationen mit den neuen Informationen aus der Aussage verknüpft werden, um dann ein neues oder aktualisiertes Urteil zu bilden („memory-based"). Dazu passt auch der Befund aus Befragungsstudien, dass mit on-line Urteilen kürzere, mit memory-based Urteilen dagegen längere Reaktionszeiten bei der Beantwortung von Einstellungsfragen einhergehen (Matthes, Wirth & Schemer, 2007). Ähnliche Interpretationen in Anlehnung an Zwei-Routen-Modelle der Informationsverarbeitung (z.B. Chaiken, 1980; Petty & Cacioppo, 1986) sind ebenfalls möglich. Hier sprächen Effekte im Intercept für eine periphere oder heuristische Verarbeitung bestimmter Cues in den Aussagen, während Effekte im Slope als Indikatoren für die zentrale Route mit einer systematischen Verarbeitung der Kandidatenaussage aufgefasst würden. Schließlich erlaubt die Modellierung der Veränderung über die Zeit bessere Prognosen darüber, zu welchem Umfang ein Effekt sich im Verlauf einer Aussage kumuliert. Die Analyse der unmittelbaren Kandidatenbewertungen in Abhängigkeit von den Voreinstellungen der Rezipienten und den Relationen in den direkten Antworten der Kandidaten auf Fragen der Moderatoren in Kapitel 6.3.2 demonstriert exemplarisch die analytischen Potenziale der kreuzklassifizierten Wachstumskurvenmodelle. Weniger ertragreich ist die Anwendung der Modellklasse für die Untersuchung von Veränderungen der Kandidatenbewertungen nach Wechseln der Relation (Kapitel 6.3.3). Dies ist jedoch nicht der Modellklasse geschuldet. Vielmehr vermuten wir, wie oben ausführlich diskutiert, dass die Prämissen, auf denen solche Analysen aufbauen, nicht haltbar sind.

Limitationen der vorgestellten Modelle und mögliche Erweiterungen

Auch wenn die in dieser Arbeit vorgeschlagenen Mehrebenenmodelle der unmittelbaren Kandidatenbewertungen bereits in der vorliegenden Form erhebliche Vorteile gegenüber den bislang zur deduktiven Analyse eingesetzten aggregationsbasierten Verfahren bringen, haben sie einige Limitationen. Einige

dieser Einschränkungen lassen sich durch einfachere Erweiterungen der Modelle beheben. Andere erfordern größere Modifikationen oder sind mit der gegenwärtig verfügbaren Standard-Software wenig praktikabel bzw. gar nicht möglich. Insofern ist die folgende Liste auch als eine Anregung für weitere Forschung in der hier begonnenen Richtung zu verstehen.

Modellierung der (zeitlichen) Assoziationen zwischen den Analyseeinheiten des Debatteninhalts Die vorgestellten Modelle vernachlässigen die Information, in welcher Reihenfolge die Turns bzw. Antworten der Kandidaten in der Debatte aufeinander folgen. Auch die kreuzklassifizierten Wachstumskurvenmodelle berücksichtigen nur die dynamischen Veränderungen der RTR-Messungen *innerhalb* einer Antwort. Denkbar sind zwei Möglichkeiten, die zeitliche Struktur auch auf den höheren Ebenen zu berücksichtigen. Zum einen können zeitliche Assoziationen der L2-Einheiten des Debatteninhalts (hier: Turns bzw. Antworten) modelliert werden. Die einfachste Möglichkeit ist die Erweiterung um einen L2-Prädiktor, der die Information enthält, an welcher Stelle eine gegebene Einheit in der Reihenfolge aller Einheiten liegt. So könnte untersucht werden, ob die Kandidatenbewertungen über die Einheiten hinweg einem Trend folgen, also beispielsweise die Bewertungen eines Kandidaten im Debattenverlauf immer positiver werden. Wenn dieser Prädiktor als Random Effect spezifiziert wird, der Trend also zwischen den Rezipienten frei variieren darf, könnten interindividuelle Unterschiede zwischen den Rezipienten wiederum durch deren Merkmale (z.B. deren Voreinstellungen) erklärt werden. Ebenfalls dieser Logik folgend könnten den L2-Einheiten Eigenschaften der vorangegangenen Einheiten als weitere L2-Prädiktoren zugeordnet werden. Damit würde beispielsweise geprüft, ob ein Angriff anders wirkt, wenn er auf einen Angriff des Gegenkandidaten folgt (also ein Gegenangriff ist). Oder es könnte die Vermutung getestet werden, dass sich der Effekt bestimmter Stilmittel bei mehrfacher Wiederholung abnutzt.

Zum anderen könnte die zeitliche Assoziation zwischen den Kandidatenbewertungen selbst über die Einheiten hinweg innerhalb der individuellen Rezipienten modelliert werden. Dies entspräche einer in der Zeitreihenanalyse typischen Prozessmodellierung durch autoregressive, integrierte oder Moving-Average-Prozesse (sog. ARIMA-Modelle, vgl. z.B. Scheufele, 1999), jedoch im Gegensatz zu den von Nagel (2012) diskutierten Modellen auf Niveau der mit den Messmodellen erfassten Bewertungen der individuellen Rezipienten. Einer solchen Modellierung läge die konzeptionelle Vorstellung zugrunde, dass die Bewertung während einer Einheit auch davon beeinflusst wird, wie ein Rezipient den Kandidaten in den Einheiten zuvor bewertet hat. Formal wäre es

in einem Modell recht einfach möglich, die funktionale Form der zeitlichen Prozesse zwischen den Rezipienten variieren zu lassen. Allerdings folgen daraus Modellspezifikationen mit sehr vielen Parametern, was sehr wahrscheinlich zu Problemen bei der praktischen Schätzung der Modelle führen würde.

Neben den zu erwartenden praktischen Problemen haben wir auf diese Erweiterungen verzichtet, da die TV-Duell-Studie nicht-experimentell angelegt ist. Daraus ergibt sich, dass einige inhaltliche Merkmale der Debatte, z.B. die zeitliche Abfolge der Themen, mit dem Debattenverlauf korreliert sind. Daher können zeitliche Effekte oft nur schwierig in Hinblick auf ihre Kausalität eingeordnet werden: Der (hypothetische) Befund, dass ein Kandidat im Zeitverlauf immer besser bewertet wird, kann zum einen auf seine kumulierte Überzeugungsleistung zurückzuführen sein. Zum anderen könnte dieser Befund zustande kommen, da zu Beginn vor allem Themen diskutiert wurden, bei denen der Kandidat im Nachteil war, während zum Ende der Debatte seine Issue-Owner-Themen behandelt wurden. Um diese Frage zu klären, müssten die thematischen Blöcke von mehreren Gruppen in randomisierten Reihenfolgen rezipiert werden. In einer solchen randomisierten Versuchsanordnung sollte dann eine (einfach parametrisierte) Modellierung solcher Effekte in Erwägung gezogen werden.

Explizite Modellierung der Fehlerterme (Residuen) Konzeptionell ähnlich der Modellierung zeitlicher Abhängigkeiten in den beobachteten Daten ist die explizite Modellierung (zeitlicher) Abhängigkeiten in den Residuen. Bei solchen Überlegungen ist zuerst zu bedenken, dass in einer Mehrebenenmodellierung immer auch eine Modellierung der Residuen inbegriffen ist. Im Gegensatz zur einfachen Regression, die eine vollständige Unabhängigkeit der Residuen fordert, erlaubt das Mehrebenenmodell Korrelationen der Residuen innerhalb einer durch die Mehrebenenstruktur gebildeten Gruppe (Gelman & Hill, 2006, S. 8). Die Abhängigkeit der RTR-Messungen, die einem Rezipienten, einem Turn bzw. einer Antwort oder einer Kombination von Rezipienten und Turn bzw. Antwort zuzuordnen sind, ist ja gerade ein ausschlaggebender Grund dafür, dass wir das Mehrebenenmodell als geeignete Modellklasse vorschlagen. Eine ganz wesentliche Modellierung der Residuen ist der Kern jedes Mehrebenenmodells.

Es ist jedoch auch darüber hinaus möglich, die Form der Residuen explizit in der Modellformulierung zu berücksichtigen. Eine gute Einführung in solche Modellerweiterungen geben Pinheiro und Bates (2000, S. 206-267). Dabei ist zwischen Modellen der Residualkovarianz und der Residualvarianz zu unterscheiden. Erstgenannte beziehen sich auf die Korrelationen der Residuen

untereinander, zweitgenannte auf die Assoziation der Residualvarianz mit den Prädiktoren. In unserer Anwendung käme vor allem eine Modellierung der L1-Residuen, also der Abweichungen der einzelnen RTR-Messungen von den durch die Wachstumskurvenmodelle geschätzten latenten Kandidatenbewertungen, infrage. Die Residualkovarianzen werden meist durch Autokorrelationen bzw. andere ARIMA-Prozesse beschrieben. Das AR(1)-Modell als einfachste und verbreitetste Spezifikation unterstellt einen autoregressiven Prozess erster Ordnung. Es besagt in unserem Beispiel, dass die Abweichung einer beobachteten Kandidatenbewertung eines Rezipienten in einer Analyseeinheit von der durch das Modell geschätzten Kandidatenbewertung zum Zeitpunkt t von der Abweichung zum Zeitpunkt $t-1$ abhängt. Die inhaltliche Interpretation eines AR(1)-Modells der Residuen ist im Gegensatz zu einem äquivalenten Modell der Daten schwierig. Der Parameter des Modells quantifiziert, wie groß der Modellfehler in einer Sekunde in Abhängigkeit vom Umfang des Modellfehlers in der Sekunde zuvor ausfällt. Vereinfacht könnte dies als eine Diagnose verstanden werden, wie sehr sich die Fehler in Abhängigkeit von den vorangegangenen Fehlern kumulieren.

Die Autokorrelation (oder andere Abhängigkeiten der Residuen voneinander) stellen eine Verletzung der Prämissen des allgemeinen Regressionsmodells dar. Sie weisen auf eine Fehlspezifikation des Modells hin und können die Inferenzen auf Basis dieses Modells verschlechtern oder verzerren. Zuvorderst sollten solche Abhängigkeiten als ein Hinweis verstanden werden, die zeitspezifische Komponente des Modells weiter zu ergänzen. Es besteht jedoch auch die Möglichkeit, die Fehlspezifikation durch ein Modell der Residuen zu beheben. Dieses Vorgehen ist allerdings nicht unumstritten. In einem Artikel mit dem plakativen Titel „A simple message for autocorrelation correctors: Don't" warnt Mizon (1995, S. 267) vor einem unreflektierten Korrigieren für autoregressive Prozesse in den Residuen durch Residuen-Modelle. Mit solchen Korrekturen gehen implizit starke Prämissen einher, die jedoch in den seltensten Fällen von den Anwendern auf ihre inhaltliche Passung zu den konzeptionellen Modellen geprüft werden. Wenn aber die Prämissen nicht erfüllt werden, führt die Korrektur zu inhaltlich wie statistisch verzerrten Inferenzen. Er argumentiert, dass eine explizite Modellierung der beobachteten Daten in Einklang mit den Annahmen des zu prüfenden theoretischen Prozessmodells fast immer zu bevorzugen ist. In unserer Anwendung ist die Modellierung von Wachstumskurvenmodellen als L1-Messmodelle eine solche explizite Modellierung der beobachteten Daten. Wie in den Kapiteln 6.1 und 6.3 ausführlich dargelegt, berücksichtigt der frei geschätzte Slope-Parameter die zeitliche Abhängigkeit der aufeinander folgenden unmittelbaren Kandidatenbewertungen, indem er die Veränderungen von einem Messzeitpunkt auf den nächsten quantifiziert.

Die verbleibende Autokorrelation in den L1-Residuen sollte in diesem Sinne vor allem als ein Hinweis darauf verstanden werden, dass die einfachen Wachstumskurven die Veränderungsprozesse nur ungenau abbilden – was in den grafischen Darstellungen der Modelle auch unschwer zu erkennen ist (vgl. z.B. Abbildung 6.6, S. 187).

Die Assoziation der Residualvarianz mit Prädiktoren des Modells wird als Heteroskedastizität bezeichnet. Sie ist in unseren Wachstumskurvenmodellen der Kandidatenbewertungen während der Antworten, wie bereits in Kapitel 6.1 diskutiert, an den im Zeitverlauf weiter werdenden Konfidenzintervallen um die vorhergesagten RTR-Verläufe zu erkennen. Formal ausgedrückt besteht eine Korrelation zwischen dem Zeit-Prädiktor und der Varianz der Residuen. Sie ergibt sich aus der Tatsache, dass die Varianz der beobachteten unmittelbaren Kandidatenbewertungen im Zeitverlauf größer wird. Es handelt sich damit *per se* nicht um eine Fehlspezifikation, sondern ein real existierendes Charakteristikum von dynamischen Veränderungen ausgehend von einem neutralen Startwert, der besagt, dass die meisten Rezipienten während der Fragen der Moderatoren keinen Eindruck von den Kandidaten haben bzw. abgeben (vgl. auch die Modellspezifikation von Iyengar et al., 2010, S. 24). Im Gegensatz zur Modellierung von Abhängigkeiten zwischen den Residuen, z.B. mit einem AR(1)-Modell der Residuen, kann hier ein direkter inhaltlicher Bezug zu den Daten und zum konzeptionellen Modell hergestellt werden. Daher ist eine Berücksichtigung der Heteroskedastizität in einer Spezifikation der Residuen unseres Erachtens eine erwägenswerte Erweiterung der hier vorgestellten Modelle. Da bei der Berechnung der vorhergesagten Verläufe die Kovarianzen zwischen den anderen L2-Prädiktoren und dem Zeit-Prädiktor berücksichtigt werden, sind die in dieser Arbeit graphisch präsentierten Ergebnisdarstellungen nicht fehlspezifiziert. Die fehlende Spezifikation eines Modells für die Heteroskedastizität der Residuen kann sich jedoch für die inferenzstatistischen Tests der Koeffizienten des Zeit-Prädiktors und der Interaktionsterme, die diesen Prädiktor enthalten, als problematisch erweisen.

In den hier vorgestellten Modellen verzichten wir darauf, weitergehende formale Modelle der Residuen zu spezifizieren. Damit folgen wir theoretischen und praktischen Erwägungen. Aus den genannten Gründen erscheint uns die Modellierung einer Kovarianzstruktur zwischen den Residuen, z.B. nach einem AR(1)-Prozess, inhaltlich wenig ertragreich. Mit einer solchen Spezifikation könnten wir lediglich lernen, dass die hier genutzten Wachstumskurvenmodelle den beobachteten Veränderungsprozess nur sehr vereinfacht abbilden. Diese Erkenntnis ist angesichts der visuellen Inspektion der beobachteten und geschätzten RTR-Verläufe wenig überraschend. Ohne ein besseres Verständnis des tatsächlichen datengenerierenden Prozesses verzichten wir darauf, ein

Modell zur Korrektur der Autokorrelation zu spezifizieren, da dies unter den gegebenen Umständen im Zweifel mehr Schaden als Nutzen bringt. Und hätten wir ein tieferes Verständnis des Prozesses, so wäre es empfehlenswerter, ihn direkt in Bezug auf die beobachteten RTR-Messungen zu modellieren (Mizon, 1995). Eine Erweiterung des Modells zur Berücksichtigung der Heteroskedastizität wäre dagegen wünschenswert. Hierauf müssen wir aus dem praktischen Grund verzichten, dass die eingesetzte Software *lme4* keine Modellierung der Residuen vorsieht. Unseres Wissens ist auch keine andere Software in der Lage, ein Modell mit der Komplexität der hier vorgestellten kreuzklassifizierten Wachstumskurvenmodelle unter Erweiterung um ein Modell der Residuen zu schätzen. Daher müssen wir zum gegebenen Zeitpunkt die Konsequenzen tragen. Sie sind jedoch hinreichend akzeptabel: Der Verzicht auf ein spezifisches Modell für die Residuen erhöht zwar die Gefahr einer fehlerhaften Schätzung der Varianzkomponenten. Die Schätzung der Fixed-Effect-Koeffizienten und ihrer Standardfehler – und damit der Koeffizienten, die für unsere inhaltlichen Inferenzen zentral sind – ist nach Simulationsstudien jedoch kaum beeinträchtigt (Ferron, Dailey & Yi, 2002; Wolfinger, 1993).

Einfache Wachstumskurvenmodelle, mit denen wie im einführend demonstrierten Beispiel nur die RTR-Messungen während eines einzelnen Stimulus untersucht werden (vgl. Kapitel 6.1), sollten gegebenenfalls um ein Modell zur Berücksichtigung der Heteroskedastizität der L1-Residuen ergänzt werden. Pinheiro und Bates (2000, S. 206-267) beschreiben dieses Vorgehen ausführlich für das *R* Paket *nlme* (Pinheiro, Bates, DebRoy, Sarkar & R Core Team, 2013). Bei Peugh und Enders (2005) findet sich eine entsprechende Einführung mit *SPSS* und *SAS*. In kreuzklassifizierten Wachstumskurvenmodellen scheint die Nutzung einer bayesianischen Schätzung möglich (Iyengar et al., 2010). Schließlich entwickelte Koller (2013) in seiner Dissertation kürzlich einige robuste Schätzverfahren, die weniger anfällig für die Verletzung der Modellprämissen sind. Die Verfahren sind als Erweiterung für *lme4* im *R* Paket *robustlmm* (Koller, 2014) verfügbar. Da die numerische Evaluation der robusten Schätzverfahren (bisher) jedoch deutlich weniger effizient arbeitet, ist ihre Anwendung für die hier vorgestellten komplexen Modelle zurzeit wenig praktikabel. Für einfachere Modelle sollte diese Möglichkeit jedoch genutzt werden.

Verbesserung von Interpretierbarkeit und Datenpassung durch echte nichtlineare Modelle Ein großer Nutzen des Wachstumskurvenmodells mit einem latenten Intercept und einem latenten linearen Slope bei der Analyse der unmittelbaren Kandidatenbewertungen während der Antworten ist die differenzierte Interpretation der beiden Parameter. Allerdings ist die interpretative

Zuordnung auch zu einem gewissen Grad fehleranfällig. Der Vergleich der Datenpassung unterschiedlicher L1-Spezifikationen sowie die visuelle Inspektion der beobachteten und vorhergesagten RTR-Verläufe weist zudem darauf hin, dass eine nicht-lineare Annäherung angemessener wäre. In der vorliegenden Arbeit haben wir nicht-lineare RTR-Verläufe durch polynomiale Terme realisiert. Diese erweisen sich jedoch als schwer interpretierbar und führen zudem bei der numerischen Evaluation komplexerer Modelle zu Problemen.

Eine Alternative hierzu sind echte nicht-lineare Modelle. So beschreiben Iyengar et al. (2010) in einem Konferenzbeitrag RTR-Bewertungen während politischer Werbespots, die zu Beginn des Spots am neutralen Mittelpunkt der Skala beginnen, durch eine Gompertz-Funktion mit drei distinkten Parametern. Der erste Parameter gibt den geschätzten Zeitpunkt an, an dem sich die mittlere RTR-Bewertung vom neutralen Skalenmittel entfernt. Der zweite Parameter beschreibt die geschätzte Asymptote, die die mittlere RTR-Bewertung im Zeitverlauf erreicht. Der dritte Parameter gibt schließlich die Wachstumsrate an und quantifiziert damit die Geschwindigkeit, mit der sich die mittlere RTR-Bewertung von der ersten Bewegung weg vom Skalenmittelpunkt hin zur Asymptote bewegt (vgl. zu dieser und anderen nicht-linearen Funktionen auch Pinheiro & Bates, 2000, S. 511-521). Eine solche Parametrisierung hat den großen Vorteil, dass sich jeder der drei Parameter direkt inhaltlich anhand einer *a priori* definierten Funktion interpretieren und sich so einem theoretisch fundierten Prozess zuordnen lässt. Die Nicht-Linearität ermöglicht zudem eine bessere Datenpassung der vorhergesagten RTR-Verläufe.

