Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey

Complexity, Learning Effects and Plausibility of Vignettes in the Factorial Survey Design

Katrin Auspurg, Thomas Hinz und Stefan Liebig

Zusammenfassung

Der Faktorielle Survey gilt als eine Erhebungsmethode, bei der sich die Vorteile der Umfrageforschung mit denen experimenteller Designs verbinden. Statt einzelner Items bewerten die Befragten hypothetische Objekt- oder Situationsbeschreibungen. Indem in diesen ‚Vignetten' einzelne Merkmalsausprägungen experimentell variiert werden, lässt sich ihr Einfluss auf die abgefragten Urteile oder Entscheidungen exakt bestimmen und damit das Gewicht von Faktoren isolieren, die in der Realität oftmals konfundiert sind. Bislang liegen allerdings nur sehr wenige Methodenstudien zur Validität der erzielten Messungen vor. Der Beitrag gibt zunächst einen knappen Überblick zum Einsatz des Faktoriellen Surveys in der sozialwissenschaftlichen Forschung und benennt anschließend bislang ungeklärte methodische Probleme. Die mit einer eigenen experimentellen Datenerhebung durchgeführten Analysen beziehen sich auf die Stabilität des Urteilsverhaltens der Befragten in Abhängigkeit von der Anzahl der in den Vignetten abgebildeten Dimensionen, möglichen Lerneffekten sowie von ‚unplausiblen' oder ‚unlogischen' Fällen (Vignettentexte für Situationen, die in der Realität sehr selten oder gar nicht vorkommen und die Befragten daher irritieren könnten). Getestet werden verschiedene Hypothesen zur Komplexität der Erhebungssituation und der Kohärenz der Urteile. Nach

Abstract

The factorial survey is a method of data collection that combines the advantages of survey research and the advantages of experimental designs. Respondents react to hypothetical descriptions of objects or situations (vignettes) instead of answering single-item questions. By varying each dimension of the vignettes in an experimental design, the dimensions' impact on respondents' judgments or decisions can be estimated accurately. Thus, the method is able to identify the effect of single factors which are often confounded in reality. So far, only few methodological studies address questions of measurement validity when a factorial survey design is used. The article provides a brief overview of the use of the factorial design in the social sciences and points our still unresolved methodological questions. Using experimental data specifically designed for this purpose our analyses consider the stability of respondents' judgments with respect to the number of dimensions presented in the vignettes, possible learning effects and ‚implausible' or ‚illogical' cases (vignettes describing objects or situations which are rare or even impossible). We test several hypotheses regarding the complexity of vignettes and the consistency of judgments. According to our results, a high complexity of vignettes and implausible cases cause respondents to consider

unseren Ergebnissen führen eine hohe Komplexität der Vignetten und unplausible Fälle zu einem weniger Vignettendimensionen einbeziehenden Urteilsverhalten, damit geringeren Einflussstärken einzelner Vignettenmerkmale bei gleich bleibender Konsistenz. Abschließend diskutieren wir die praktischen Konsequenzen dieser Befunde.

fewer dimensions in their judgments; we find smaller influences of vignette variables while the consistency of the judgments remains the same. Finally, we discuss the practical consequences of these results.

1 Einleitung[1]

Der Faktorielle Survey ist eine in Umfragen einsetzbare experimentelle Methode, bei der den Befragten hypothetische Objekt- oder Situationsbeschreibungen (*Vignetten*) vorgelegt werden.[2] Die Vignetten unterscheiden sich nach Merkmalen (*Dimensionen*), die in ihren Ausprägungen (*levels*) variieren. Solche hypothetischen Fälle und Szenarien, die Befragte beurteilen oder bewerten, werden heute in verschiedenen akademischen und nicht-akademischen Forschungszusammenhängen vermehrt eingesetzt, neben den Sozialwissenschaften etwa auch in den Gesundheitswissenschaften, der Rechtswissenschaft, der Psychologie und der Marktforschung. Thematisch zeigen die Studien in der Soziologie eine beachtliche Breite. In der Norm- und Werteforschung beschäftigen sie sich mit der Messung von Status und Prestige von Individuen und Haushalten (Rossi 1979; Rossi et al. 1974; Meudell 1982; Nock 1982), den Vorstellungen über ein gerechtes Erwerbseinkommen (Alves/Rossi 1978; Hermkens/Boerman 1989; Jann 2003; Jasso 1994; Jasso/Webster 1997, 1999; Shepelak/Alwin 1986), der Bewertung von Armutsdimensionen (Will 1993), den Kriterien zur Festlegung wohlfahrtsstaatlicher Unterstützungszahlungen (Liebig/Mau 2002), gerechten Steuersätzen (Liebig/Mau 2005) und Entlassungsverfahren (Struck et al. 2008). Ebenso liegen Arbeiten vor zur Bewertung von sexuellem Missbrauch/sexueller Belästigung (Garrett 1982; Rossi/Anderson 1982; O´Toole et al. 1999), zu der Bestrafung und dem Umgang mit Straftätern (Berk/Rossi 1977;

1 Der Beitrag entstand im Rahmen des von der DFG geförderten Forschungsprojekts ‚Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen'. Projektleiter sind Thomas Hinz (Universität Konstanz) und Stefan Liebig (Universität Bielefeld). Die Autoren danken Peter Steiner sowie einem anonymen Gutachter für wertvolle Hinweise und Anmerkungen. Für die Unterstützung bei der Organisation der Feldphase bedanken wir uns bei Judith Tonner.

2 Ursprünglich wurde der faktorielle Survey in den Sozialwissenschaften 1951 von Peter H. Rossi in seiner Dissertation entwickelt und zur Einschätzung des sozialen Status von Haushalten verwendet (Alves/Rossi 1978; Rossi 1979; Rossi/Nock 1982). Rossis zentrales Anliegen war es, ein Messverfahren zu entwickeln, das es ermöglicht herauszufinden, welche Objekteigenschaften in welchem Ausmaß für soziale Einstellungen relevant sind (Rossi/Anderson 1982: 15ff.; Rossi/Nock 1982: 9ff.).

Hembroff 1987; Miller/Rossi/Simpson 1986), unterschiedlichen Kriterien der Einbürgerung (Jasso 1988), zur Vergabe medizinischer Hilfen (Hechter et al. 1999), zur Qualität von Kinderbetreuungsmaßnahmen (Shlay et al. 2005) und zum sozialen Kontext von Normgeltung (Beck/Opp 2001; Diefenbach/Opp 2007; Horne 2003; Jasso/Opp 1997). Ferner existieren Arbeiten, die der Frage möglicher Diskriminierungen nachgehen (Jann 2003; John/Bates 1990), Effekte sozialer Einbettung analysieren (Buskens/Weesie 2000) oder familiensoziologische Theorien untersuchen (Auspurg/Abraham 2007). Angesichts des großen und vielfältigen Interesses für diese Erhebungsmethode verwundert es, dass methodische und modelltheoretische Fragen sehr selten diskutiert werden (Ausnahmen: Dülmer 2001, 2007; Dülmer/Klein 2003; Steiner/Atzmüller 2006). Wenn sie thematisiert werden, so besteht das Anliegen meistens darin, die Vorteile dieser Befragungsmethode gegenüber itembasierten Abfragen oder den traditionellen experimentellen Vorgehensweisen zu unterstreichen (Hechter/Kim/Baer 2005; Jasso 1988). Die im Verfahren angelegten methodischen Probleme waren dagegen kaum Gegenstand einer expliziten Untersuchung. Dies gilt insbesondere für Probleme, die sich aus der Anlage und Durchführung eines Faktoriellen Surveys und dem Einsatz von Vignetten in Umfragen ergeben.