Die statistischen Grundlagen zur Schätzung solcher Modelle sind bereits seit Längerem vorhanden (z.B. Lindstrom & Bates, 1990). Für einfache Wachstumskurvenmodelle sind sie beispielsweise im Paket *nlme* in großem Umfang implementiert (Pinheiro & Bates, 2000, S. 274-414), und auch für kreuzklassifizierte nicht-lineare Modelle sind grundlegende Funktionen in *lme4* vorhanden. Jedoch erwies sich das Vorhaben, die RTR-Messungen während der in Kapitel 6.1 beschriebenen Antwort als nicht-lineares Modell auf Individualniveau zu schätzen, als nicht umsetzbar. Über die Gründe hierfür können wir nur spekulieren. Es ist wahrscheinlich, dass die empirisch beobachteten RTR-Messungen von zu vielen individuellen Rezipienten von der oben beschriebenen idealtypischen Gompertz-Funktion abweichen, um eine stabile Schätzung zu erreichen. Dies mag ein Charakteristikum der hier vorliegenden Daten sein, ist aber vermutlich eher auf das nicht-deterministische menschliche Verhalten zurückzuführen. Pinheiro und Bates (2000) analysieren in ihren Beispielen größtenteils naturwissenschaftliche Datensätze, bei denen die funktionalen Formen der Effekte einheitlicher und ihre Messung deutlich präziser sind. Auch die bayesianische Schätzung von Iyengar et al. (2010) löst dieses Problem nicht. Hier

werden die RTR-Bewertungen nicht für die individuellen Rezipienten, sondern für die Kombinationen von politischen Lagern und Werbespots analysiert. Da das Ausweichen auf ein Vorgehen, das die Zusammenfassung der Bewertungen von individuellen Rezipienten erfordert, dem Ziel dieser Arbeit widerspricht, haben wir die echte nicht-lineare Modellierung nicht weiter verfolgt. Wenn jedoch eine Analyse auf Aggregatniveau angestrebt wird, oder es gar gelingt, die Modelle erfolgreich auf individuelle RTR-Messungen zu übertragen, ist die Verwendung echter nicht-linearer Spezifikationen äußerst vielversprechend. Weitere Arbeiten sollten an dieser Stelle ansetzen.

Schätzung frei variierender Effekte In den vorgestellten Modellen werden die Effekte der Rezipienten- und Inhaltsmerkmale sowie ihrer Kombinationen nur als Fixed Effects spezifiziert. Damit setzen wir inhaltlich bedeutsame Prämissen. Zur Illustration ein einfaches Beispiel: In einem Modell ohne Cross-Level-Interaktionen würden wir davon ausgehen, dass die Einstellung zu Mappus während jeder Antwort denselben Effekt auf die unmittelbare Bewertung von Mappus hat. Umgekehrt würden wir davon ausgehen, dass eine Verteidigung bei jedem Rezipienten denselben Effekt hat. Die Abweichungen von den so ermittelten durchschnittlichen Effekten würden ohne weitere Spezifikation den jeweiligen Fehlerkomponenten zugerechnet. Da wir von Interaktionseffekten von Rezipienten- und Inhaltsmerkmalen ausgehen, modellieren wir darüber hinaus eine spezifische Interaktion zwischen beiden Merkmalen. Dadurch lassen wir zu, dass Rezipienten in Abhängigkeit von ihrer Voreinstellung zu Mappus unterschiedlich auf Verteidigungen und Nicht-Verteidigungen reagieren, bzw. die Voreinstellung während Verteidigungen und den übrigen Aussagen unterschiedliche Effekte hat. Alle Variabilität der Effekte von Voreinstellungen und Verteidigungen, die über diese spezifisch modellierte Interaktion hinausgeht, wird aber weiterhin den unspezifischen Fehlervarianzen zugeschlagen.

Mehrebenenmodelle bieten jedoch die Möglichkeit, die Effekte von Prädiktoren zwischen den Einheiten, nach denen der Datensatz gruppiert ist, frei variieren zu lassen. Im beschriebenen Beispiel könnte der Effekt der Voreinstellung zu Mappus zwischen den Antworten variieren. Somit würde ermittelt, während welcher Antworten die Voreinstellung einen kleineren oder größeren Einfluss auf die unmittelbare Bewertung des Kandidaten hat. Genauso könnte mit einem frei variierenden Verteidigung-Prädiktor untersucht werden, in welchem Ausmaß verschiedene Rezipienten unterschiedlich auf diese Relation reagieren. Entsprechend des in diesem Kapitel demonstrierten Vorgehens zur Reduzierung der unspezifischen Fehlervarianz auf Ebene der Personen, Turns

bzw. Antworten und Messmodelle könnten dann wiederum die Varianzen um die frei variierenden Prädiktoren durch die Aufnahme weiterer Merkmale in das Modell erklärt werden.

Die Berücksichtigung zwischen den Einheiten variierender Effekte ist ein großes Potenzial der Mehrebenenanalyse (Gelman & Hill, 2006, S. 6), das in dieser Arbeit aber ungenutzt bleiben muss. Dies ist vor allem dem Rahmen der Arbeit geschuldet, aber auch der bereits ohne diese Erweiterung bestehenden Komplexität der vorgestellten Modelle. Im Kontext weiterer Analysen bietet es sich an, auch dieses Potenzial zu nutzen.

Beschränkung der (kreuzklassifizierten) Wachstumskurvenmodelle auf Ausschnitte der Debatte (Antworten) Eine letzte Limitation, die hier genannt werden muss, bezieht sich nicht direkt auf das Analyseverfahren der (kreuzklassifizierten) Wachstumskurvenmodelle, sondern auf die für den ertragreichen Einsatz notwendige Beschränkung auf die Untersuchung der unmittelbaren Kandidatenbewertungen während direkter Antworten auf Fragen der Moderatoren. Diese Beschränkung hat sich als notwendig erwiesen, da nur in diesen Debattenausschnitten die dynamischen Veränderungen der Urteile in einem relativ standardisierten Kontext untersucht werden kann. Für das deduktive Unterfangen, allgemeine Annahmen über die Effekte von Eigenschaften der Kandidatenaussagen auf die unmittelbaren Urteile unter Berücksichtigung von Rezipienteneigenschaften zu prüfen, ist diese Limitation weniger bedeutsam. Wenn allerdings auch ein Schluss auf die Wirkung der relationalen Strategie der Kandidaten in der gesamten Debatte angestrebt werden soll, kann sich diese Auswahl von einigen Debattenausschnitten als wesentliche Einschränkung herausstellen. Eine (sicherlich übermäßig vereinfachende) Frage wie die, ob es sich für Mappus denn nun „gelohnt hat“, dass er sich für einen Amtsinhaber mit vielen Angriffen relativ aggressiv verhalten hat (Bachl, Käfferlein & Spieker, 2013b), lässt sich mit der vorliegenden Analyse nur unvollständig beantworten. Nicht alle Angriffe von Mappus fielen in den ersten 20 Sekunden einer Antwort auf eine Frage der Moderatoren. Und von den hier empirisch untersuchten Angriffen sollte nur sehr vorsichtig auf die Angriffe außerhalb des Untersuchungsmaterials geschlossen werden. Wir könnten an dieser Stelle lediglich feststellen, dass die untersuchten Angriffe im rot-grünen Lager etwas schlechter ankamen als einfache Selbstpräsentationen, während es für die Reaktionen der eigenen Anhänger und der Unentschiedenen kaum einen Unterschied machte, ob er angriff oder seine eigenen Pläne und Leistungen thematisierte.

Da sich die Analyse der Veränderungen infolge von Relationswechseln mit diesen Modellen als wenig ertragreich erwiesen hat, steht für die empirisch

gestützte Beantwortung einer solchen Frage nach den gesamten Debatteninhalten weiterhin nur die Peak-Spike-Analyse mit allen ihren Stärken und Schwächen zur Verfügung. Bereits in den Empfehlungen zu diesem Verfahren (Kapitel 5.2.4) haben wir angeregt, für Passagen einer Debatte, die zu anhand einer aggregierten RTR-Zeitreihe identifizierten Peaks führen, die RTR-Bewertungen der individuellen Rezipienten zumindest visuell zu inspizieren. Ein solches Vorgehen ließe sich auch um eine auf Ebene der individuellen Rezipienten ansetzende quantitative Komponente ergänzen. So wäre es möglich, die individuellen RTR-Verläufe vom Beginn des Turns bis zum Erreichen des Aggregat-Peaks durch ein Wachstumskurvenmodell abzubilden und interindividuelle Differenzen durch Merkmale der Rezipienten zu erklären. In einem zweiten Schritt könnte dieses Modell auch zu einem kreuzklassifizierten Wachstumskurvenmodell mit den einzelnen Peaks als zweiter Klassifikationsebene erweitert werden, um Systematiken über die Peaks hinweg herauszuarbeiten. Eine solche Analyse ist noch immer induktiver Natur, da das zu untersuchende Material auf Basis der RTR-Zeitreihe identifiziert wird. Alle der induktiven Logik inhärenten Limitationen bleiben so erhalten, insbesondere die Vernachlässigung der Passagen, die nicht zu bemerkenswerten Reaktionen im Aggregat führen. Aber immerhin wird die Analyse so um eine systematische, quantifizierende Beschreibung erweitert, die auf dem für viele theoretische Erklärungen angemessenen individuellen Datenniveau ansetzt.

7 Fazit und Ausblick

> Remember that all models are wrong; The practical question is how wrong they have to be to not be useful (Box & Draper, 1987, S. 74).

Ziel dieser Arbeit war es, Verfahren für die Analyse von unmittelbaren Kandidatenbewertungen in TV-Debatten zu diskutieren und auf ihre Eignung für die Beantwortung verschiedener kommunikationswissenschaftlicher Fragestellungen zu prüfen. Wie der renommierte Statistiker George E. P. Box in seinem vielzitierten Bonmot feststellt, sind alle (statistischen) Modelle zu einem gewissen Grad falsch. Sie müssen es auch sein, da sie die komplexe Realität vereinfacht abbilden und so die den empirischen Beobachtungen zugrunde liegenden Mechanismen einer allgemeineren Erklärung zugänglich machen sollen. Dementsprechend sind auch alle in dieser Arbeit diskutierten Modelle – die etablierten Verfahren wie auch die vorgeschlagenen Mehrebenenmodelle – vereinfachte, gewissermaßen falsche Annäherungen an den Prozess, der erklärt, wie welche Rezipienten die Kandidaten während welcher Aussagen bewerten. Vor diesem Hintergrund wollen wir in diesem Kapitel abschließend abwägen, welche Verfahren für die Beantwortung welcher Forschungsfragen nützlich sind. Es geht also letztlich um die Frage, ob die den jeweiligen Verfahren inhärenten Probleme in Anbetracht des Nutzens akzeptabel sind, oder ob sie zu irreführenden Schlussfolgerungen bezüglich einer Forschungsfrage verleiten.

Im Folgenden gehen wir zuerst auf die Limitationen der vorliegenden Arbeit ein, um eine Einordnung der Befunde vor ihrem Hintergrund zu ermöglichen. Dann fassen wir die Ergebnisse in Bezug auf die in Kapitel 2 herausgearbeiteten Forschungsfragen, die mit TV-Duell-Studien beantwortet werden sollen, zusammen. Abschließend geben wir einen Ausblick, wie sich insbesondere die vorgeschlagenen Mehrebenenmodelle auch in Studien als nützlich erweisen können, die sich, wie in Kapitel 3 dargestellt, von der hier behandelten TV-Duell-Studie unterscheiden.

Limitationen

Die vorliegende Arbeit hat, wie jede wissenschaftliche Arbeit, Limitationen, die bei der Einordnung der Befunde berücksichtigt werden müssen. Da sich große Teile dieser Arbeit mit der Diskussion von Limitationen der Messmethode RTR

und den Verfahren zur Analyse von RTR-Messungen befassen, soll an dieser Stelle eine knappe Zusammenfassung mit dem Verweis auf die entsprechenden Kapitel ausreichen. Wie jede Studie, die auf der rezeptionsbegleitenden Erhebung von Publikumsurteilen mit RTR-Messungen aufbaut, ist auch diese von den Problemen der Messmethode betroffen (vgl. Kapitel 3.3). Zu nennen sind diesbezüglich zum einen die bisher nur in Teilen abgesicherte Reliabilität und Validität der RTR-Messungen. Anhand einiger Indikatoren können wir zwar nachweisen, dass die Qualität unserer Daten nach den auch in anderen RTR-Studien angelegten Kriterien akzeptabel erscheint (vgl. Kapitel 4.2). Bei der Auseinandersetzung mit der methodologischen Literatur und der Evaluation der eigenen Daten wird aber auch deutlich, dass sowohl hinsichtlich der Qualität der Messungen selbst als auch hinsichtlich geeigneter Indikatoren zur Feststellung ihrer Qualität noch einiger Forschungsbedarf besteht. Zum anderen weist der methodologische Forschungsstand darauf hin, dass die RTR-Messung die externe Validität der Studie einschränkt. Diese Limitation ist noch zusätzlich zu den ohnehin gegebenen Problemen einer Laborstudie, bei der sehr viele Teilnehmer den Stimulus gemeinsam und in einer künstlichen Situation rezipieren, zu beachten.

Weiter ist die Stichprobenziehung der Studie zu problematisieren. Bezüglich der Personenstichprobe gelten die Einwände, die gegen fast alle kontrollierten Rezeptionsstudien zu TV-Debatten vorzubringen sind (vgl. Kapitel 2 und 3.4). Die Verfügbarkeit der RTR-Hardware beschränkt die Stichprobenziehung regional auf das Umfeld der Erhebungsstandorte Stuttgart und Ravensburg. Aus der Anforderung an die Teilnehmer, das TV-Duell an einem Abendtermin an den Hochschulen zu verfolgen, folgt eine erhöhte Selbstselektion, die zu Verzerrungen zugunsten höher Gebildeter und politisch stärker Interessierter führt. Insgesamt konnten wir aber durch eine Quotierung wichtiger Merkmale, allen voran der politischen Voreinstellung, eine diverse Stichprobe realisieren. Sie hat zudem im Vergleich zu vielen Vorgängerstudien einen beachtlichen Umfang auch in den für die Aggregatanalysen relevanten Subgruppen. Bedenklicher für die Verallgemeinerung unserer Befunde ist dagegen die Beschränkung des Untersuchungsgegenstands auf ein einziges TV-Duell, an dem zwei bestimmte Kandidaten teilnahmen und das in einem spezifischen politischen Kontext stattfand. Auch diese Limitation teilen wir mit einem Großteil der Publikationen zu TV-Debatten. Im Verlauf der Arbeit machen wir jedoch mehrmals deutlich, dass für einige Forschungsfragen, die anhand einer TV-Duell-Studie beantwortet werden sollen, auch die Debatteninhalte als Stichproben aus größeren Grundgesamtheiten politischer Botschaften aufgefasst werden. Die Inhalte einer Debatte sind aber natürlich in vielerlei Hinsicht von ihren Rahmenbedingungen abhängig. Es sei an dieser Stelle daher nochmals explizit darauf

hingewiesen, dass gerade die inhaltlichen Befunde dieser Arbeit – etwa zu den Effekten des Issue Ownership und der Relationen – auch in anderen Kontexten repliziert werden müssen, um ihre Generalisierbarkeit zu prüfen.

Auch für das grundlegende Ziel dieser Arbeit – die Evaluation und Weiterentwicklung von Analyseverfahren für die unmittelbaren Kandidatenbewertungen – ist die Beschränkung auf nur einen Datensatz nicht völlig unproblematisch. Zwar ist es unwahrscheinlich, dass sich die Datensätze aus ähnlichen TV-Duell-Studien in wesentlichen Charakteristika von den vorliegenden Daten unterscheiden. Da in dieser einzelnen Fallstudie ein Teil der Urteile über die Eignung der Verfahren auch daran festgemacht wird, dass wir substantiell sinnvolle Befunde erhalten, ist auch die Evaluation der Verfahren nicht völlig losgelöst vom Kontext des untersuchten TV-Duells zu betrachten. Jedoch achten wir gerade bei der Parametrisierung der Mehrebenenmodelle darauf, nur wenige Entscheidungen in Abhängigkeit vom vorliegenden Datensatz zu treffen. Grundsätzlich bevorzugen wir zudem verhältnismäßig sparsame Modelle, um eine Überanpassung zu vermeiden. Daher sind wir zuversichtlich, dass sich die Befunde zur Eignung der Modellklassen generalisieren lassen. Um dies zu erhärten, sind selbstverständlich weitere Replikationen anhand anderer TV-Duell-Studien wünschenswert. Ergänzend sind Simulationsstudien zu empfehlen, um insbesondere die statistischen Eigenschaften der sehr komplexen kreuzklassifizierten Wachstumskurvenmodelle eingehender zu erforschen.

Schließlich werden die Limitationen der Analyseverfahren bereits an anderer Stelle sehr ausführlich diskutiert, was an dieser Stelle nicht nochmals wiederholt werden soll. Die auch statistisch ausgearbeitete Kritik an den etablierten Verfahren findet sich in Kapitel 5. Grenzen und mögliche Erweiterungen der Mehrebenenmodelle zeigen wir in Kapitel 6.4. Im anschließenden Abschnitt gehen wir zusammenfassend auf die praktischen Konsequenzen dieser Auseinandersetzungen für die Beantwortung verschiedener Forschungsfragen anhand der Analyse unmittelbarer Kandidatenbewertungen in TV-Debatten ein.

Zusammenfassung mit Bezug auf die Ziele von RTR-Studien zu TV-Debatten

In Kapitel 2 werden fünf allgemeine Forschungsfragen benannt, die mit der Analyse der unmittelbaren Kandidatenbewertungen während der Debatte beantwortet werden können. Die Fragen unterscheiden sich danach, welche Inferenzschlüsse zu ihrer Beantwortung notwendig sind. Wenn wir an einer einfachen Deskription der Kandidatenbewertungen durch die Studienteilnehmer im Debattenverlauf interessiert sind (Frage 1), beispielsweise, um Hypothesen über Reaktionen auf bestimmte Merkmale des Debatteninhalts zu generieren,

so ist die verbreitete visuelle Inspektion der aggregierten RTR-Zeitreihen ein geeignetes Analysewerkzeug. Auch die Peak-Spike-Analyse zur Identifikation bemerkenswerter Debattenausschnitte ist hier hilfreich. Die ausführliche Diskussion und die Visualisierungen in Kapitel 5.2 offenbaren aber, dass mit der Betrachtung der aggregierten Zeitreihe lediglich Aussagen über Aggregate, also Gruppen von Personen, getroffen werden können. Der sich kontinuierlich verändernde Verlauf der RTR-Zeitreihen ist ausschließlich ein Aggregatphänomen. Die individuellen RTR-Verläufe verändern sich dagegen abrupt und behalten dann eine Ausprägung für einige Zeit bei. Während bei deskriptiven Auswertungen aus inferenzstatistischer Perspektive selbstverständlich nichts gegen die Inspektion der aggregierten RTR-Zeitreihen spricht, so ist es doch zu empfehlen, auch die individuellen RTR-Verläufe zumindest während wichtiger Passagen in die Betrachtung mit einzubeziehen. Dies gilt vor allem dann, wenn das explorative, hypothesengenerierende Vorgehen auch auf Erklärungen abzielt, die einen theoretischen Bezug zur individuellen Wahrnehmung und Verarbeitung der Debatteninhalte herstellen.

Häufig sollen die unmittelbaren Kandidatenbewertungen durch das Testpublikum (oder seine Teilgruppen) auch als ein Indikator für die Bewertung der Kandidaten durch das gesamte Publikum (oder seine Teilgruppen) genutzt werden (Frage 2). Hier ist ein klassischer Inferenzschluss von den Studienteilnehmern auf eine Grundgesamtheit außerhalb des Labors gefragt. Der Stimulus, auf den sich die Bewertungen beziehen, stimmt in Labor und Grundgesamtheit überein. Ein im eigentlichen Sinne repräsentativer Schluss ist in Anbetracht der bereits ausführlich dargestellten Einschränkungen bei der Ziehung der Personenstichprobe und der externen Validität der Studien nicht möglich. Wenn wir aber trotzdem einen Schluss von den beobachteten Bewertungen auf eine Grundgesamtheit zulassen, müssen wir dabei die statistische Unsicherheit berücksichtigen. Für die Beschreibung der Kandidatenbewertungen im Debattenverlauf ist dies einfach möglich, indem Konfidenzintervalle um die aggregierten RTR-Zeitreihen konstruiert werden. Deren Umfang ist erwartungsgemäß relativ weit, was die relativ geringen Fallzahlen einer RTR-Studie sowie die erhebliche Variabilität der individuellen Bewertungen innerhalb der Zuschauergruppen widerspiegelt. Mit den Intervallen können signifikante Abweichungen vom neutralen Skalenmittelpunkt und anderen als bedeutsam eingestuften Grenzwerten zu einem Zeitpunkt bestimmt werden. So kann beispielsweise die Frage beantwortet werden, während welcher Aussagen die Gruppe der Unentschiedenen einen Kandidaten signifikant im Vor- oder Nachteil gesehen hat. Wegen des Messwiederholungscharakters der RTR-Messungen ist es etwas schwieriger, die Kandidatenbewertungen durch eine Gruppe im Zeitverlauf unter Berücksichtigung der statistischen Unsicherheit zu vergleichen, um etwa

nach dem Peak-Spike-Ansatz die relativ am besten bewerteten Aussagen zu identifizieren. Wie wir in Kapitel 5.2.3 zeigen, kann dieses Problem jedoch durch ein Bootstrap-Verfahren umgangen werden. Für die Beantwortung von Fragen zur Bewertung der Kandidaten durch eine Grundgesamtheit anderer Rezipienten, bei denen die Inhalte der Debatte als gegeben angesehen werden, ist die Betrachtung der RTR-Zeitreihen und die darauf aufbauende Peak-Spike-Analyse ein nützliches Verfahren, wenn dabei die statistische Unsicherheit um die Zeitreihen berücksichtigt wird. Es muss allerdings beachtet werden, dass auch hier die Dynamik in den RTR-Zeitreihen nur als Aggregatphänomen interpretiert werden darf.

Die drei weiteren Forschungsfragen lösen sich nicht nur von der Personenstichprobe, sondern auch von der spezifischen Debatte, zu der die Bewertungen erfasst werden. Stattdessen sollen hier allgemeine Annahmen über die Effekte von bestimmten Charakteristika der politischen Botschaften am Beispiel des untersuchten Duells geprüft werden. Von Interesse sind hier die Effekte von Merkmalen der Aussagen dieser Kandidaten im Wahlkampf (Frage 3), Effekte von Merkmalen von Politikeraussagen in TV-Duellen im Allgemeinen (Frage 4), oder gar Effekte von Merkmalen (massenmedial vermittelter) (politischer) Botschaften im Allgemeinen (Frage 5). Gemeinsam ist diesen Forschungsfragen, dass sie neben dem Inferenzschluss auf eine Grundgesamtheit von Rezipienten auch einen Schluss auf eine Grundgesamtheit von Stimuli anstreben. Wie valide diese Inferenzschlüsse auf Basis der Inhalte einer TV-Debatte aus inhaltlicher Perspektive sein können, ist im Einzelfall zu klären (vgl. Kapitel 2). Damit sie aber überhaupt möglich sind, muss die Datenanalyse diese doppelte Stichprobenziehung und die damit einhergehende Unsicherheit auf Ebene der Rezipienten und auf Ebene der Inhalte berücksichtigen. Die Analysen aggregierter RTR-Messungen sind hierzu nicht in der Lage (vgl. Kapitel 5.3). Wenn die Kandidatenbewertungen zu Personenaggregaten zusammengefasst werden, gehen die Informationen verloren, welche zur Überprüfung von Annahmen über die Personenstichprobe hinaus notwendig sind. Formal betrachtet wird hier analysiert, wie das *untersuchte* Testpublikum andere Kandidatenaussagen in Abhängigkeit von deren Merkmalen bewerten würde. Dieser Inferenzschluss ist in der Medienwirkungsforschung jedoch nur selten relevant, da wir in aller Regel an Effekten auf Rezipienten im Allgemeinen interessiert sind. Zu dem gleichen Problem, jedoch aus Perspektive der Medieninhalte, führt das Zusammenfassen von Messzeitpunkten über Kandidatenaussagen hinweg. Hier wird vernachlässigt, dass die Rezipienten auf das Vorkommen desselben Merkmals in mehreren Aussagen unterschiedlich reagieren. Wie wir anhand der relativen Bewertung der Kandidaten in den Themenblöcken zeigen (vgl. Kapitel 6.2.2), führt die Vernachlässigung der Variabilität der Bewertungen zwischen den

Kandidatenaussagen zu einer Überschätzung der Präzision, mit der die mittlere Kandidatenbewertung festgestellt werden kann. Insgesamt müssen wir zu dem Schluss kommen, dass die bisher zur deduktiven Analyse eingesetzten Verfahren „zu falsch“ sind, um für die Prüfung von allgemeinen Annahmen über Effekte von Rezipienten- und Inhaltsmerkmalen nützlich zu sein.

Bei der Beantwortung der Fragen, die Inferenzen auf Rezipienten- und auf Inhaltsebene erfordern, kommen die Stärken der in Kapitel 6.2 und 6.3 vorgestellten kreuzklassifizierten (Wachstumskurven-) Modelle zu tragen. Die kreuzklassifizierte Struktur berücksichtigt explizit, dass die Messungen in zwei Stichprobeneinheiten – Rezipienten und Kandidatenaussagen – geschachtelt sind. Da für beide Merkmalsträger gleichzeitig Prädiktoren in den Modellen berücksichtigt werden können, ist es möglich, Annahmen über die Erklärung der unmittelbaren Kandidatenbewertungen durch Eigenschaften der Rezipienten, Eigenschaften der Kandidatenaussagen und spezifische Interaktionen zwischen den Eigenschaften von Rezipienten und Kandidatenaussagen zu prüfen. Wenn zudem Annahmen über die dynamischen Veränderungen der individuellen Kandidatenbewertungen getestet werden sollen, ist für die Untersuchung einer einzelnen Aussage das einfache Wachstumskurvenmodell (vgl. Kapitel 6.1), für die systematische Untersuchung vieler Aussagen das kreuzklassifizierte Wachstumskurvenmodell geeignet. Dabei sind jedoch zwei wichtige Einschränkungen zu beachten: Erstens ist die Form der relativ sparsam spezifizierten Wachstumskurven nur eine sehr vereinfachte Annäherung. Von ihr sollte nicht direkt auf die Form der beobachteten individuellen Verläufe geschlossen werden. In dieser Hinsicht sind auch die (kreuzklassifizierten) Wachstumskurvenmodelle noch relativ falsche Modelle der individuellen Dynamiken. Sie stellen aber immerhin eine nützliche Approximation für die weiteren Analysen zur Verfügung. Zweitens bewährt sich nur die Analyse der dynamischen Bewertungsveränderungen während der direkten Antworten der Kandidaten auf Fragen der Moderatoren – und damit nur während bestimmter Ausschnitte der Debatte. Wenn wie in Frage 2 die Kandidatenbewertung während der gesamten Debatte relevant ist, sind solche Analysen nur von begrenztem Nutzen.

Abschließend ist noch auf einen Befund hinzuweisen, der die Analysen zu allen genannten Fragen und mit allen diskutierten Verfahren betrifft. Durchweg zeigt sich, dass die unmittelbaren Kandidatenbewertungen sehr stark von den Voreinstellungen der Rezipienten beeinflusst werden. Dieses Ergebnis ist nicht neu, sondern wurde in der empirischen Literatur zur Erklärung der unmittelbaren Kandidatenbewertungen immer wieder herausgestellt (vgl. Kapitel 3.4). Allerdings wurden daraus nicht immer Konsequenzen für die detaillierteren Analysen der Kandidatenbewertungen in Abhängigkeit von den Debatteninhalten gezogen. So finden sich zahlreiche Auswertungen (inklusive unserer

eigenen, z.B. Bachl, 2013a; Bachl & Brettschneider, 2013; Brettschneider & Bachl, 2012), in denen die unmittelbaren Bewertungen auch ohne Beachtung der Voreinstellungen untersucht werden. Auf Grundlage der in dieser Arbeit präsentierten Befunde ist von solchen Analysen, gleich mit welchem Verfahren, abzuraten. Die Ergebnisse zeigen deutlich, dass die Debatteninhalte in Interaktion mit den Voreinstellungen der Rezipienten wirken. Aggregierte RTR-Zeitreihen sollten daher nur für Teilstichproben nach den politischen Voreinstellungen der Rezipienten gebildet werden. Um in diesen Analysen Interaktionen zu erkennen, müssen die Veränderungen der einzelnen Zeitreihen nicht nur isoliert, sondern auch im Kontrast zu den anderen Zeitreihen betrachtet werden. In den deduktiven Verfahren müssen Interaktionen zwischen Voreinstellungen der Rezipienten und Merkmalen der Kandidatenaussagen explizit modelliert werden. Hier erweisen sich die kreuzklassifizierten (Wachstumskurven-) Modelle als besonders flexibel, da die Voreinstellungen mit mehreren und quasi-metrisch skalierten Variablen detaillierter in die Modelle einfließen können und die Interaktionen explizit getestet werden.

Die zwei wesentlichen Befunde bezüglich der Frage, welche Verfahren sich für die Beantwortung welcher Fragen mit RTR-Studien zu TV-Debatten als nützlich erweisen, sind hier nochmals komprimiert zusammengefasst:

- Für Analysen, die rein deskriptiv angelegt sind oder nur einen Inferenzschluss in Richtung anderer Rezipienten der untersuchten Debatte anstreben, sind die nach wichtigen Voreinstellungen aggregierten Zeitreihen der Kandidatenbewertungen weiterhin nützlich. Sie müssen aber entsprechend des Aggregatniveaus der Daten auch inhaltlich auf Aggregatniveau interpretiert werden. Für eine Annäherung an die individuellen Urteilsprozesse während der Debatte ist zusätzlich eine visuelle Inspektion der individuellen RTR-Verläufe zu empfehlen. Der Inferenzschluss in Richtung einer Grundgesamtheit von Rezipienten der Debatte muss darüber hinaus die statistische Unsicherheit der Daten berücksichtigen – z.B. durch Konfidenzintervalle um die Zeitreihen oder mit einem Bootstrap-Verfahren zur Identifikation relativer Peaks.

- Für Analysen, die Annahmen über die (kombinierten) Effekte von Eigenschaften der Rezipienten und der Kandidatenaussagen prüfen, können die auf einer Aggregation über Personen oder Messzeitpunkte aufbauenden Verfahren nicht weiter empfohlen werden. Mehrebenenmodelle der unmittelbaren Kandidatenbewertungen sind hier vielversprechende Alternativen: das einfache Wachstumskurvenmodell zur Untersuchung der individuellen Veränderungen der Bewertungen während einer Aussage;

das kreuzklassifizierte Modell zur Untersuchung der individuellen summarischen Bewertungen während vieler Aussagen; und das kreuzklassifizierte Wachstumskurvenmodell zur Untersuchung der individuellen Veränderungen der Bewertungen während vieler Aussagen. Dabei ist zu empfehlen, die Effekte der Inhalte konditional von den Voreinstellungen der Rezipienten zu modellieren.

Ausblick: Übertragung der Befunde auf andere RTR-Studien

Die Sichtung der Studien mit RTR-Messungen zeigt, dass die Technik der RTR-Messung in wesentlichen Eigenschaften von der hier behandelten Messung mit RTR-Dails im *latched mode* abweichen kann (vgl. Kapitel 3.1). Auch werden RTR-Studien zu anderen Fragestellungen jenseits der TV-Debatten durchgeführt, und es werden andere Designs mit anderen Anordnungen der Stimuli verwendet (vgl. Kapitel 3.2). Zum Abschluss der Arbeit geben wir einen Ausblick, wie die Befunde dieser Arbeit auch in anderen Kontexten nutzbar gemacht werden können. Dabei konzentrieren wir uns auf die in Kapitel 6 vorgestellten Mehrebenenmodelle. Die Erweiterungen der Peaks-Spike-Analyse aus Kapitel 5 können ohne weiteres auf Anwendungen, in denen dieses Analyseverfahren bisher schon eingesetzt wurde, übertragen werden.

Analyse von mit Push-Button-Geräten erfassten RTR-Messungen Neben Dial- und Slider-Geräten, deren Funktionsweise und erzeugte Datenstruktur sich stark ähneln, kommen häufiger auch Push-Button-Geräte zur rezeptionsbegleitenden Messung zum Einsatz. Die Daten einer RTR-Messung mit Push-Button-Geräten weisen auf der individuell-längsschnittlichen Ebene eine binäre Struktur auf. Für jede Messeinheit (z.B. eine Sekunde), jeden Probanden und jeden Knopf des Geräts ist die Information gespeichert, ob der Knopf gerade gedrückt ist oder nicht. In den individuellen Verläufen existiert daher kein Wachstum im Sinne einer monotonen Veränderung. Daher kann für die Rohdaten kein Wachstumskurvenmodell geschätzt werden. Wenn die grundsätzliche Analyselogik, in der die einzelnen Messungen als gleichzeitig in Rezipienten und Stimuluseinheiten (hier: Antworten oder Turns) gruppiert betrachtet werden, beibehalten werden soll, bieten sich zwei einfache Möglichkeiten an. Zum einen kann das in Kapitel 6.2 beschriebene einfache kreuzklassifizierte Modell zu einem verallgemeinerten linearen Modell für dichotome Daten erweitert werden (vgl. z.B. Hox, 2010, Kap. 6). So würde in Abhängigkeit von Eigenschaften der Rezipienten und der Stimuluseinheiten die Wahrscheinlichkeit geschätzt, dass ein Knopf in einer Sekunde gedrückt ist. Zum anderen können die Messungen innerhalb der Kombinationen von Rezipienten und Stimuluseinheiten

transformiert werden, um den Modellen die zeitliche Dynamik der Messungen innerhalb der Kombinationen zugänglich zu machen. Denkbar wäre hier die Bildung von über die Zeit kumulierten Summen oder Saldi der Messungen. Diese Maße würden dann eine stetige Veränderung aufweisen und könnten mit (kreuzklassifizierten) Wachstumskurvenmodellen beschrieben werden. Da die Push-Button-Geräte im *reset mode* messen (vgl. Kapitel 3.1), eignen sie sich womöglich besser für die Untersuchung von Urteilsveränderungen während nicht formal definierter Einheiten des Stimulus. Es wäre zu prüfen, ob eine Analyse wie die der Veränderungen der Kandidatenbewertungen infolge von Relationswechseln mit solchen Messungen ertragreicher ist.

Analyse von anderen mit RTR-Messungen erfassten Konstrukten Die in Kapitel 3.2 vorgestellten Konstrukte können wir hinsichtlich zweier Dimensionen ordnen. Zum einen ist zu differenzieren, ob die Rezipienten ihren Eindruck vom gesamten bzw. von bestimmten Teilen des Stimulus wiedergeben, oder ob die RTR-Messungen einer Selbstbeobachtung, z.B. bezüglich des emotionalen Erlebens, dienen. Zum anderen kann zwischen allgemeiner formulierten Vorgaben, z.B. zum allgemeinen Eindruck oder zum allgemeinen Wohlbefinden, und spezifischeren Vorgaben, z.B. zum Eindruck von der Verständlichkeit oder zum Empfinden einer bestimmten Emotion, unterschieden werden. Ob die in dieser Arbeit behandelten Mehrebenenmodelle taugen, um auch RTR-Messungen anderer Konstrukte zu untersuchen, ist davon abhängig, ob die individuellen Verläufe sich wesentlich von den hier untersuchten unmittelbaren allgemeinen Eindrücken von Politikern unterscheiden. Da uns selbst keine RTR-Messungen zu anderen Konstrukten vorliegen und in Publikationen kaum auf die Charakteristika der individuellen RTR-Verläufe eingegangen wird, können wir an dieser Stelle nur eine begründete Spekulation anstellen. Wir vermuten, dass die individuellen Verläufe stärker und mit einer höheren Frequenz oszillieren, wenn sie eine Selbstbeobachtung erfassen und wenn sie ein spezifischeres Konstrukt repräsentieren. So ist beispielsweise davon auszugehen, dass sich der emotionale Zustand einer Person häufig und schnell ändern kann, wenn ein Stimulus eine Emotion variabel induziert (Larsen & Fredrickson, 1999).

Wenn es für die zu beantwortende Forschungsfrage ausreichend ist, die Messungen summarisch über vorgegebene Einheiten des Stimulus zu betrachten, ist das kreuzklassifizierte Modell mit seiner Logik, einfach die mittlere Ausprägung der individuellen Messungen während einer Einheit zu schätzen, technisch zur Analyse geeignet. Allerdings ist der geschätzte Mittelwert umso weniger aussagekräftig für eine individuelle Messung, je stärker diese während der Einheit schwankt. Problematischer ist die Übertragung von Modellen, die

wie das (kreuzklassifizierte) Wachstumskurvenmodell auch die Veränderungen innerhalb der Stimuluseinheiten beschreiben sollen. Bereits für die interindividuell relativ einheitlichen Kandidatenbewertungen während der Antworten ist das einfache Wachstumskurvenmodell mit latentem Intercept und latentem Slope eine zwar pragmatisch nützliche, jedoch trotzdem nur sehr ungenaue Annäherung. Sollte es das Ziel einer Arbeit sein, innerhalb der Stimuluseinheiten oszillierende individuelle Verläufe systematisch zu beschreiben, müsste statt des Wachstumskurvenmodells ein nicht-lineares Modell entsprechend des erwarteten Verlaufs spezifiziert werden. Dies ist, wie in Kapitel 6.4 kurz angerissen, statistisch recht einfach, dürfte sich in der Anwendung jedoch als Herausforderung erweisen. Die in der vorliegenden Arbeit diskutierten Wachstumskurvenmodelle eignen sich vor allem für die Beschreibung monotoner Veränderungsprozesse innerhalb der Kombinationen von Rezipienten und Stimuluseinheiten. Diese dürften sich wahrscheinlicher bei allgemeiner formulierten RTR-Items zum Eindruck der Rezipienten vom Stimulus zeigen.