Wir verfolgen daher das Ziel, drei miteinander verbundene und als besonders relevant geltende methodische Probleme zu diskutieren und anhand von empirischen Tests zu untersuchen. Dies sind *erstens* die Effekte der Komplexität der den Befragten geschilderten Situation und *zweitens* die hiermit in Verbindung stehenden Lerneffekte bei wiederholter Präsentation von Vignetten. Da das Risiko von unplausiblen Fällen mit der Komplexität steigt und diese zudem als ursächlich für Lerneffekte in Form vereinfachter Entscheidungsheuristiken gelten, analysieren wir *drittens* die Auswirkungen unplausibler Vignetten auf das Urteilsverhalten. Diese Aspekte wurden nach unserem Wissen für Vignettenstudien allesamt noch nicht gezielt untersucht. Eine Beschäftigung mit methodischen Effekten scheint für eine Verbesserung der Datenqualität jedoch dringend angeraten, auch um möglichen Fehlschlüssen gezielt vorzubeugen (die andernfalls beim Vergleich verschieden komplexer Vignettenstudien oder kognitiv unterschiedlich belastbarer Probandengruppen zu befürchten sind) – sei es durch ihren Einbezug bei der Vignettenkonstruktion, Datenauswertung und/oder Ergebnisinterpretation.

Die Gliederung ist wie folgt: Zunächst werden die Verfahrensweise des Faktoriellen Surveys sowie der Stand der Methodendiskussion knapp vorgestellt (Abschnitt 2). Dann werden ausgehend vom Forschungsstand Hypothesen zu den genannten Problemstellungen abgeleitet (Abschnitt 3) und auf der Grundlage einer experimentellen Online-Erhebung getestet (Abschnitte 4 und 5). Schließlich werden die Ergebnisse diskutiert und weiterer Analysebedarf aufgezeigt (Abschnitt 6).

2 Faktorieller Survey: Aufbau, Motivation und Probleme

Der Faktorielle Survey zielt darauf ab, die *relativen* Gewichte einzelner Objekt- oder Situationsmerkmale für Einstellungen, Bewertungen oder Entscheidungen zu bestimmen (für detaillierte Einführungen Beck/Opp 2001; Jasso 2006; Rossi/Anderson 1982). Dazu sind zunächst die in den Vignetten enthaltenen Merkmalsdimensionen und ihre Ausprägungen nach theoretischen Vorüberlegungen auszuwählen. In den Befragungssituationen werden diese Ausprägungen dann experimentell variiert, um zu prüfen, ob die gezielt erzeugte Variation der Objekt- und Situationsmerkmale eine entsprechende Variation der Urteile der Befragten nach sich zieht. In den Auswertungen lassen sich damit die exakten Beziehungen zwischen den Merkmalen und den Urteilen der Befragten ermitteln.

In der Durchführung Faktorieller Surveys werden die Befragten in der Regel also mit mehreren, zufällig oder systematisch ausgewählten Vignetten konfrontiert.[3] Die Befragungsmethode hat gegenüber itembasierten Survey-Studien vier wesentliche Vorteile. *Erstens* erlaubt sie eine Konstruktion von Objekten und Situationen, bei denen eine Mehrzahl solcher Merkmale zusammentreten, die in der Realität oft stark miteinander korrelieren und deswegen keine getrennte Einschätzung ihrer Bedeutung erlauben. Im experimentellen Design des Faktoriellen Surveys lassen sich diese Faktoren isolieren, im technischen Sinn zueinander orthogonal setzen. Die so erzeugte Unkorreliertheit der Merkmale ermöglicht eine separate Bestimmung ihres jeweiligen Einflusses auf Urteil und Entscheidung. *Zweitens* können entsprechende Forschungshypothesen im Unterschied zur klassischen Laborforschung auf der Grundlage größerer (Zufalls-)Stichproben in Bevölkerungsumfragen überprüft werden. *Drittens* eröffnen sich interessante Analysemöglichkeiten, wenn den Befragten mehrere Vignetten vorgelegt und deshalb pro Befragten mehrere Urteile erzielt werden. Dadurch entsteht eine hierarchische Mehrebenenstruktur, die genutzt werden kann, um zwischen ‚between-' und ‚within-subject'-Faktoren zu unterscheiden. Es ist möglich, die Kovariation des Einflusses von Vignetten- und