Analyse von RTR-Messungen aus experimentellen Designs und Messwiederholungsdesigns mit einem oder mehreren Stimuli In Kapitel 3.2 zeigen wir verschiedene Ansätze, um in RTR-Studien Effekte von Stimulusmerkmalen auf die rezeptionsbegleitend erfassten Konstrukte zu untersuchen. Zum einen kann das relevante Merkmal des Stimulus experimentell variiert werden. Zum anderen werden Variationen relevanter Merkmale im Stimulus mit Variationen in den RTR-Messungen verknüpft, um die Eigenschaft der RTR-Messungen als Messwiederholungen desselben Konstrukts bei denselben Probanden zu unterschiedlichen Stimulusmerkmalen auszunutzen. Die Erweiterung der vorgestellten Modelle um einen Test des Effekts einer experimentellen Variation ist einfach. Angenommen, wir hätten in unserer Studie bei einer Experimentalgruppe die Videoübertragung ausgeblendet, sodass sie das TV-Duell lediglich hören konnte. Um zu prüfen, ob der Rezeptionsmodus einen Effekt auf die Bewertung von Mappus während der in Kapitel 6.1 untersuchten Antwort hat, würde der Rezeptionsmodus als zusätzliches Merkmal im Modell berücksichtigt. Der Koeffizient des Merkmals und sein inferenzstatistischer Test geben dann Auskunft über Richtung und Signifikanz des Effekts der Manipulation.

Um Variationen relevanter Merkmale im Stimulusinhalt mit den RTR-Messungen zu verknüpfen und damit das Messwiederholungsdesign zu nutzen, haben wir uns für eine Zuordnung der Merkmale auf Ebene der Stimuluseinheiten Antwort, Turn und Relationswechsel entschieden. Die Analyse der Veränderung der unmittelbaren Kandidatenbewertungen nach den Relationswechseln erweist sich jedoch zumindest im hier untersuchten Beispiel als

problematisch. Die Befunde deuten darauf hin, dass eine auf inhaltliche Kriterien zurückgehende Identifikation der Einheiten von den Rezipienten nicht nachvollzogen wird. Dagegen hat sich die Einheit der Antwort zur Analyse von Bewertungsveränderungen ausgehend von einem neutralen Ausgangswert bewährt. Die Turns eignen sich aufgrund ihrer klaren formalen Definition ebenfalls gut zur Abgrenzung der einzelnen Einheiten. Da ihre Länge jedoch teils deutlich variiert und sich innerhalb der Einheiten sehr unterschiedliche Formen von individuellen RTR-Verläufen finden, können die Veränderungsprozesse während der Turns nicht durch einfache Wachstumskurven beschrieben werden. Daher analysieren wir hier nur die summarische Bewertung.

Folglich lassen sich die hier vorgestellten Modelle besonders gut auf die Analyse von RTR-Bewertungen während formal voneinander abgegrenzten Stimuli übertragen. Studien, die Reaktionen auf eine größere Zahl von Stimuli im Spotformat – z.B. kommerzielle oder politische Werbespots – untersuchen, eignen sich ideal für die Analyse mit kreuzklassifizierten Wachstumskurvenmodellen. Wir können davon ausgehen, dass die Rezipienten zu Beginn jedes Spots einen neutralen Eindruck haben, da sie noch keinerlei Informationen über den Stimulus besitzen. Ausgehend von diesem Punkt können die individuellen RTR-Verläufe durch relativ einfache Wachstumskurven abgebildet werden, da Spots meist relativ kurz sind und eine durchgängige, in sich geschlossene Botschaft vermitteln. Auch zur Analyse der RTR-Messungen während anderer multiepisodischer Stimuli, wie beispielsweise einer Nachrichtensendung mit mehreren Beiträgen, sollte diese Modellklasse gut geeignet sein. Wenn eine solche Analyse angestrebt wird, muss jedoch bereits bei der Planung der Studie berücksichtigt werden, dass in diesen Modellen auch die Stimuli – hier also die Spots bzw. Nachrichtenbeiträge – als Stichproben aus einer größeren Grundgesamtheit behandelt werden. Auch diese Stichprobe muss einen entsprechenden Umfang haben, um abgesicherte Schlüsse zu ermöglichen. Allerdings ist es nicht notwendig, dass alle Rezipienten auch alle Stimuli sehen, was bei einer Studie zu einer Nachrichtensendung mit mehreren Beiträgen die externe Validität gefährden würde. Stattdessen könnten die Beiträge zu mehreren Nachrichtensendungen zusammengestellt werden, die jeweils von einer Rezipientengruppe gesehen werden. In einem entsprechenden Mehrebenenmodell würde dann berücksichtigt, dass zuerst Beiträge in Sendungen und dann RTR-Messungen in Beiträgen geschachtelt sind, wobei die RTR-Messungen in den Sendungen von unterschiedlichen Rezipienten stammen.

Schwieriger dürfte sich die Übertragung der Modelle auf die Analyse von RTR-Messungen während längerer, formal nur wenig strukturierter Stimuli darstellen. Aus datenanalytischer Perspektive können die kreuzklassifizierten Wachstumskurvenmodelle auch dann eingesetzt werden, wenn die RTR-

Messungen bereits zu Beginn einer Stimuluseinheit variieren (vgl. Kapitel 6.3.3). Wenn aber die Rezipienten die Abgrenzung der Einheiten nicht wahrnehmen oder die Einteilung für sie nicht relevant ist, kann dies zu inhaltlich invaliden Ergebnissen führen. Als Beispiel sei an eine Studie gedacht, in der das Unterhaltungserleben während der Rezeption eines Spielfilms untersucht werden soll. Zu prüfen wäre die Annahme, dass sich das Unterhaltungserleben in Abhängigkeit von Eigenschaften der Rezipienten und der Szenen erklären lässt. Voraussetzung dafür, dass die Veränderung des Unterhaltungserlebens von einer Szene auf die andere durch ein kreuzklassifiziertes Wachstumskurvenmodell erklärt werden kann, ist, dass die Rezipienten einen Szenenwechsel wahrnehmen und als bedeutsam für ihr Unterhaltungserleben empfinden. Ob das überhaupt eine theoretisch plausible Annahme ist, wäre von einschlägigen Unterhaltungsforschern zu beantworten. Dies wäre dann auch mit entsprechenden Studien zu prüfen. In Ermangelung solcher Daten können wir an dieser Stelle unter Kenntnis der Urteilsveränderungen während der Rezeption eines TV-Duells lediglich Zweifel daran äußern, dass individuelle Rezeptionsprozesse sich nach formal nicht oder nur ungenau zu erkennenden Einheiten strukturieren lassen. Für Analysen, bei denen die RTR-Messungen sich besser als Abbildung eines kontinuierlichen Prozesses über den gesamten, längeren Stimulus auffassen lassen, erscheint die (sicherlich schwierigere) Übertragung des zeitreihenanalytischen Vorgehens von Nagel (2012) auf die Ebene der individuellen RTR-Verläufe angemessener.

Abschließend wollen wir der Hoffnung Ausdruck verleihen, dass diese Arbeit dazu beitragen kann, die RTR-Messungen der individuellen Rezipienten stärker in den Fokus der empirischen Kommunikationsforschung zu rücken – unabhängig von der Technik der RTR-Messung, den gemessenen Konstrukten oder dem Design der Studien. Kommunikationswissenschaftliche Ansätze zu Prozessen der Wahrnehmung und Verarbeitung von kontinuierlichen Medienstimuli setzen häufig auf der Ebene der individuellen Rezipienten an. Das große Potenzial der RTR-Methode ist, dass sie Daten erhebt, die in ihrer längsschnittlichen und individuellen Messung der Theorieebene angemessen sind (Biocca et al., 1994). Auch die statistische Beschreibung und Erklärung der so erfassten Daten muss aber entsprechend dem Individualniveau der Theorien und Messungen erfolgen, um das Potenzial der Methode voll ausschöpfen zu können. Einige Verfahren dazu haben wir in dieser Arbeit diskutiert, weitere Entwicklungen sind darüber hinaus wünschenswert. Doch auch die Beschreibung und Darstellung der aggregierten RTR-Messungen kann je nach Forschungsfrage nützlich sein, um Prozesse und Veränderungen auf Gruppenebene deutlich zu machen. In diesem Fall muss mit der Aggregation der Messungen jedoch auch eine Verlagerung der theoretischen Ansätze auf die Gruppenebene stattfinden.

Literatur

Aaker, D. A., Stayman, D. M. & Hagerty, M. R. (1986). Warmth in advertising: Measurement, impact, and sequence effects. *Journal of Consumer Research*, *12* (4), 365-381. doi: 10.2307/254299

Abeele, P. V. & MacLachlan, D. L. (1994). Process tracing of emotional responses to TV ads: Revisiting the Warmth Monitor. *Journal of Consumer Research*, *20* (4), 586-600. doi: 10.2307/2489761

Algie, J. & Rossiter, J. R. (2010). Fear patterns: A new approach to designing road safety advertisements. *Journal of Prevention & Intervention in the Community*, *38* (4), 264-279. doi: 10.1080/10852352.2010.509019

Aristoteles. (Übers. 2007). *Rhetorik* (2. Aufl.; G. Krapinger, Übers.). Stuttgart: Reclam.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66* (3), 411-421. doi: 10.1097/01.psy.0000127692.23278.a9

Bacherle, P., Schneider, F. M. & Krause, S. (2012, Oktober). *Continuous response measurement from a bird's-eye view: Integrating evidence for validity across 13 studies.* Vortrag auf der 4. European Communication Conference der ECREA, Istanbul.

Bachl, M. (2013a). Die Wahrnehmung des TV-Duells. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 135-169). Wiesbaden: VS.

Bachl, M. (2013b). Die Wirkung des TV-Duells auf die Bewertung der Kandidaten und die Wahlabsicht. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 171-198). Wiesbaden: VS.

Bachl, M. & Brettschneider, F. (2011). The German national election campaign and the mass media. *German Politics*, *20* (1), 51-74. doi: 10.1080/09644008.2011.554100

Bachl, M. & Brettschneider, F. (2013). Das TV-Duell Mappus gegen Schmid – Wahrnehmung und Wirkungen. In U. Wagschal, U. Eith & M. Wehner (Hrsg.), *Der historische Machtwechsel: Grün-Rot in Baden-Württemberg* (S. 93-118). Baden-Baden: Nomos.

Bachl, M., Brettschneider, F., Kercher, J., Spieker, A. & Vögele, C. (2013a). Befragung 1 vor der Duell-Rezeption. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (Online-Anhang). Wiesbaden: VS.

Bachl, M., Brettschneider, F., Kercher, J., Spieker, A. & Vögele, C. (2013b). Befragung 2 nach der Duell-Rezeption. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (Online-Anhang). Wiesbaden: VS.

Bachl, M., Brettschneider, F. & Ottler, S. (2013a). *Das TV-Duell in Baden-Württemberg 2011*. Wiesbaden: VS.

Bachl, M., Brettschneider, F. & Ottler, S. (2013b). Die TV-Duell-Studie Baden-Württemberg 2011. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 7-27). Wiesbaden: VS.

Bachl, M., Käfferlein, K., Krafft, A., Schmalz, I. & Vögele, C. (2013). Transkript und Echtzeitbewertung des TV-Duells Mappus gegen Schmid vor der Landtagswahl 2011 in Baden-Württemberg. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (Online-Anhang). Wiesbaden: VS.

Bachl, M., Käfferlein, K. & Spieker, A. (2013a). Codebuch zur Inhaltsanalyse des TV-Duells. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (Online-Anhang). Wiesbaden: VS.

Bachl, M., Käfferlein, K. & Spieker, A. (2013b). Die Inhalte des TV-Duells. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 57-86). Wiesbaden: VS.

Bachl, M. & Spieker, A. (2010, Juli). *Opening the ‚black-box': Exploring immediate audience responses to rhetorical strategies in televised debates.* Vortrag auf der Jahrestagung der International Association for Media and Communication Research (IAMCR), Braga.

Bachl, M. & Vögele, C. (2013). „Ich habe die Möglichkeiten in diesem großartigen Land bekommen durch eine tolle Bildung". Inhalte, Wahrnehmung und Wirkungen des bildungspolitischen Debattenteils im TV-Duell vor der Landtagswahl 2011 in Baden-Württemberg. *Studies in Communication | Media*, *2* (3), 367-400.

Baggaley, J. (1987). Continual response measurement: Design and validation. *Canadian Journal of Educational Communication*, *16* (3), 217-38.

Baggaley, J., Salmon, C., Siska, M., Lewis-Hardy, R., Tambe, P. B., Jorgensen, C., … Jason, J. (1992). Automated evaluation of AIDS messages with high-risk, low-literacy audiences. *Journal of Educational Television*, *18* (2-3), 83-95. doi: 10.1080/0260741920180202

Bates, D. (2013a). Computational methods for mixed models. *R Package Vignette*. doi: http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf

Bates, D. (2013b). Linear mixed model implementation in lme4. *R Package Vignette*. doi: http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf

Bates, D., Maechler, M. & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes (R package version 0.999999-0, http://CRAN.R-project.org/package=lme4) [Software].

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-0, http://lme4.r-forge.r-project.org/) [Software].

Baumgartner, H., Sujan, M. & Padgett, D. (1997). Patterns of affective reactions to advertisements: The integration of moment-to-moment responses into overall judgments. *Journal of Marketing Research*, *34* (2), 219-232. doi: 10.2307/3151860

Benoit, K., Laver, M. & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, *53* (2), 495-513. doi: 10.1111/j.1540-5907.2009.00383.x

Benoit, W. L. (1999). *Seeing spots: A functional analysis of presidential television advertisements, 1952-1996*. Westport: Praeger.

Benoit, W. L. (2007). *Communication in political campaigns*. New York: Peter Lang.

Benoit, W. L. (2013). *Political election debates: Informing voters about policy and character*. Plymouth: Lexington Books.

Benoit, W. L. & Airne, D. (2005). A functional analysis of American vice presidential debates. *Argumentation & Advocacy*, *41* (4), 225-236.

Benoit, W. L. & Benoit-Bryan, J. M. (2013). Debates come to the United Kingdom: A functional analysis of the 2010 British prime minister election debates. *Communication Quarterly*, *61* (4), 463-478. doi: 10.1080/01463373.2013.799513

Benoit, W. L., Blaney, J. R. & Pier, P. (1998). *Campaign '96: A functional analysis of acclaiming, attacking, and defending*. Westport: Praeger.

Benoit, W. L., Blaney, J. R. & Pier, P. M. (2000). Acclaiming, attacking, and defending: A functional analysis of U.S. nominating convention keynote speeches. *Political Communication*, *17* (1), 61-84. doi: 10.1080/105846000198512

Benoit, W. L. & Brazeal, L. M. (2002). A functional analysis of the 1988 Bush-Dukakis presidential debates. *Argumentation & Advocacy*, *38* (4), 219-233.

Benoit, W. L. & Hansen, G. J. (2001). Presidential debate questions and the public agenda. *Communication Quarterly*, *49* (2), 130-141. doi: 10.1080/01463370109385621

Benoit, W. L., Hansen, G. J. & Verser, R. M. (2003). A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs*, *70* (4), 335-350. doi: 10.1080/0363775032000179133

Benoit, W. L. & Harthcock, A. (1999). Functions of the great debates: Acclaims, attacks, and defenses in the 1960 presidential debates. *Communication Monographs*, *66* (4), 341-357. doi: 10.1080/03637759909376484

Benoit, W. L. & Henson, J. R. (2007). A functional analysis of the 2006 Canadian and 2007 Australian election debates. *Argumentation & Advocacy*, *44* (1), 36-48.

Benoit, W. L. & Klyukovski, A. A. (2006). A functional analysis of 2004 Ukrainian presidential debates. *Argumentation*, *20* (2), 209-225. doi: 10.1007/s10503-006-9007-x

Benoit, W. L., McHale, J. P., Hansen, G. J., Pier, P. M. & McGuire, J. P. (2003). *Campaign 2000: A functional analysis of presidential campaign discourse.* Lanham: Rowman & Littlefield.

Benoit, W. L., Pier, P. M., Brazeal, L. M., McHale, J. P., Klyukovski, A. & Airne, D. (2002). *The primary decision: A functional analysis of debates in presidential primaries*. Westport: Praeger.

Benoit, W. L. & Sheafer, T. (2006). Functional theory and political discourse: Televised debates in Israel and the United States. *Journalism & Mass Communication Quarterly*, *83* (2), 281-297. doi: 10.1177/107769900608300204

Benoit, W. L., Wen, W.-C. & Yu, T.-h. (2007). A functional analysis of 2004 Taiwanese political debates. *Asian Journal of Communication*, *17* (1), 24-39. doi: 10.1080/01292980601114521

Beretvas, S. N. (2010). Cross-classified and multiple membership models. In J. Hox & J. Roberts (Hrsg.), *The handbook of advanced multilevel analysis* (S. 313-334). New York: Routledge.

Beretvas, S. N. & Murphy, D. L. (2013). An evaluation of information criteria use for correct cross-classified random effects model selection. *The Journal of Experimental Education*, *81* (4), 429-463. doi: 10.1080/00220973.2012.745467

Biocca, F., David, P. & West, M. (1994). Continuous response measurement (CRM): A computerized tool for research on the cognitive processing of communication messages. In A. Lang (Hrsg.), *Measuring psychological responses to media messages* (S. 15-64). Hillsdale: Routledge.

Blais, A. & Perrella, A. M. L. (2008). Systemic effects of televised candidates' debates. *The International Journal of Press/Politics*, *13* (4), 451-464. doi: 10.1177/1940161208323548

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation* (4. Aufl.). Heidelberg: Springer Medizin.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und aktualisierte Aufl.). Heidelberg: Springer.

Bos, A. L., van Doorn, B. W. & Smanik, A. C. (2012). The effects of HDTV on perceptions of Obama and McCain in a 2008 presidential debate. *Communication Research Reports*, *29* (2), 161-168. doi: 10.1080/08824096.2012.666769

Box, G. E. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Bradley, S. D. (2007). Examining the eyeblink startle reflex as a measure of emotion and motivation to television programming. *Communication Methods and Measures*, *1* (1), 7-30. doi: 10.1080/19312450709336658

Brambor, T., Clark, W. R. & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14* (1), 63-82. doi: 10.1093/pan/mpi014

Brettschneider, F. (2005a). Bundestagswahlkampf und Medienberichterstattung. *Aus Politik und Zeitgeschichte* (51/52), 19-26.

Brettschneider, F. (2005b). Massenmedien und Wählerverhalten. In J. Falter & H. Schoen (Hrsg.), *Handbuch Wahlforschung* (S. 473-500). Wiesbaden: VS.

Brettschneider, F. & Bachl, M. (2009). Die Bundestagswahl 2009 und die Medien. *Politische Studien*, *60*, 46-55.

Brettschneider, F. & Bachl, M. (2012). Das TV-Duell Mappus gegen Schmid – Wahrnehmung und Wirkungen. *Der Bürger im Staat*, *62* (3), 141-148.

Brittin, R. V. (1996). Listeners' preference for music of other cultures: Comparing response modes. *Journal of Research in Music Education*, *44* (4), 328-340. doi: 10.2307/3345445

Bryk, A. S. & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101* (1), 147-158. doi: 10.1037/0033-2909.101.1.147

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33* (2), 261-304. doi: 10.1177/0049124104268644

Carlin, D. B., Morris, E. & Smith, S. (2001). The influence of format and questions on candidates' strategic argument choices in the 2000 presidential debates. *American Behavioral Scientist*, *44* (12), 2196-2218. doi: 10.1177/00027640121958276

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39* (5), 752-766. doi: 10.1037/0022-3514.39.5.752

Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). *Graphical methods for data analysis*. Monterey: Wadsworth.

Choi, Y. S. & Benoit, W. L. (2013). A functional analysis of the 2007 and 2012 French presidential debates. *Journal of Intercultural Communication Research*, *42* (3), 215-227. doi: 10.1080/17475759.2013.827584

Chou, C., Bentler, P. M. & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5* (3), 247-266. doi: 10.1080/10705519809540104

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78* (1), 98-104. doi: 10.1037/0021-9010.78.1.98

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38* (4), 529-569. doi: 10.1207/s15327906mbr3804_5

Dalakas, V. (2006a). The effect of cognitive appraisals on emotional responses during service encounters. *Services Marketing Quarterly*, *27* (1), 23-41. doi: 10.1300/J396v27n01_02

Dalakas, V. (2006b). The importance of a good ending in a service encounter. *Services Marketing Quarterly*, *28* (1), 35-53. doi: 10.1300/J396v28n01_03

Davis, C. J., Bowers, J. S. & Memon, A. (2011). Social influence in televised election debates: A potential distortion of democracy. *PLoS ONE*, *6* (3), e18154. doi: 10.1371%2Fjournal.pone.0018154

Dehm, U. (2002). Fernsehduelle im Urteil der Zuschauer. Eine Befragung des ZDF zu einem neuen Sendungsformat bei der Bundestagswahl 2002. *Media Perspektiven*, *o.Jg.* (12), 600-609.

de Leeuw, J. & Meijer, E. (2008). *Handbook of multilevel analysis*. New York: Springer.

Delli Carpini, M. X., Keeter, S. & Webb, S. (1997). The impact of presidential debates. In P. Norris (Hrsg.), *Politics and the press: The news media and their influences* (S. 145-164). Boulder: Rienner.

Donsbach, W. (2002). Sechs Gründe gegen Fernsehduelle. *Die politische Meinung*, *47* (396), 19-25.

Donsbach, W., Jandura, O. & Hastall, M. (2002). Neues aus der Fernsehdemokratie - Wahrnehmung und Wirkung des ersten TV-Duells. In H. Oberreuther (Hrsg.), *Der versäumte Wechsel. Eine Bilanz des Wahljahres* (S. 136-156). München: Olzog.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54-75.

Efron, B. & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, *253* (5018), 390-395. doi: 10.1126/science.253.5018.390

Egermann, H., Nagel, F., Altenmüller, E. & Kopiez, R. (2009). Continuous measurement of musically-induced emotion: A web experiment. *International Journal of Internet Science*, *4* (1), 4-20.

Faas, T. & Huber, S. (2010). Experimente in der Politikwissenschaft: Vom Mauerblümchen zum Mainstream. *Politische Vierteljahresschrift*, *51* (4), 721-749. doi: 10.1007/s11615-010-0039-3

Faas, T. & Maier, J. (2004a). Mobilisierung, Verstärkung, Konversion? Ergebnisse eines Experiments zur Wahrnehmung der Fernsehduelle im Vorfeld der Bundestagswahl 2002. *Politische Vierteljahresschrift*, *45* (1), 55-72. doi: 10.1007/s11615-004-0004-0

Faas, T. & Maier, J. (2004b). Schröders Stimme, Stoibers Lächeln: Wahrnehmungen von Gerhard Schröder und Edmund Stoiber bei Sehern und Hörern der Fernsehdebatten im Vorfeld der Bundestagswahl 2002. In T. Knieper & M. Müller (Hrsg.), *Visuelle Wahlkampfkommunikation* (S. 186-209). Köln: von Halem.

Faas, T., Maier, J., Maier, M. & Brettschneider, F. (2009, September). *Das TV-Duell 2009.* Vortrag auf der 24. Jahrestagung der Deutschen Gesellschaft für Politische Wissenschaft, Kiel.

Fahr, A. (2006). Fernsehen fühlen. Ein Ansatz zur Messung von Rezeptionsemotionen. In W. Wirth, H. Schramm & V. Gehrau (Hrsg.), *Unterhaltung durch Medien: Theorien und Messung* (S. 204-226). Köln: Halem.

Fahr, A. (2008). Real-Time ratings (RTR). In W. Donsbach (Hrsg.), *The international encyclopedia of communication* (S. 4121-4124). Malden: Blackwell Publishing.

Fahr, A. (2009). *Politische Talkshows aus Zuschauersicht*. Baden-Baden: Nomos.

Fahr, A. & Fahr, A. (2009). Reactivity of real-time response measurement: The influence of employing RTR techniques on processing media content. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 45-61). Frankfurt a.M.: Peter Lang.

Fahr, A. & Hofer, M. (2013). Psychophysiologische Messmethoden. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 347-365). Wiesbaden: Springer Fachmedien.

Fenwick, I. & Rice, M. D. (1991). Reliability of continuous measurement copy-testing methods. *Journal of Advertising Research*, *31* (1), 23-29.

Ferron, J., Dailey, R. & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, *37* (3), 379-403. doi: 10.1207/S15327906MBR3703_4

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4. Aufl.). London: Sage.

Field, A. P., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Sage.

Franzmann, S. (2006). Parteistrategien auf Oligopolistischen Issue-Märkten. Eine Empirische Analyse der Wahlprogrammatik in Deutschland, Dänemark, Österreich und den Niederlanden mit Hilfe des Gutenberg-Modells. *Politische Vierteljahresschrift*, *47* (4), 571-594. doi: 10.1007/s11615-006-0342-1

Fredrickson, B. L. & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65* (1), 45-55. doi: 10.1037/0022-3514.65.1.45

Früh, H. (2010). *Emotionalisierung durch Nachrichten: Emotionen und Informationsverarbeitung in der Nachrichtenrezeption*. Baden-Baden: Nomos.

Früh, H. & Fahr, A. (2006). Erlebte Emotionen. *Publizistik*, *51* (1), 24-38. doi: 10.1007/s11616-006-0003-9

Funke, F. (2010). *Internet-based measurement with visual analogue scales: An experimental investigation*. Doktorarbeit, Universität Tübingen. Zugriff auf `http://tobias-lib.uni-tuebingen.de/volltexte/2010/5282`

Funke, F. & Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, *24* (3), 310-327. doi: 10.1177/1525822x12444061

Gabriel, O. W. & Kornelius, B. (2011). Die baden-württembergische Landtagswahl vom 27. März 2011: Zäsur und Zeitenwende? *Zeitschrift für Parlamentsfragen*, *42* (2), 784-804.

Gatz, D. F. & Smith, L. (1995a). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. *Atmospheric Environment*, *29* (11), 1185-1193. doi: 10.1016/1352-2310(94)00210-C

Gatz, D. F. & Smith, L. (1995b). The standard error of a weighted mean concentration—II. Estimating confidence intervals. *Atmospheric Environment*, *29* (11), 1195-1200. doi: 10.1016/1352-2310(94)00209-4

Geese, S., Zubayr, C. & Gerhard, H. (2005). Berichterstattung zur Bundestagswahl 2005 aus Sicht der Zuschauer. *Media Perspektiven*, *o.Jg.* (12), 613-626.

Geiser, C. (2010). *Datenanalyse mit Mplus*. Wiesbaden: VS.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A. & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60* (4), 328-331. doi: 10.1198/000313006X152649

Goldstein, H. (2011). *Multilevel statistical models* (4. Aufl.). Chichester: Wiley.

Greene, S. (2002). The social-psychological measurement of partisanship. *Political Behavior*, *24* (3), 171-197. doi: 10.1023/A:1021859907145

Grewe, O., Nagel, F., Kopiez, R. & Altenmüller, E. (2007). Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, *7* (4), 774-788. doi: 10.1037/1528-3542.7.4.774

Gscheidle, C. & Gerhard, H. (2013). Berichterstattung zur Bundestagswahl 2013 aus Sicht der Zuschauer. *Media Perspektiven*, *o.Jg.* (12), 558-573.

Harrell Jr, F. E. (2012). Hmisc: Harrell Miscellaneous (with contributions from Charles Dupont and many others) (R package version 3.10-1) [Software].

Hart, R. P. & Jarvis, S. E. (1997). Political debate. *American Behavioral Scientist*, *40* (8), 1095-1122. doi: 10.1177/0002764297040008010

Hayes, A. F. (2005). *Statistical methods for communication science*. Mahwah: Lawrence Erlbaum.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32* (4), 385-410. doi: 10.1111/j.1468-2958.2006.00281.x

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1* (1), 77-89. doi: 10.1080/19312450709336664

Hinrichs, J. P. (2002). Wir bauen einen Themenpark. Wähler werden doch mit Inhalten gewonnen – durch Issues Management. In M. Althaus (Hrsg.), *Kampagne! Neue Strategien für Wahlkampf, PR und Lobbying* (3. Aufl., S. 45-64). Münster: Lit.

Hofrichter, J. (2004). Die Rolle der TV-Duelle im Bundestagswahlkampf 2002. In F. Brettschneider, J. Deth & E. Roller (Hrsg.), *Die Bundestagswahl 2002* (S. 51-73). Wiesbaden: VS.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2. Aufl.). New York: Routledge.

Hox, J. & Roberts, J. K. (2011). *Handbook of advanced multilevel analysis*. New York: Routledge.

Hox, J. & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everitt & D. C. Howell (Hrsg.), *Encyclopedia of statistics in behavioral science* (Bd. 3, S. 1296-1305). Chichester: Wiley & Sons.

Hughes, G. D. (1992). Realtime response measures redefine advertising wearout. *Journal of Advertising Research*, *32* (3), 61-77.

Hutcherson, C. A., Goldin, P. R., Ochsner, K. N., Gabrieli, J. D., Barrett, L. F. & Gross, J. J. (2005). Attention and emotion: Does rating emotion alter neural responses to amusing and sad films? *NeuroImage*, *27* (3), 656-668. doi: 10.1016/j.neuroimage.2005.04.028

Iyengar, S. (2011). Experimental designs for political communication research: Using new technology and online participant pools to overcome the problem of generalizability. In E. P. Bucy & R. L. Holbert (Hrsg.), *Sourcebook for political communication research. Methods, measures, and analytical techniques* (S. 129-148). New York: Routledge.

Iyengar, S., Jackman, S., Hahn, K. S. & Lim, J. (2010, Jun 22). *Polarization in less than 30 seconds: Continuous monitoring of voter response to campaign advertising*. Vortrag auf der 60. Jahrestagung der International Communication Association (ICA), Singapore.

Iyengar, S. & Simon, A. F. (2000). New perspectives and evidence on political communication and campaign effects. *Annual Review of Psychology*, *51* (1), 149-169. doi: 10.1146/annurev.psych.51.1.149

Jackob, N., Petersen, T. & Roessing, T. (2008). Strukturen der Wirkung von Rhetorik. *Publizistik*, *53* (2), 215-230. doi: 10.1007/s11616-008-0076-8

Jansen, C. & Maier, J. (2013, März). *Negativity in German televised debates, 1997-2012. A content analysis of candidate messages.* Vortrag auf der 41st ECPR Joint Sessions of Workshops, Mainz.

Jarman, J. W. (2005). Political affiliation and presidential debates. *American Behavioral Scientist*, *49* (2), 229-242. doi: 10.1177/0002764205280921

Jarvis, B. (2012). MediaLab (v2012) [Software].

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47* (2), 263-291.

Kaid, L. L. (2009). Immediate responses to political television spots in U.S. elections. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 137-153). Frankfurt a.M.: Peter Lang.

Kercher, J. (2013). *Verstehen und Verständlichkeit von Politikersprache*. Wiesbaden: VS.

Kercher, J., Bachl, M., Vögele, C. & Vohle, F. (2012, März). *The MediaLiveTracker. A new online-tool for real-time-response-measurement.* Vortrag auf der 14. Jahrestagung der Deutschen Gesellschaft für Onlineforschung (GOR), Mannheim.

Kipp, M. (2011). Anvil. The video annotation research tool (Version 5.0.20, http://www.anvil-software.org/) [Software].

Kipp, M. (2014). ANVIL: A universal video research tool. In J. Durand, U. Gut & G. Kristofferson (Hrsg.), *The Oxford handbook of corpus phonology* (S. 420-436). Oxford: Oxford University Press.

Kirchgässner, G. & Wolters, J. (2007). *Introduction to modern time series analysis*. Heidelberg: Springer.

Klein, M. (2005). Der Einfluss der beiden TV-Duelle im Vorfeld der Bundestagswahl 2002 auf die Wahlbeteiligung und die Wahlentscheidung. Eine log-lineare Pfadanalyse auf der Grundlage von Paneldaten. *Zeitschrift für Soziologie*, *34* (3), 207-222.

Klein, M. & Rosar, U. (2007). Wirkungen des TV-Duells im Vorfeld der Bundestagswahl 2005 auf die Wahlentscheidung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *59* (1), 81-104. doi: 10.1007/s11577-007-0004-3

Klimmt, C. & Weber, R. (2013). Das Experiment in der Kommunikationswissenschaft. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 125-144). Wiesbaden: Springer Fachmedien.

Koller, M. (2013). *Robust estimation of linear mixed models*. Doktorarbeit, Eidgenössische Technische Hochschule Zürich. Zugriff auf `http://dx.doi.org/10.3929/ethz-a-007632241`

Koller, M. (2014). robustlmm: Robust linear mixed effects models (R package version 1.4-2, http://CRAN.R-project.org/package=robustlmm) [Software].

Krafft, A. & Zaiss, V. (2013). Das TV-Duell aus Sicht der Wahlkämpfer – Ein Blick in die Kampagnenpraxis. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 237-250). Wiesbaden: VS.

Krause, B. & Gehrau, V. (2007). Das Paradox der Medienwirkung auf Nichtnutzer. *Publizistik*, *52* (2), 191-209. doi: 10.1007/s11616-007-0083-1

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2013a). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) (R package version 1.2-0, http://CRAN.R-project.org/package=lmerTest) [Software].

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2013b). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) (R package version 2.0-0, http://CRAN.R-project.org/package=lmerTest) [Software].

Lambooij, M., Ijsselsteijn, W. A. & Heynderickx, I. (2011). Visual discomfort of 3D TV: Assessment methods and modeling. *Displays*, *32* (4), 209-218. doi: 10.1016/j.displa.2011.05.012

Lang, A., Sanders-Jackson, A., Wang, Z. & Rubenking, B. (2013). Motivated message processing: How motivational activation influences resource allocation, encoding, and storage of TV messages. *Motivation and Emotion*, *37* (3), 508-517. doi: 10.1007/s11031-012-9329-y

Langer, W. (2009). *Mehrebenenanalyse. Eine Einführung für Forschung und Praxis* (2. Aufl.). Wiesbaden: VS.

Larsen, R. J. & Fredrickson, B. L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener & N. Schwarz (Hrsg.), *Well-being: The foundations of hedonic psychology* (S. 40-60). New York: Sage.

Lavine, H. (2002). On-line versus memory-based process models of political evaluation. In K. R. Monroe (Hrsg.), *Political psychology* (S. 225-274). Mahwah: Lawrence Erlbaum.

Lee, C. & Benoit, W. L. (2005). A functional analysis of the 2002 Korean presidential debates. *Asian Journal of Communication*, *15* (2), 115-132. doi: 10.1080/01292980500118193

Lee, S. & Lang, A. (2013). Redefining media content and structure in terms of available resources: Toward a dynamic human-centric theory of communication. *Communication Research*. doi: 10.1177/0093650213488416

Lemmink, J. & Mattsson, J. (1998). Warmth during non-productive retail encounters: the hidden side of productivity. *International Journal of Research in Marketing*, *15* (5), 505-517. doi: 10.1016/S0167-8116(98)00016-0

Lemmink, J. & Mattsson, J. (2002). Employee behavior, feelings of warmth and customer perception in service encounters. *International Journal of Retail & Distribution Management*, *30* (1), 18-33. doi: 10.1108/09590550210415239

Levenson, R. W. & Gottman, J. M. (1983). Marital interaction: physiological linkage and affective exchange. *Journal of Personality and Social Psychology*, *45* (3), 587-597. doi: 10.1037//0022-3514.45.3.587

Levy, M. R. (1982). The Lazarsfeld-Stanton Program Analyzer: An historical note. *Journal of Communication*, *32* (4), 30-38. doi: 10.1111/j.1460-2466.1982.tb02516.x

Lindstrom, M. & Bates, D. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*, 673-687.

Luo, W. (2013). *The impact of misspecifying cross-classified random effects models in cross-sectional and longitudinal multilevel data: a Monte Carlo study*. Phd thesis, Texas A&M University. Zugriff auf `http://hdl.handle.net/1969.1/ETD-TAMU-1507`

Madsen, C. K. (1998). Emotion versus tension in Haydn's Symphony no. 104 as measured by the two-dimensional continuous response digital interface. *Journal of Research in Music Education*, *46* (4), 546-554. doi: 10.2307/3345350

Maier, J. (2007). Erfolgreiche Überzeugungsarbeit. Urteile über den Debattensieger und die Veränderung der Kanzlerpräferenz. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 90-109). Wiesbaden: VS.

Maier, J. (2009). "Frau Merkel wird doch noch Kritik ertragen können…": Inhalt, Struktur, Wahrnehmung und Wirkung des wirtschaftspolitischen Teils der Fernsehdebatte 2005. In

O. W. Gabriel, B. Weßels & J. W. Falter (Hrsg.), *Wahlen und Wähler. Analysen aus Anlass der Bundestagswahl 2005* (S. 177-201). Wiesbaden: VS.

Maier, J. (2011). Führt der Einsatz von Real-Time-Response-Technik zu einer anderen Wahrnehmung von Fernsehdebatten? – Ergebnisse zweier Experimente zur externen Validität von Echtzeitmessungen. *Politische Psychologie*, *2* (1), 7-21.

Maier, J. (2013). Rezeptionsbegleitende Erfassung individueller Reaktionen auf Medieninhalte: Bedeutung, Varianten, Qualität und Analyse von Real-Time-Response-Messungen. *ESSA-CHESS–Journal for Communication Studies*, *6* (1), 169-184.

Maier, J. & Faas, T. (2003). The affected German voter: Televised debates, follow-up communication and candidate evaluations. *Communications*, *28* (4), 383-404. doi: 10.1515/comm.2003.025

Maier, J. & Faas, T. (2004). Debattenwahrnehmung und Kandidatenorientierung. *Zeitschrift für Medienpsychologie*, *16* (1), 26-35. doi: 10.1026/1617-6383.16.1.26

Maier, J. & Faas, T. (2009). Measuring spontaneous reactions to media messages the traditional way: uncovering political information processing with push button devices. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 15-26). Frankfurt a.M.: Peter Lang.

Maier, J. & Faas, T. (2011). 'Miniature campaigns' in comparison: The German televised debates, 2002–09. *German Politics*, *20* (1), 75-91. doi: 10.1080/09644008.2011.554102

Maier, J. & Maier, M. (2007). Audience reactions to negative campaign spots in the 2005 German national elections: the case of two ads called „the Ball ". *Human Communication*, *10*, 329-344.