3 Dies stellt auch das Standardvorgehen in der vornehmlich in der Marktforschung verwendeten und dem Faktoriellen Survey ähnlichen Conjoint-Analyse dar (Carroll/Green 1995). Hier werden den Probanden meist simulierte oder echte Produktbeschreibungen vorgelegt und anschließend die relativen Nutzenwerte je Produktmerkmal ermittelt. Die Produkte weisen wie die Vignetten ein mehrfaktorielles Merkmalsbündel auf (Klein 2002; Orme 2006). Geht es um die Ermittlung von Entscheidungen, werden dagegen zum Teil nur wenige, in manchen Fällen nur eine einzige Vignette präsentiert. Es gibt durchaus Argumente, bei randomisierter Verteilung der Vignetten auf die Befragten nur eine einzige Vignette zu präsentieren: Die Effekte sozialer Erwünschtheit sowie die in diesem Aufsatz thematisierten Lerneffekte werden vermindert. In solchen Studien muss auch kein Mehrebenendesign bemüht werden (Jann 2003).

Befragtenmerkmalen auf die Urteile zu ermitteln. *Viertens* kann mit Faktoriellen Surveys einem gewichtigen Vorwurf an die konventionelle Einstellungsmessung begegnet werden, die Analyse lediglich einzelner Item-Werte würde der komplexen Struktur von Einstellungen nicht gerecht (Jasso/Opp 1997: 949; Liebig/Mau 2002: 114-116). Im Faktoriellen Survey sind komplexe Beurteilungs- und Entscheidungsprobleme simulierbar, indem eine Vielzahl von Merkmalen gekreuzt wird. Dies gilt insbesondere für solche Objekte und Situationen, bei denen verschiedene Objekt- oder Situationsmerkmale in unterschiedlichem Grad urteilsrelevant werden und bei denen der soziale Kontext einer Entscheidungssituation eine wichtige Rolle spielt. So wird beispielsweise die Höhe eines als gerecht empfundenen Erwerbseinkommens für einen Erwerbstätigen oder das gerechte Strafmaß für einen Verurteilten an das Vorliegen verschiedener Bedingungen gekoppelt sein. Genau diese Bedingungen können im Rahmen des Faktoriellen Surveys berücksichtigt und ‚alltagsnah' simuliert werden. Durch eine solche ‚Verbundmessung' könne – so die Argumentation einiger Autoren (Hechter/Kim/Baer 2005; Jasso 1988; Dülmer/Klein 2003) – eine validere Messung von Einstellungen erzielt werden als durch itembasierte Verfahren. Denn die Einstellungen zu den einzelnen Dimensionen werden nicht sequenziell, sondern in der Situationsbeschreibung gemeinsam erfragt. Darüber hinaus verhindere die wiederholte Bewertung einer größeren Anzahl von Objekten und Situationen, dass Befragte ein ‚falsches' oder ‚künstliches' Bild ihrer Einstellungen zeichnen (Hechter et al. 1999). Tatsächlich haben Vergleiche von item- und vignettenbasierten Messungen gezeigt, dass über Faktorielle Surveys erfasste Einstellungen weniger durch soziale Erwünschtheit verzerrt werden (Jann 2003; Liebig/Mau 2002; Smith 1986). Vor diesem Hintergrund resümieren Dülmer/Klein (2003), dass über die Vignettenanalyse eine vergleichsweise exakte Einstellungsmessung möglich sei (siehe auch Hechter/Kim/Baer 2005: 103; Jasso 1988).