Maier, J. & Maier, M. (2013). Serving different agendas. In E. Czerwick (Hrsg.), *Politische Kommunikation in der repräsentativen Demokratie der Bundesrepublik Deutschland* (S. 149-164). Wiesbaden: Springer Fachmedien.

Maier, J., Maier, M., Maurer, M., Reinemann, C. & Meyer, V. (2009). *Real-time response measurement in the social sciences. Methodological perspectives and applications*. Frankfurt a. M. et al.: Peter Lang.

Maier, J., Maurer, M., Reinemann, C. & Faas, T. (2007). Reliability and validity of real-time response measurement: A comparison of two studies of a televised debate in Germany. *International Journal of Public Opinion Research*, *19* (1), 53-73. doi: 10.1093/ijpor/edl002

Maier, M. & Maier, J. (2009). Measuring the perception and the impact of verbal and visual content of televised political ads. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 63-84). Frankfurt a.M.: Peter Lang.

Maier, M. & Strömbäck, J. (2009). Responses to televised debates: A comparative study of Germany and Sweden. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 97-116). Frankfurt a.M.: Peter Lang.

Manor, O. & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, *46* (4), 801-817. doi: 10.1016/j.csda.2003.10.005

Martel, M. (1983). *Political campaign debates: Images, strategies, and tactics*. New York: Longman.

Matthes, J., Wirth, W. & Schemer, C. (2007). Measuring the unmeasurable? Toward operationalizing on-line and memory-based political judgments in surveys. *International Journal of Public Opinion Research*, *19* (2), 247-257. doi: 10.1093/ijpor/edm001

Maurer, M. (2007). Themen, Argumente, rhetorische Strategien. Die Inhalte des TV-Duells. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 33-52). Wiesbaden: VS.

Maurer, M. (2009). Sagen Bilder mehr als tausend Worte? *Medien und Kommunikationswissenschaft*, *57* (2), 198-216.

Maurer, M. (2012). Die Kombination von Inhaltsanalyse- und Befragungsdaten in der Medienwirkungsforschung: Theoretische Überlegungen und methodische Entscheidungsprozesse. In

W. Loosen & A. Scholl (Hrsg.), *Methodenkombinationen in der Kommunikationswissenschaft: Methodologische Herausforderungen und empirische Praxis* (S. 89-101). Köln: von Halem.

Maurer, M. (2013a). Grundlagen: Designs und Forschungslogik in der Medienwirkungsforschung. In W. Schweiger & A. Fahr (Hrsg.), *Handbuch Medienwirkungsforschung* (S. 549-563). Wiesbaden: Springer Fachmedien.

Maurer, M. (2013b). Real-Time Response Messung: Kontinuierliche Befragung in Echtzeit. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 219-234). Wiesbaden: Springer Fachmedien.

Maurer, M. & Reinemann, C. (2003). *Schröder gegen Stoiber: Nutzung, Wahrnehmung und Wirkung der TV-Duelle*. Wiesbaden: Westdeutscher.

Maurer, M. & Reinemann, C. (2007a). TV-Duelle als Instrumente der Wahlkampfkommunikation: Mythen und Fakten. In N. Jackob (Hrsg.), *Wahlkämpfe in Deutschland* (S. 317-331). Wiesbaden: VS.

Maurer, M. & Reinemann, C. (2007b). Warum TV-Duelle Wahlen entscheiden können. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 229-246). Wiesbaden: VS.

Maurer, M. & Reinemann, C. (2009). RTR measurement in the social sciences: Applications, benefits, and some open questions. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 1-13). Frankfurt a.M.: Peter Lang.

Maurer, M., Reinemann, C., Maier, J. & Maier, M. (2007). *Schröder gegen Merkel*. Wiesbaden: VS.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 601-613). Wiesbaden: VS.

Mazerolle, M. J. (2013). AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c) (R package version 1.32, http://CRAN.R-project.org/package=AICcmodavg) [Software].

McKinney, M. (2007). Debates. In L. L. Kaid & C. Holtz-Bacha (Hrsg.), *Encyclopedia of political communication* (S. 159-165). Thousand Oaks: Sage.

McKinney, M. & Carlin, D. (2004). Political campaign debates. In L. L. Kaid (Hrsg.), *Handbook of political communication research* (S. 203-234). Mahwah: Lawrence Erlbaum.

McKinney, M., Kaid, L. L. & Robertson, T. A. (2001). The front-runner, contenders, and also-rans: Effects of watching a 2000 republican primary debate. *American Behavioral Scientist*, *44* (12), 2232-2251. doi: 10.1177/00027640121958294

McKinnon, L. M. & Tedesco, J. C. (1993). The third 1992 presidential debate: Channel and commentary effects. *Argumentation & Advocacy*, *30* (2), 106-118.

McKinnon, L. M. & Tedesco, J. C. (1999). The influence of medium and media commentary on presidential debate effects. In L. L. Kaid (Hrsg.), *The electronic election: Perspectives on the 1996 campaign* (S. 191-206). Mahwah: Erlbaum.

Millard, W. J. (1992). A history of handsets for direct measurement of audience response. *International Journal of Public Opinion Research*, *4* (1), 1. doi: 10.1093/ijpor/4.1.1

Mizon, G. E. (1995). A simple message for autocorrelation correctors: Don't. *Journal of Econometrics*, *69* (1), 267-288. doi: 10.1016/0304-4076(94)01671-L

Müller, M. F. (2003). „Der oder ich!" Eine Analyse der Kandidatenduelle im Bundestagswahlkampf 2002. In A. M. Wüst (Hrsg.), *Politbarometer* (S. 295-315). Opladen: Westdeutscher.

Nagel, F. (2012). *Die Wirkung verbaler und nonverbaler Kommunikation in TV-Duellen: Eine Untersuchung am Beispiel von Gerhard Schröder und Angela Merkel*. Wiesbaden: VS.

Nagel, F., Kopiez, R., Grewe, O. & Altenmüller, E. (2007). EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, *39* (2), 283-290. doi: 10.3758/bf03193159

Nagel, F., Maurer, M. & Reinemann, C. (2012). Is there a visual dominance in political communication? How verbal, visual, and vocal communication shape viewers' impressions of political

candidates. *Journal of Communication*, *62* (5), 833-850. doi: 10.1111/j.1460-2466.2012.01670.x

Ottler, S. (2013). RTR-Messung: Möglichkeiten und Grenzen einer sozialwissenschaftlichen Methode. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 113-134). Wiesbaden: VS.

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahwah: Lawrence Erlbaum.

Papastefanou, G. (2013). Reliability and validity of RTR measurement device. *GESIS-Working Papers* (27).

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, 825-850.

Petrocik, J. R., Benoit, W. L. & Hansen, G. J. (2003). Issue ownership and presidential campaigning, 1952–2000. *Political Science Quarterly*, *118* (4), 599-626. doi: 10.1002/j.1538-165X.2003.tb00407.x

Petty, R. & Cacioppo, J. (1986). *Communication and persuasion: central and peripheral routes to attitude change*. New York: Springer.

Peugh, J. L. & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement*, *65* (5), 717-741. doi: 10.1177/0013164405278558

Pinheiro, J. & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. (2013). nlme: Linear and nonlinear mixed effects models (R package version 3.1-108, http://CRAN.R-project.org/package=nlme) [Software].

Racine Group. (2002). White paper on televised political campaign debates. *Argumentation & Advocacy*, *38* (4), 199-218.

Ramanathan, S. & McGill, A. (2007). Consuming with others: Social influences on moment-to-moment and retrospective evaluations of an experience. *Journal of Consumer Research*, *34* (4), 506-524. doi: 10.1086/520074

Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19* (4), 337-350. doi: 10.3102/10769986019004337

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R. & Weßels, B. (2012). German Longitudinal Election Study – Wahlkampf-Panel, 10.07.-07.10.2009, ZA5305. *GESIS*. doi: 10.4232/1.11131

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B. & Wolf, C. (2013). *German Longitudinal Election Study (http://gles.eu/)*. GESIS. Zugriff auf `http://gles.eu/`

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, *18* (4), 321-349. doi: 10.3102/10769986018004321

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2. Aufl.). Thousand Oaks: Sage.

Reinemann, C. (2007). Völlig anderer Ansicht. Die Medienberichterstattung über das TV-Duell. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 167-194). Wiesbaden: VS.

Reinemann, C., Maier, J., Faas, T. & Maurer, M. (2005). Reliabilität und Validität von RTR-Messungen. *Publizistik*, *50* (1), 56-73. doi: 10.1007/s11616-005-0118-4

Reinemann, C. & Maurer, M. (2005). Unifying or polarizing? Short-term effects and postdebate consequences of different rhetorical strategies in televised debates. *Journal of Communication*, *55* (4), 775-794. doi: 10.1111/j.1460-2466.2005.tb03022.x

Reinemann, C. & Maurer, M. (2007a). Kandidatenwahrnehmung in Echtzeit. Anlage und Methoden der TV-Duell-Studie 2005. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 19-31). Wiesbaden: VS.

Reinemann, C. & Maurer, M. (2007b). Populistisch und unkonkret. Die unmittelbare Wahrnehmung des TV-Duells. In M. Maurer, C. Reinemann, J. Maier & M. Maier (Hrsg.), *Schröder gegen Merkel* (S. 53-89). Wiesbaden: VS.

Reinemann, C. & Maurer, M. (2007c). Schröder gegen Merkel. Wahrnehmung und Wirkung des TV-Duells. In F. Brettschneider, O. Niedermayer & B. Weßels (Hrsg.), *Die Bundestagswahl 2005* (S. 197-217). Wiesbaden: VS.

Reinemann, C. & Maurer, M. (2008). Televised debates. In W. Donsbach (Hrsg.), *The international encyclopedia of communication* (S. 5057-5063). Malden: Blackwell Publishing.

Reinemann, C. & Maurer, M. (2009). Is RTR biased towards verbal message components? An experimental test of the external validity of RTR measurements. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 27-44). Frankfurt a.M.: Peter Lang.

Reinemann, C. & Maurer, M. (2010). Leichtgläubig und manipulierbar? Die Rezeption persuasiver Wahlkampfbotschaften durch politisch Interessierte und Desinteressierte. In T. Faas, K. Arzheimer & S. Roßteutscher (Hrsg.), *Information – Wahrnehmung – Emotion* (S. 239-257). Wiesbaden: VS.

Reips, U.-D. & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40* (3), 699-704. doi: 10.3758/BRM.40.3.699

Roessing, T., Jackob, N. & Petersen, T. (2009). The explanatory power of RTR graphs. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 85-95). Frankfurt a.M.: Peter Lang.

Rossiter, J. R. & Thornton, J. (2004). Fear-pattern analysis supports the fear-drive model for antispeeding road-safety TV ads. *Psychology and Marketing*, *21* (11), 945-960. doi: 10.1002/mar.20042

Rust, L. W. (1985). Using test scores to guide the content analysis of TV materials. *Journal of Advertising Research*, *25* (5), 17-23.

Scharkow, M. (2012). *Automatische Inhaltsanalyse und maschinelles Lernen*. Berlin: epubli.

Schemer, C. (2012). Reinforcing spirals of negative affects and selective attention to advertising in a political campaign. *Communication Research*, *39* (3), 413-434. doi: 10.1177/0093650211427141

Schemer, C., Matthes, J. & Wirth, W. (2009). Applying latent growth models to the analysis of media effects. *Journal of Media Psychology*, *21* (2), 85-89. doi: 10.1027/1864-1105.21.2.85

Schenk, M. (2007). *Medienwirkungsforschung* (3. Aufl.). Tübingen: Mohr.

Scheufele, B. (1999). *Zeitreihenanalysen in der Kommunikationsforschung. Eine praxisorientierte Einführung in die uni-und multivariate Zeitreihenanalyse mit SPSS for Windows*. Stuttgart: Döbler & Rössler.

Scheufele, B., Schünemann, J. & Brosius, H.-B. (2005). Duell oder Berichterstattung? *Publizistik*, *50* (4), 399-421. doi: 10.1007/s11616-005-0141-5

Schill, D. & Kirk, R. (2009). Applied dial testing: Using real-time response to improve media coverage of debates. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 155-173). Frankfurt a.M.: Peter Lang.

Schill, D. & Kirk, R. (2013). Courting the swing voter: "Real time" insights into the 2008 and 2012 U.S. presidential debates. *American Behavioral Scientist*, *online first*. doi: 10.1177/0002764213506204

Schneider, F. M., Erbsen, J., Satzl, I., Altzschner, R.-S., Kockler, T. & Petzold, S. (2011). Die Übungsseqzenz macht den Meister…? In M. Suckfüll, H. Schramm & C. Wünsch (Hrsg.), *Rezeption und Wirkung in zeitlicher Perspektive* (S. 253-270). Baden-Baden: Nomos.

Schoen, H. & Weins, C. (2005). Der sozialpsychologische Ansatz zur Erklärung von Wahlverhalten. In J. Falter & H. Schoen (Hrsg.), *Handbuch Wahlforschung* (S. 187-242). Wiesbaden: VS.

Schrott, P. R. (1990a). Electoral consequences of "winning" televised campaign debates. *Public Opinion Quarterly*, *54* (4), 567-585. doi: 10.1086/269228

Schrott, P. R. (1990b). Wahlkampfdebatten im Fernsehen von 1972 bis 1987: Politikerstrategien und Wählerreaktion. In M. Kaase & H. D. Klingemann (Hrsg.), *Wahlen und Wähler. Analysen aus Anlaß der Bundestagswahl 1987* (S. 647-674). Opladen: Westdeutscher.

Schrott, P. R. & Lanoue, D. J. (2013). The power and limitations of televised presidential debates: Assessing the real impact of candidate performance on public opinion and vote choice. *Electoral Studies*, *32* (4), 684-692. doi: 10.1016/j.electstud.2013.03.006

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51* (3), 154-165. doi: 10.1080/00049539908255353

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music perception*, *21* (4), 561-585.

Schulz, W. (2011). *Politische Kommunikation. Theoretische Ansätze und Ergebnisse empirischer Forschung* (3. Aufl.). Wiesbaden: VS.

Shapiro, M. A. & Chock, T. M. (2003). Psychological processes in perceiving reality. *Media Psychology*, *5* (2), 163-198. doi: 10.1207/s1532785xmep0502_3

Skowronski, J. J. & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: a review of explanations. *Psychological Bulletin*, *105*, 131-142. doi: 10.1037/0033-2909.105.1.131

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17* (3), 281-303. doi: 10.1111/j.1468-2885.2007.00296.x

Slater, M. D. & Hayes, A. F. (2010). The influence of youth music television viewership on changes in cigarette use and association with smoking peers: A social identity, reinforcing spirals perspective. *Communication Research*, *37* (6), 751-773. doi: 10.1177/0093650210375953

Slater, M. D., Henry, K. L., Swaim, R. C. & Anderson, L. L. (2003). Violent media content and aggressiveness in adolescents: A downward spiral model. *Communication Research*, *30* (6), 713-736. doi: 10.1177/0093650203258281

Slater, M. D., Snyder, L. & Hayes, A. F. (2006). Thinking and modeling at multiple levels: The potential contribution of multilevel modeling to communication theory and research. *Human Communication Research*, *32* (4), 375-384. doi: 10.1111/j.1468-2958.2006.00292.x

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, *30* (4), 405-426. doi: 10.1007/bf00170145

Snijders, T. & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2. Aufl.). London: SAGE.

Sparks, J. V. & Lang, A. (2010). An initial examination of the post-auricular reflex as a physiological indicator of appetitive activation during television viewing. *Communication Methods and Measures*, *4* (4), 311-330. doi: 10.1080/19312458.2010.527872

Spieker, A. (2011). Licht ins Dunkel der TV-Duelle: Rhetorische Strategien und ihre Wirkungen im TV-Duell 2009. In J. F. Haschke & A. M. Moser (Hrsg.), *Politik–Deutsch, Deutsch–Politik: Aktuelle Trends und Forschungsergebnisse* (S. 75-93). Berlin: Frank & Timme.

Spieker, A. & Bachl, M. (2010, Oktober). *Opening the ‚black-box': Exploring immediate audience responses to rhetorical strategies in televised debates.* Vortrag auf der 3. European Communication Conference der ECREA, Hamburg.

Stayman, D. M. & Aaker, D. A. (1993). Continuous measurement of self-report of emotional response. *Psychology and Marketing*, *10* (3), 199-214. doi: 10.1002/mar.4220100304

Stoel, R. D. & Galindo Garre, F. (2011). Growth curve analysis using multilevel regression and structural equation modeling. In J. Hox & J. K. Roberts (Hrsg.), *Handbook of advanced multilevel analysis* (S. 97-111). New York: Routledge.

Strömbäck, J., Maier, M. & Maier, J. (2009, Mai). *The use and effectiveness of negativity in televised debates.* Vortrag auf der 59. Jahrestagung der International Communication Association, Chicago.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th Aufl.). Boston: Pearson.

Tapper, C. & Quandt, T. (2006). „Trotzdem nochmal nachgefragt, Frau Kirchhof...“ Eine dialoganalytische Untersuchung des Fernseh-Duells im Wahlkampf 2005. In C. Holtz-Bacha (Hrsg.), *Die Massenmedien im Wahlkampf* (S. 246-276). Wiesbaden: VS.

Tapper, C. & Quandt, T. (2010). „Ich beantworte die Fragen so, wie ich mir das vorgenommen habe...“ Eine dialoganalytische Untersuchung des Fernseh-Duells im Wahlkampf 2009. In C. Holtz-Bacha (Hrsg.), *Die Massenmedien im Wahlkampf* (S. 283-312). Wiesbaden: VS.

Tedesco, J. C. & Ivory, A. (2009). Health message primes and sexual health campaign messages. In J. Maier, M. Maier, M. Maurer, C. Reinemann & V. Meyer (Hrsg.), *Real-time response measurement in the social sciences* (S. 175-192). Frankfurt a.M.: Peter Lang.

Trepte, S. & Wirth, W. (2004). Externe versus interne Validität in kommunikationswissenschaftlichen Experimenten. In W. Wirth, E. Lauf & A. Fahr (Hrsg.), *Forschungslogik und -design in der Kommunikationswissenschaft* (Bd. 1, S. 60-87). Köln: von Halem.

Van Der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity*, *32* (1), 15-29. doi: 10.1023/a:1004233225855

von Pape, T. & Meyer, V. (2011). Emotionen, die hochkochen? Zum Einfluss von Rezeptionsemotionen auf die Härte von Strafeinstellungen. In M. Suckfüll, H. Schramm & C. Wünsch (Hrsg.), *Rezeption und Wirkung in zeitlicher Perspektive* (S. 145-165). Baden-Baden: Nomos.

Vögele, C. (2013). Das TV-Duell Mappus gegen Schmid – die Ausgangslage. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 47-55). Wiesbaden: VS.

Vögele, C., Brettschneider, F. & Bachl, M. (2013). Parteien, Massenmedien, Wähler und TV-Debatten in Landtagswahlkämpfen. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 29-46). Wiesbaden: VS.

Vögele, C. & Schmalz, I. (2013). „Bildung, Bildung und nochmals Bildung.“ Die Bildungspolitik im TV-Duell. In M. Bachl, F. Brettschneider & S. Ottler (Hrsg.), *Das TV-Duell in Baden-Württemberg 2011* (S. 219-236). Wiesbaden: VS.

Walls, T. A. & Schafer, J. L. (2006). Introduction: Intensive longitudinal data. In T. A. Walls & J. L. Schafer (Hrsg.), *Models for intensive longitudinal data* (S. xi-xxii). New York: Oxford University Press.

Wang, Z., Lang, A. & Busemeyer, J. R. (2011). Motivational processing and choice behavior during television viewing: An integrative dynamic approach. *Journal of Communication*, *61* (1), 71-93. doi: 10.1111/j.1460-2466.2010.01527.x

Wang, Z., Morey, A. C. & Srivastava, J. (2012). Motivated selective attention during political ad processing: The dynamic interplay between emotional ad content and candidate evaluation. *Communication Research*. doi: 10.1177/0093650212441793

Wang, Z., Solloway, T., Tchernev, J. M. & Barker, B. (2012). Dynamic motivational processing of antimarijuana messages: Coactivation begets attention. *Human Communication Research*, *38* (4), 485-509. doi: 10.1111/j.1468-2958.2012.01431.x

Weaver III, J. B., Huck, I. & Brosius, H.-B. (2009). Biasing public opinion: Computerized continuous response measurement displays impact viewers' perceptions of media messages. *Computers in Human Behavior*, *25* (1), 50-55. doi: 10.1016/j.chb.2008.06.004

Wehner, M. (2013). Die historische Niederlage der CDU. Ursachen für das Scheitern. In U. Wagschal, U. Eith & M. Wehner (Hrsg.), *Der historische Machtwechsel: Grün-Rot in Baden-Württemberg* (S. 119-142). Baden-Baden: Nomos.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40* (1), 1-29.

Wolf, B. (2010). *Beurteilung politischer Kandidaten in TV-Duellen*. Baden-Baden: Nomos.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics - Simulation and Computation*, *22* (4), 1079-1106. doi: 10.1080/03610919308813143

Wolling, J. & Wirth, W. (2012). Die Verknüpfung von Umfrage- und Inhaltsanalysedaten in der Medienwirkungsforschung. In W. Loosen & A. Scholl (Hrsg.), *Methodenkombinationen in der Kommunikationswissenschaft: Methodologische Herausforderungen und empirische Praxis* (S. 68-88). Köln: von Halem.

Woltman Elpers, J. L. C. M., Mukherjee, A. & Hoyer, W. (2004). Humor in television advertising: A moment-to-moment analysis. *Journal of Consumer Research*, *31* (3), 592-598.

Woltman Elpers, J. L. C. M., Wedel, M. & Pieters, R. G. M. (2003). Why do consumers stop viewing television commercials? Two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research*, *40* (4), 437-453. doi: 10.2307/30038877

Wünsch, C. (2006a). Unterhaltung als Performance. Überlegungen und erste Anwendungserfahrungen mit einem Messinstrument zur dynamischen Erfassung von Unterhaltungserleben. In W. Wirth, H. Schramm & V. Gehrau (Hrsg.), *Unterhaltung durch Medien. Theorie und Messung* (S. 174-203). Köln: von Halem.

Wünsch, C. (2006b). *Unterhaltungserleben. Ein hierarchisches Zwei-Ebenen-Modell affektiv-kognitiver Informationsverarbeitung*. Köln: von Halem.

Yanovitzky, I. & Greene, K. (2009). Quantitative methods and causal inference in media effects research. In R. L. Nabi & M. B. Oliver (Hrsg.), *The SAGE handbook of media processes and effects* (S. 35-52). Thousand Oaks: Sage.

ZA & ZUMA. (2012). Sympathie-Skalometer. In A. Glöckner-Rist (Hrsg.), *ZIS Version 15.00*. Bonn: GESIS.

Zaller, J. (1992). *The nature and origins of mass opinion*. New York: Cambridge University Press.

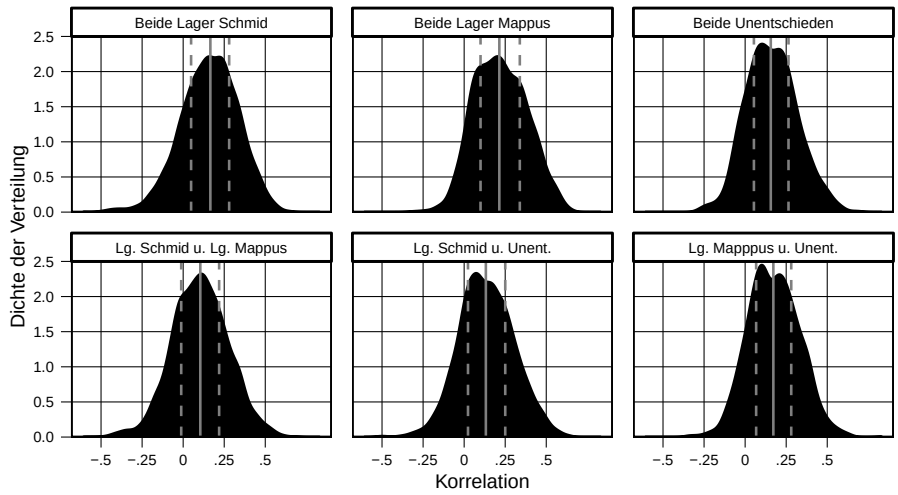
Zhu, J.-H., Milavsky, J. R. & Biswas, R. (1994). Do televised debates affect image perception more than issue knowledge? A study of the first 1992 presidential debate. *Human Communication Research*, *20* (3), 302-333. doi: 10.1111/j.1468-2958.1994.tb00325.x

Zubayr, C., Geese, S. & Gerhard, H. (2009). Berichterstattung zur Bundestagswahl 2009 aus Sicht der Zuschauer. *Media Perspektiven*, *o.Jg.* (12), 637-650.

Zubayr, C. & Gerhard, H. (2002). Berichterstattung zur Bundestagswahl 2002 aus Sicht der Zuschauer. *Media Perspektiven*, *o.Jg.* (12), 586-599.

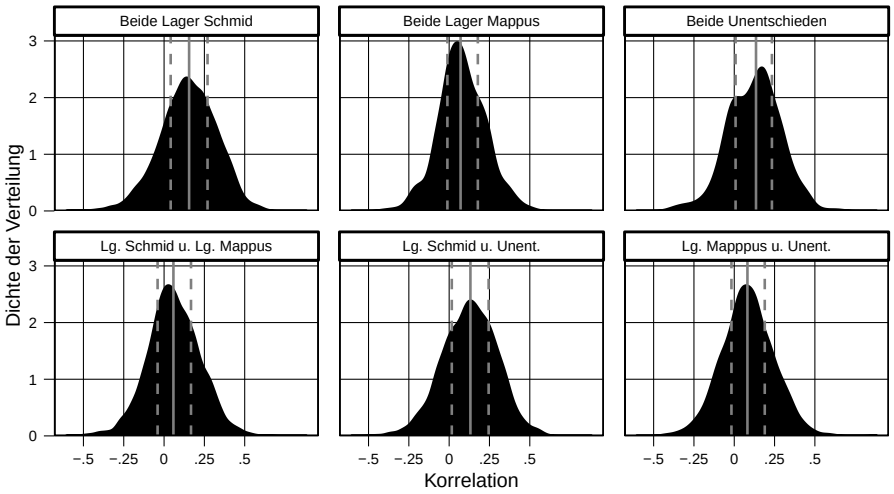
A Zusätzliche Tabellen und Abbildungen

A.1 Zu Kapitel 4



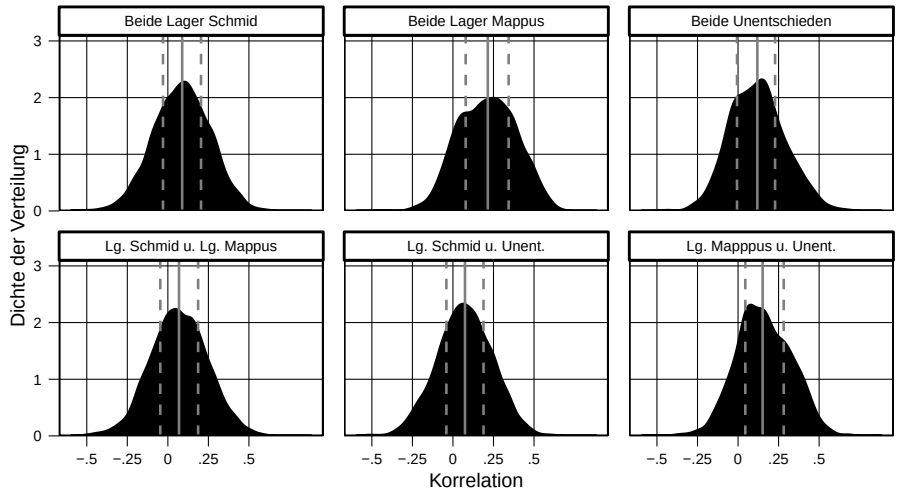
Anmerkungen
Dichte der Verteilung der Korrelationen (Pearsons *r*) zwischen den individuellen RTR-Zeitreihen aller 15400 Rezipienten-Paare (Gesamte Zeitreihe) geordnet nach Lagerzugehörigkeit der Rezipienten. *Durchgezogene Linie*: Median; *Gestrichelte Linien*: 25- und 75-Perzentile, Interquartilsrange.

Abbildung A.1: Verteilung der Korrelationen zwischen den individuellen Zeitreihen (Gesamte Zeitreihen)



Anmerkungen
Dichte der Verteilung der Korrelationen (Pearsons *r*) zwischen den individuellen RTR-Zeitreihen aller 15400 Rezipienten-Paare (Ausschnitte der Zeitreihe mit Schmid als Sprecher) geordnet nach Lagerzugehörigkeit der Rezipienten. *Durchgezogene Linie*: Median; *Gestrichelte Linien*: 25- und 75-Perzentile, Interquartilsrange.

Abbildung A.2: Verteilung der Korrelationen zwischen den individuellen Zeitreihen (Schmid als Sprecher)



Anmerkungen
Verteilung der Korrelationen (Pearsons *r*) zwischen den individuellen RTR-Zeitreihen aller 15400 Rezipienten-Paare (Ausschnitte der Zeitreihe mit Mappus als Sprecher) geordnet nach Lagerzugehörigkeit der Rezipienten. *Durchgezogene Linie*: Median; *Gestrichelte Linien*: 25- und 75-Perzentile, Interquartilsrange.

Abbildung A.3: Verteilung der Korrelationen zwischen den individuellen Zeitreihen (Mappus als Sprecher)

A.2 Zu Kapitel 6

Tabelle A.1: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Voreinstellungen (M1)

	M1.0	M1.1	M1.2	M1.3	M1.4	M1.5
AICc	28758	28743	28738	28732	28735	28730
Deviance	28752	28732	28727	28721	28715	28720
ΔDeviance (χ^2)		19	24	30	36	−4
Freiheitsgrade		2	2	2	6	−4
Signifikanz		<.001	<.001	<.001	<.001	0.365
$R^2_{2_{\text{Slope}}}$		.12	.15	.19	.20	.20

Anmerkungen
Dargestellt sind das Informationskriterium *AICc*, die Kenngrößen der Likelihood-Ratio-Tests und die Reduzierung der personenbezogenen Varianzkomponente ($R^2_{2_{\text{Slope}}}$) im Vergleich zu M1.0. Die Modelle M1.1 bis M1.4 werden mit Likelihood-Ratio-Tests mit M1.0 verglichen. Das finale Modell M1.5 wird mit dem vollen Modell M1.4 verglichen, um zu zeigen, dass der Verzicht auf die entsprechenden Koeffizienten das Modell nicht signifikant verschlechtert. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.2: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M1)

	M1.0	M1.1	M1.2	M1.3	M1.4	M1.5
Slope	0.13^{*}	0.21	0.30^{***}	0.22^{***}	0.26^{*}	0.37^{***}
	(0.06)	(0.13)	(0.07)	(0.07)	(0.14)	(0.07)
Slope x Lg. Mappus		0.32^{*}			0.15	
		(0.18)			(0.18)	
Slope x Lg. Schmid		-0.32^{*}			−0.01	
		(0.16)			(0.17)	
Slope x Sk. Mappus			0.09^{***}		0.05^{*}	0.09^{***}
			(0.02)		(0.03)	(0.02)
Slope x Sk. Schmid			-0.10^{**}		0.01	
			(0.04)		(0.05)	
Slope x Sk. CDU				0.08^{***}	0.04	
				(0.02)	(0.03)	
Slope x Sk. SPD				-0.12^{***}	-0.11^{**}	-0.11^{***}
				(0.03)	(0.04)	(0.03)

Anmerkungen
Dargestellt sind die REML-Koeffizienten β mit ihren Standardfehlern (s.e) und Signifikanzniveau (Einseitige T-Tests mit Freiheitsgraden nach Satterthwaite-Approximation): $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Lg.: Lager; Sk.: Skalometer. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.3: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Voreinstellungen (M2)

	M2.0	M2.1	M2.2	M2.3	M2.4	M2.5
AICc	28001	27989	27984	27978	27989	27977
Deviance	27988	27967	27963	27956	27947	27955
ΔDeviance (χ^2)		21	25	32	41	−8
Freiheitsgrade		4	4	4	12	−8
Signifikanz		<.001	<.001	<.001	<.001	.437
$R^2_{2_{\text{Intercept}}}$		.04	.05	.09	.09	.09
$R^2_{2_{\text{Slope}}}$		.08	.11	.11	.12	.12

Anmerkungen
Dargestellt sind das Informationskriterium *AICc*, die Kenngrößen der Likelihood-Ratio-Tests und die Reduzierung der personenbezogenen Varianzkomponente (R^2_2) im Vergleich zu M2.0. Die Modelle M2.1 bis M2.4 werden mit Likelihood-Ratio-Tests mit M2.0 verglichen. Das finale Modell M2.5 wird mit dem vollen Modell M2.4 verglichen, um zu zeigen, dass der Verzicht auf die entsprechenden Koeffizienten das Modell nicht signifikant verschlechtert. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.4: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M2)

	M2.0	M2.1	M2.2	M2.3	M2.4	M2.5
Intercept	0.22	1.82	1.23*	0.74	1.78	1.78*
	(0.64)	(1.38)	(0.72)	(0.69)	(1.49)	(0.75)
Slope	0.11*	0.11	0.24***	0.18**	0.17	0.28***
	(0.06)	(0.12)	(0.06)	(0.06)	(0.13)	(0.06)
Lg. Mappus		0.28			−0.85	
		(1.83)			(1.87)	
Lg. Schmid		−3.20*			−0.69	
		(1.63)			(1.82)	
Slope x Lg. Mappus		0.30*			0.19	
		(0.15)			(0.16)	
Slope x Lg. Schmid		−0.16			0.02	
		(0.14)			(0.16)	
Sk. Mappus			0.61**		0.30	0.62**
			(0.22)		(0.30)	(0.21)
Sk. Schmid			−0.29		0.65	
			(0.37)		(0.50)	
Slope x Sk. Mappus			0.06**		0.03	0.06***
			(0.02)		(0.03)	(0.02)
Slope x Sk. Schmid			−0.08**		−0.02	
			(0.03)		(0.04)	
Sk. CDU				0.63**	0.48	
				(0.21)	(0.31)	
Sk. SPD				−0.76**	−1.06*	−0.64*
				(0.29)	(0.42)	(0.28)
Slope x Sk. CDU				0.05**	0.02	
				(0.02)	(0.03)	
Slope x Sk. SPD				−0.08***	−0.06	−0.07**
				(0.02)	(0.04)	(0.02)

Anmerkungen
Dargestellt sind die REML-Koeffizienten β mit ihren Standardfehlern (s.e) und Signifikanzniveau (Einseitige T-Tests mit Freiheitsgraden nach Satterthwaite-Approximation): $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Lg.: Lager; Sk.: Skalometer. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.5: Vergleich der Modelle zur Erklärung der Bewertung von Mappus während der Antwort durch die Voreinstellungen (M3)

	M3.0	M3.1	M3.2	M3.3	M3.4	M3.5
AICc	26784	26773	26767	26763	26774	26761
Deviance	26771	26752	26745	26741	26732	26739
ΔDeviance (χ^2)		20	26	30	39	−7
Freiheitsgrade		4	4	4	12	−8
Signifikanz		.001	<.001	<.001	<.001	.514
$R^2_{2_{\text{Slope}}}$		.08	.10	.14	.15	.15
$R^2_{2_{\text{Slope}^2}}$		.04	.05	.08	.08	.09

Anmerkungen
Dargestellt sind das Informationskriterium *AICc*, die Kenngrößen der Likelihood-Ratio-Tests und die Reduzierung der personenbezogenen Varianzkomponente (R^2_2) im Vergleich zu M3.0. Die Modelle M3.1 bis M3.4 werden mit Likelihood-Ratio-Tests mit M3.0 verglichen. Das finale Modell M3.5 wird mit dem vollen Modell M3.4 verglichen, um zu zeigen, dass der Verzicht auf die entsprechenden Koeffizienten das Modell nicht signifikant verschlechtert. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.6: Effekte der Voreinstellungen auf die Bewertung von Mappus während der Antwort (M3)

	M3.0	M3.1	M3.2	M3.3	M3.4	M3.5
Slope	0.17	0.40	0.52**	0.36*	0.49	0.67***
	(0.15)	(0.32)	(0.17)	(0.16)	(0.34)	(0.17)
Slope²	0.00	−0.01	−0.01*	−0.01	−0.01	−0.01**
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
Slope x Lg. Mappus		0.52			0.18	
		(0.42)			(0.43)	
Slope² x Lg. Mappus		−0.01			0.00	
		(0.01)			(0.01)	
Slope x Lg. Schmid		−0.71*			−0.04	
		(0.38)			(0.41)	
Slope² x Lg. Schmid		0.02			0.00	
		(0.01)			(0.01)	
Slope x Sk. Mappus			0.19***		0.11	0.19***
			(0.05)		(0.07)	(0.05)
Slope² x Sk. Mappus			0.00**		0.00	0.00**
			(0.00)		(0.00)	(0.00)
Slope x Sk. Schmid			−0.14*		0.12	
			(0.09)		(0.11)	
Slope² x Sk. Schmid			0.00		0.00	
			(0.00)		(0.00)	
Slope x Sk. CDU				0.17***	0.10	
				(0.05)	(0.07)	
Slope² x Sk. CDU				0.00**	0.00	
				(0.00)	(0.00)	
Slope x Sk. SPD				−0.25***	−0.29***	−0.22**
				(0.07)	(0.09)	(0.06)
Slope² x Sk. SPD				0.01**	0.01**	0.00**
				(0.00)	(0.00)	(0.00)

Anmerkungen
Dargestellt sind die REML-Koeffizienten β mit ihren Standardfehlern (s.e) und Signifikanzniveau (Einseitige T-Tests mit Freiheitsgraden nach Satterthwaite-Approximation): $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Lg.: Lager; Sk.: Skalometer. Für alle Modelle L2: n = 136 Rezipienten; L1: n = 4080 RTR-Messungen.