Von Kritikern des Faktoriellen Surveys werden aber auch eine ganze Reihe von Nachteilen bzw. Unzulänglichkeiten genannt. Grundsätzliche Einwände beziehen sich zunächst auf den vergleichsweise hohen zeitlichen Befragungsaufwand und die daraus resultierenden Opportunitätskosten bezüglich der Erhebung alternativer Items (Sniderman/Grob 1996). Die Bewertung von zehn und mehr Vignetten ist zeitlich aufwändiger als eine entsprechende itembasierte Abfrage der Dimensionen (Dülmer/Klein 2003; Liebig/Mau 2002). Als problematischer wird jedoch angesehen, dass bei Vignettenstudien vergleichsweise starke Antworteffekte plausibel sind, die sich aus der Auswahl der Beispiele (z. B. Kontrasteffekte), deren Reihenfolge (*carry-over*-Effekte) oder aus der Komplexität der präsentierten Beispiele ergeben können. Mit Faktoriellen Surveys erhobene Einstellungen wären daher höchst instabil und letztlich Artefakte. Letzteres sei insbesondere dann zu erwarten, wenn

die Befragten aufgrund der hohen Komplexität der Bewertungsaufgabe überfordert seien. Sie würden mitunter solche Dimensionen in ihr Antwortverhalten einfließen lassen, denen sie ‚eigentlich' gar keine Bedeutung zumessen. Kritisch angemerkt wird diesbezüglich zudem die Gefahr einer zu starken oder ausschließlichen Konzentration der Befragten auf ein in sich stimmiges Antwortverhalten (Faia 1980; Seyde 2005). Ferner können mögliche Kontexteffekte durch Namen, Begriffe oder Bezeichnungen entstehen und Störeffekte hervorrufen, die aus den individuellen Erfahrungen der Befragten stammen (welche den unterschiedlichsten Alltagssituationen inhärent sind) und kaum kontrollierbar sind.[4] Diese Einwände konnten bislang aufgrund fehlender Methodenstudien weder bestätigt noch entkräftet werden.

Der vorliegende Beitrag bezieht sich auf derartige Forschungslücken und entstand in der ersten Phase eines breiter ansetzenden, von der Deutschen Forschungsgemeinschaft (DFG) finanzierten Projekts der Universitäten Konstanz und Bielefeld.[5] Die hier präsentierten Analysen konzentrieren sich auf folgende drei Aspekte: (1) Zunächst geht es um die Bestimmung einer noch handhabbaren Komplexität der geschilderten Situationen (Beck/Opp 2001: 287; Rossi/Anderson 1982: 59). Diese wird anhand der Menge an variablen Dimensionen untersucht. Mögliche kognitive Über- bzw. Unterforderungen sind allerdings nicht unabhängig von Lerneffekten durch die wiederholte Bearbeitung von Vignetten zu beurteilen, weshalb wir als weiteren Aspekt (2) die Konsistenz des Urteilsverhaltens im Bearbeitungsverlauf analysieren. Schließlich adressieren wir (3) die Auswirkung von unplausiblen Fällen, die – wie noch ausführlicher begründet – ebenfalls in Wechselwirkung mit diesen beiden anderen Aspekten zu sehen ist.

4 Fraglich ist schließlich auch die prognostische Validität des Verfahrens (Rooks et al. 2000), da die Befragten nur hypothetische und nicht aktuelle Entscheidungen treffen (dazu Hechter/Kim/Baer 2005; für Versuche einer externen Validierung Eifler 2007; Groß/Börensen 2009; Nisic/Auspurg 2009).

5 Im Rahmen dieses DFG-Forschungsprojekts werden vielfältige experimentelle Variationen zur Komplexität der Erhebungssituation (Anzahl der Merkmalsdimensionen und Vignetten sowie Relevanz von möglichen Reihenfolgeeffekten) und zur Bedeutung von Darstellungsformen (Bandbreite der Ausprägungen bzw. ‚range'-Effekte, Einflüsse verschiedener Beurteilungsskalen und Präsentationsformen) untersucht. Außerdem gilt die Aufmerksamkeit der zeitlichen Stabilität der Messungen. Umfangreiche Experimentalreihen werden mit einem Studierenden-Sample bearbeitet, es geht darauf aufbauend im Projekt aber ebenso um die Tauglichkeit der Befragungsmethode in *allgemeinen* Bevölkerungsumfragen. Um die Belastbarkeit der Befragten und den Zeitaufwand alters- und bildungsübergreifend einschätzen zu können, werden unterschiedlich komplexe Designs an einer bevölkerungsrepräsentativen Stichprobe in zwei Surveysituationen (‚face-to-face' und schriftlich) getestet. Für nähere Informationen: http://www.uni-konstanz.de/hinz/?cont=faktorieller_survey&lang=de.