Tabelle A.7: Effekte des Themas und der Lagerzugehörigkeit auf die Bewertung von Schmid während seiner Turns

	M3					M4				
	β	s.e.	t	df	p	β	s.e.	t	df	p
Intercept	4.73	2.02	2.34	56	.023	3.60	2.15	1.67	71	.099
EnBW	0.58	2.85	0.20	19	.844	1.76	3.22	0.55	31	.586
Arbeitsmarkt	1.76	2.29	0.77	19	.451	3.52	2.58	1.36	31	.184
Kita/Kiga	3.10	2.86	1.09	19	.289	3.57	3.22	1.11	31	.275
Schule	−3.49	2.29	−1.52	19	.145	−1.15	2.59	−0.44	32	.663
Studiengebühren	2.82	3.74	0.76	19	.456	1.24	4.22	0.29	32	.774
Finanzen	−2.69	2.16	−1.25	19	.226	−1.98	2.44	−0.81	32	.424
Persönliches	−1.74	2.49	−0.70	19	.492	−1.84	2.82	−0.65	32	.520
S21/Bürgerbeteiligung	−4.36	2.07	−2.11	19	.048	−1.77	2.34	−0.75	32	.459
Lager Schmid	8.51	1.61	5.29	169	<.001	11.42	1.90	6.01	325	<.001
Lager Mappus	−3.80	1.82	−2.09	169	.038	−5.03	2.15	−2.34	325	.020
EnBW X Lg. Schmid						−2.18	2.05	−1.06	4904	.289
Arbeitsmarkt X Lg. Schmid						−2.86	1.65	−1.74	4955	.082
Kita/Kiga X Lg. Schmid						−3.31	2.06	−1.61	4926	.107
Schule X Lg. Schmid						−3.81	1.66	−2.29	4999	.022
Studiengeb. X Lg. Schmid						1.00	2.72	0.37	4928	.711
Finanzen X Lg. Schmid						−2.14	1.57	−1.37	4968	.171
Persönliches X Lg. Schmid						−4.00	1.81	−2.21	5057	.027
S21/Bürgerb. X Lg. Schmid						−5.70	1.50	−3.80	4982	<.001
EnBW X Lg. Mappus						−0.37	2.34	−0.16	4903	.873
Arbeitsmarkt X Lg. Mappus						−1.17	1.85	−0.63	4955	.529
Kita/Kiga X Lg. Mappus						4.52	2.34	1.93	4925	.054
Schule X Lg. Mappus						−1.58	1.89	−0.84	4999	.401
Studiengeb. X Lg. Mappus						4.12	3.01	1.37	4927	.171
Finanzen X Lg. Mappus						1.35	1.76	0.77	4968	.441
Persönliches X Lg. Mappus						7.81	2.04	3.84	5056	<.001
S21/Bürgerb. X Lg. Mappus						1.15	1.69	0.68	4981	.497

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade. n_{Turns} = 34, $n_{Rezipienten}$ = 172, $n_{Messmodelle}$ = 5456, $n_{RTR\text{-}Messungen}$ = 224923.

Tabelle A.8: Effekte der Relation und der Lagerzugehörigkeit auf die Bewertung von Schmid während seiner Antworten

	β	s.e.	t	df	p
Intercept	−0.48	0.62	−0.77	167	.442
Lager Schmid	1.71	0.64	2.68	201	.008
Lager Mappus	0.85	0.72	1.18	203	.240
Angriff	−0.38	0.73	−0.52	69	.606
Verteidigung	0.82	1.11	0.74	72	.462
Lg. Sch. X Angr.	0.44	0.57	0.77	4564	.444
Lg. Map. X Angr.	−0.55	0.65	−0.85	4571	.396
Lg. Sch. X Vert.	−0.72	0.87	−0.83	4617	.408
Lg. Map. X Vert.	−1.51	0.99	−1.52	4618	.129
Zeit	0.21	0.07	3.00	176	.003
Lg. Sch. X Zeit	0.28	0.07	3.88	200	<.001
Lg. Map. X Zeit	−0.08	0.08	−0.95	202	.342
Angr. X Zeit	−0.03	0.08	−0.40	63	.692
Vert. X Zeit	0.01	0.12	0.10	65	.924
Lg. Sch. X Angr. X Zeit	0.07	0.06	1.22	4400	.225
Lg. Map. X Angr. X Zeit	−0.03	0.07	−0.49	4411	.625
Lg. Sch. X Vert. X Zeit	−0.09	0.09	−1.02	4478	.308
Lg. Map. X Vert. X Zeit	−0.34	0.10	−3.40	4482	.001

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.

Tabelle A.9: Effekte der Relation und der Lagerzugehörigkeit auf die Bewertung von Mappus während seiner Antworten

	β	s.e.	t	df	p
Intercept	−0.23	0.59	−0.39	197	.697
Lager Schmid	−0.75	0.63	−1.19	238	.234
Lager Mappus	0.33	0.71	0.46	243	.648
Angriff	−0.46	0.81	−0.57	101	.569
Verteidigung	−0.53	0.64	−0.83	94	.410
Lg. Sch. X Angr.	−2.11	0.73	−2.88	4532	.004
Lg. Map. X Angr.	0.58	0.83	0.70	4529	.483
Lg. Sch. X Vert.	−1.30	0.57	−2.30	4531	.021
Lg. Map. X Vert.	1.09	0.65	1.67	4536	.095
Zeit	0.18	0.08	2.43	207	.016
Lg. Sch. X Zeit	−0.21	0.08	−2.58	204	.011
Lg. Map. X Zeit	0.31	0.09	3.37	206	.001
Angr. X Zeit	0.10	0.09	1.11	82	.271
Vert. X Zeit	−0.05	0.07	−0.73	77	.469
Lg. Sch. X Angr. X Zeit	−0.08	0.07	−1.11	4319	.267
Lg. Map. X Angr. X Zeit	−0.04	0.08	−0.53	4317	.596
Lg. Sch. X Vert. X Zeit	−0.17	0.06	−2.93	4326	.003
Lg. Map. X Vert. X Zeit	−0.04	0.07	−0.64	4328	.522

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.

Tabelle A.10: Effekte der Relation und der Voreinstellungen auf die Bewertung von Schmid während seiner Antworten

	β	s.e.	t	df	p
Intercept	−0.97	0.60	−1.61	148	.110
Zeit	0.17	0.07	2.51	162	.013
Lager Schmid	0.99	0.61	1.61	159	.110
Lager Mappus	1.41	0.69	2.05	168	.042
Skalometer Schmid	0.58	0.14	4.13	162	<.001
Skalometer Mappus	−0.24	0.10	−2.27	188	.024
Verteidigung	0.39	0.92	0.43	35	.672
Angriff	−0.38	0.61	−0.62	34	.539
Vert. X Lg. Mappus	−1.30	0.76	−1.70	4604	.089
Angr. X Sk. Mappus	−0.06	0.08	−0.73	4554	.468
Zeit X Lg. Schmid	0.20	0.07	2.81	166	.006
Zeit X Lg. Mappus	−0.01	0.08	−0.12	172	.906
Zeit X Sk. Schmid	0.02	0.02	1.49	168	.138
Zeit X Sk. Mappus	−0.03	0.01	−2.54	189	.012
Zeit X Verteidigung	−0.05	0.10	−0.55	34	.586
Zeit X Angriff	−0.04	0.07	−0.63	33	.531
Zeit X Vert. X Lg. Map.	−0.27	0.08	−3.55	4448	<.001
Zeit X Angr. X Sk. Map.	−0.03	0.01	−3.59	4370	<.001

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.

Tabelle A.11: Effekte der Relation und der Voreinstellungen auf die Bewertung von Mappus während seiner Antworten

	β	s.e.	t	df	p
Intercept	−0.29	0.47	−0.62	119	.535
Zeit	0.31	0.06	5.10	157	<.001
Lager Schmid	0.40	0.55	0.72	174	.469
Skalometer Schmid	−0.27	0.13	−1.99	211	.048
Skalometer Mappus	0.22	0.10	2.07	191	.040
Skalometer CDU	0.14	0.10	1.38	165	.169
Angriff	−0.16	0.68	−0.24	48	.812
Verteidigung	0.06	0.51	0.12	36	.906
Angriff X Lg. Schmid	−2.44	0.57	−4.27	4533	<.001
Verteidigung X Sk. Schmid	−0.34	0.13	−2.67	4540	.008
Verteidigung X Sk. Mappus	0.56	0.08	7.07	4544	<.001
Zeit X Lg. Schmid	−0.16	0.08	−2.08	169	.039
Zeit X Sk. Schmid	−0.05	0.02	−2.51	188	.013
Zeit X Sk. Mappus	0.04	0.01	2.73	178	.007
Zeit X Sk. CDU	0.03	0.01	1.95	165	.053
Zeit X Angriff	0.07	0.07	0.88	44	.384
Zeit X Verteidigung	−0.11	0.06	−1.95	35	.060
Zeit X Angr. X Lg. Sch.	−0.05	0.06	−0.80	4316	.426
Zeit X Vert. X Lg. Sch.	0.01	0.01	0.82	4327	.409
Zeit X Vert. X Sk. Map.	0.03	0.01	3.94	4332	<.001

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.

Tabelle A.12: Effekte der Relation und der Lagerzugehörigkeit auf die Veränderung der Bewertung von Schmid nach Relationswechseln

	β	s.e.	t	df	p
Intercept	1.88	1.47	1.28	216	.202
Lager Schmid	7.07	1.41	5.01	185	$<$.001
Lager Mappus	−3.91	1.58	−2.47	185	.014
Angriff	−1.98	1.96	−1.01	75	.315
Neg. Lagebeschreibung	−1.76	2.51	−0.70	75	.485
Lg. Schmid X Angriff	1.64	0.92	1.78	10425	.075
Lg. Mappus X Angriff	0.22	1.03	0.21	10425	.830
Lg. Schmid X Neg. Lagebes.	0.29	1.18	0.24	10426	.808
Lg. Mappus X Neg. Lagebes.	0.99	1.32	0.75	10424	.453
Zeit	0.21	0.08	2.64	171	.009
Zeit X Lg. Schmid	0.12	0.07	1.71	231	.088
Zeit X Lg. Mappus	−0.07	0.08	−0.92	231	.357
Zeit X Angriff	0.03	0.13	0.24	102	.814
Zeit X Neg. Lagebes.	0.31	0.17	1.83	102	.070
Zeit X Lg. Sch. X Angr.	0.09	0.09	0.95	10420	.342
Zeit X Lg. Map. X Angr.	0.02	0.10	0.18	10418	.855
Zeit X Lg. Sch. X Neg. Lagebes.	0.01	0.12	0.08	10419	.936
Zeit X Lg. Map. X Neg. Lagebes.	−0.22	0.13	−1.69	10415	.092

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.
$n_{\text{Relationswechsel}} = 61$, $n_{\text{Rezipienten}} = 176$, $n_{\text{Messmodelle}} = 10667$, $n_{\text{RTR-Messungen}} = 106212$.

Tabelle A.13: Effekte der Relation und der Lagerzugehörigkeit auf die Veränderung der Bewertung von Mappus nach Relationswechseln

	β	s.e.	t	df	p
Intercept	1.61	1.45	1.11	234	.267
Lager Schmid	−5.75	1.45	−3.96	195	<.001
Lager Mappus	4.46	1.63	2.73	195	.007
Angriff	4.61	1.61	2.86	85	.005
Verteidigung	−1.27	1.70	−0.75	85	.457
Lg. Schmid X Angriff	−2.56	0.92	−2.79	10429	.005
Lg. Mappus X Angriff	4.25	1.03	4.13	10428	<.001
Lg. Schmid X Verteidigung	2.41	0.96	2.50	10427	.012
Lg. Mappus X Verteidigung	−0.50	1.08	−0.46	10427	.646
Zeit	0.21	0.08	2.61	250	.010
Zeit X Lg. Schmid	−0.21	0.08	−2.45	249	.015
Zeit X Lg. Mappus	0.32	0.09	3.40	249	.001
Zeit X Angriff	0.05	0.11	0.46	156	.647
Zeit X Verteidigung	−0.17	0.11	−1.47	156	.143
Zeit X Lg. Sch. X Angr.	−0.08	0.09	−0.90	10435	.369
Zeit X Lg. Map. X Angr.	−0.23	0.10	−2.25	10433	.024
Zeit X Lg. Sch. X Vert.	−0.34	0.10	−3.53	10428	<.001
Zeit X Lg. Map. X Vert.	0.12	0.11	1.07	10428	.285

Anmerkungen
β: REML-Koeffizienten der Fixed Effects; s.e.: Standardfehler; df: Satterthwaite-Approximation der Nennerfreiheitsgrade.
$n_{\text{Relationswechsel}} = 61$, $n_{\text{Rezipienten}} = 176$, $n_{\text{Messmodelle}} = 10668$, $n_{\text{RTR-Messungen}} = 106113$.

B Durchführung der Bootstrap-Peak-Spike-Analyse

Ausgangspunkt

Im Folgenden ist die Durchführung einer Bootstrap-Peak-Spike-Analyse am Beispiel der aggregierten RTR-Zeitreihe der Unentschiedenen dokumentiert. Ausgangspunkt ist ein Datensatz *d* der individuellen RTR-Messungen mit einer kreuzklassifizierten Struktur. In diesem Datensatz sind die folgenden Variablen enthalten:

- *idnr*: Identifikationsnummer des Rezipienten
- *lager*: Lagerzuordnung der Rezipienten mit den Ausprägungen „Lager Schmid“, „Unentschiedene“, „Lager Mappus“
- *sec*: Sekunde der RTR-Messung
- *rtr*: Ausprägung der RTR-Messung eines Rezipienten in einer Sekunde

Benötigte *R* Pakete

- *plyr* (Wickham, 2011)
- *parallel*
- *ggplot2* (Wickham, 2009)

Umsetzung der Bootstrap-Peak-Spike-Analyse

1. Auswahl der Unentschiedenen

   ```
   d = subset(d, lager=="Unentschiedene")
   ```

2. Ziehen der k Bootstrap-Stichproben aus den Rezipienten in d (hier: $k = 1000$), Speichern in einer neuen Datenmatrix $stp.boot$

   ```
   stp.boot = replicate(1000, sample(unique(d$idnr),
       length(unique(d$idnr)), replace=T))
   ```

3. Definition einer Funktion zur Berechnung der aggregierten Zeitreihen für alle Bootstrap-Stichproben

```
f = function(i) {
  gwt = data.frame(table(stp.boot[, i]))
  names(gwt) <- c("idnr", "gwt")
  ddply(merge(d, gwt), .(sec), summarise,
      rtr_m=weighted.mean(rtr, gwt, na.rm=1))
}
```

4. Berechnung der aggregierten Zeitreihen für alle Bootstrap-Stichproben, Speichern in einem neuen Datensatz *zr.boot*

```
zr.boot = do.call(rbind,(mclapply(1:ncol(stp.boot), f,
    mc.cores=8)))
zr.boot$boot = rep(1:ncol(stp.boot),
    each=nrow(zr.boot)/ncol(stp.boot))
```

5. Speichern der Mittelwerte und Standardabweichungen aller Zeitreihen in *zr.char*, Verknüpfen der Datensätze *zr.char* und *zr.boot*, Bestimmen der relativen Grenzen für Peaks für jede Zeitreihe in *zr.boot* als Abweichung um $1.96SD$ vom Mittelwert der Zeitreihe

```
zr.char= data.frame(boot = 1:ncol(stp.boot),
    zr_m=unname(tapply(zr.boot$rtr_m, zr.boot$boot, mean)),
    zr_s=unname(tapply(zr.boot$rtr_m, zr.boot$boot, sd)))
zr.boot = merge(zr.boot, zr.char)
zr.boot$upb = zr.boot$zr_m + 1.96 * zr.boot$zr_s
zr.boot$lob = zr.boot$zr_m - 1.96 * zr.boot$zr_s
```

6. Zählen, in den Zeitreihen wie vieler Bootstrap-Stichproben eine Sekunde als Peak identifiziert wird

```
zr.boot$peak = ifelse(zr.boot$rtr_m > zr.boot$upb |
    zr.boot$rtr_m < zr.boot$lob, 1, 0)
peaks = data.frame(sec = 1:(nrow(zr.boot)/ncol(stp.boot)),
    peak_count = unname(tapply(zr.boot$peak, zr.boot$sec,
    sum)))
```

7. Identifikation der Sekunden, die mit $p < .05$ ein relativer Peak sind (*peaks*1) und der Sekunden, bei denen nicht mit $p < .05$ ausgeschlossen werden kann, dass sie ein Peak sind (*peaks*2)

```
peaks1 = subset(peaks, peak_count > 975)
peaks2 = subset(peaks, peak_count >= 25 & peak_count <= 975)
```

8. Visualisierung der Befunde (vgl. Abbildung 5.11, S. 148)

```
visdata = ddply(zr.boot, .(sec), summarise, rtr = mean(rtr_m,
    na.rm=1))
ggplot(visdata) + geom_vline(data = peaks1, aes(xintercept =
    sec), color = "red", alpha = 0.5) + geom_vline(data =
    peaks2, aes(xintercept = sec), color = "yellow", alpha =
    0.5) + geom_line(aes(sec, rtr), color = "blue")
```

Bisher in der Schwarzen Reihe erschienen

Michael Scharkow (2012): Automatische Inhaltsanalyse und maschinelles Lernen Bereits seit einigen Jahren werden verschiedene Verfahren des maschinellen Lernens für die Auswertung von digitalen Medieninhalten eingesetzt – unter anderem bei Suchmaschinen oder automatischen Übersetzungen. Was leisten diese Verfahren jedoch für die quantitative Inhaltsanalyse, wie sie in den Sozialwissenschaften angewandt wird? In diesem Buch werden die methodologischen und forschungspraktischen Besonderheiten der automatischen Inhaltsanalyse denen der klassischen manuellen Codierung gegenübergestellt. Anschließend werden die Vor- und Nachteile des maschinellen Lernens im Vergleich zu anderen computergestützten Verfahren der Textanalyse diskutiert. Praktisch wird das Potential dieses Ansatzes anhand einer umfangreichen Analyse von Online-Nachrichten evaluiert. In einer experimentellen Untersuchung stehen dabei einerseits die Klassifikationsqualität, andererseits die Effektivität des maschinellen Lernprozesses auf dem Prüfstand.
ISBN: 978-3-8442-1670-7
URL: http://opus.kobv.de/udk/frontdoor.php?source_opus=40

Marko Bachl (2014): Analyse rezeptionsbegleitend gemessener Kandidatenbewertungen in TV-Duellen Die Untersuchung der Bewertung von Kandidaten während einer TV-Debatte mit Real-Time-Response-Messungen hat sich in der politischen Kommunikationsforschung etabliert. Das Studiendesign ermöglicht es, detailliert zu erfassen, wie individuelle Rezipienten die Kandidaten infolge einzelner Aussagen bewerten. Um die Potenziale des aufwändigen Studiendesigns voll ausschöpfen zu können und der komplexen Datenstruktur der Echtzeitmessung sowohl theoretisch als auch statistisch gerecht zu werden, ist eine Reflexion über angemessene Analyseverfahren notwendig. In dieser Arbeit werden zum einen die etablierten analytischen Zugänge kritisch diskutiert und erweitert. Zum anderen wird eine Mehrebenenmodellierung vorgeschlagen, die sich in besonderer Weise eignet, die individuellen Prozesse der Kandidatenbewertungen abzubilden. Die etablierten Verfahren und die Mehrebenenmodellierung werden anhand einer Rezeptionsstudie zum TV-Duell vor der baden-württembergischen Landtagswahl 2011 praktisch demonstriert.
ISBN: 978-3-7375-0138-5