3 Forschungsstand und Hypothesen

Im Folgenden berichten wir den Forschungsstand zu den drei benannten methodischen Problemen und leiten daraus Hypothesen zu den Effekten auf das Antwortverhalten ab. Aufgrund der unzureichenden Forschungslage zu Faktoriellen Surveys ziehen wir mitunter Literatur zu verwandten Verfahren der Marktforschung und der Umwelt- und Gesundheitsökonomie heran (Conjoint- und Choice-Experimente).

3.1 Komplexität der Vignetten: Anzahl der Dimensionen

Wie bereits erwähnt, ist der Faktorielle Survey insbesondere für Fragestellungen geeignet, bei denen komplexe Bewertungen vorzunehmen sind. Der Wunsch, über viele Dimensionen eine möglichst detaillierte und ‚alltagsnahe' Beschreibung zu erhalten, kollidiert allerdings mit der eingeschränkten Verarbeitungskapazität der Befragten. Die Entscheidung für eine bestimmte Anzahl von Dimensionen ist somit von weit reichender Bedeutung (Rossi/Anderson 1982). Dies gilt, weil die Anzahl der Dimensionen über die Länge der Situationsbeschreibungen und damit die Komplexität der Bewertungsaufgabe entscheidet. Eine Vielzahl von Dimensionen erzeugt für die Befragten eine möglicherweise nicht mehr oder nur schwer handhabbare Komplexität. Die Folge wäre, dass die entsprechenden Urteile – falls es nicht zum vorzeitigen Abbruch kommt – im ungünstigsten Fall nur noch Artefakte darstellen. Jasso (2006) schlägt vor, nur solche Dimensionen auszuwählen, von denen eine Relevanz für die Bewertung bekannt ist. Dies kann durch theoretische Überlegungen, vorherige Untersuchungen oder aufgrund von Alltagsbeobachtungen geschehen. In Anknüpfung an kognitionspsychologische Arbeiten argumentiert sie zudem, dass Personen nur wenige Dimensionen zur Meinungsbildung heranziehen. Rossi und Anderson (1982) empfehlen, sich auf sechs Dimensionen zu beschränken. In den bislang durchgeführten Faktoriellen Surveys reicht die Anzahl der verwendeten Dimensionen unseres Wissens von drei (Berk/Rossi 1977) bis 21 (Shlay et al. 2005). In der Mehrzahl der Studien werden fünf bis sieben Dimensionen verwendet. Man stützt sich dabei allerdings nur auf eine ‚Daumenregel' aus den Informations- und Kognitionswissenschaften, wonach Menschen sieben plus/minus zwei Informationen am besten verarbeiten können (Zimbardo 1988: 275). Es zeigt sich also, dass die bisherige Forschungspraxis durch sehr unterschiedliche Vorgehensweisen bestimmt ist. Die in der Literatur zu findenden Empfehlungen gehen über allgemeine Ratschläge nicht wirklich hinaus, etwa wenn Beck und Opp

(2001: 287) raten, die Ausprägungen aus Hypothesen zu generieren und nur solche zu verwenden, bei deren Variation man einen tatsächlichen Einfluss vermutet.[6]

Die zunächst nahe liegende, grundsätzliche Annahme lautet, dass die kognitive Anforderung für die Befragten mit der Anzahl der Dimensionen steigt, bis hin zu einer eventuell nicht mehr handhabbaren Komplexität (Rossi/Anderson 1982; für Choice- und Conjoint-Analysen Melles 2001; DeShazo/Fermo 2002). Weitaus weniger klar ist, wie sich die dann zu erwartende Tendenz zur Vereinfachung äußert. Neben einem kompletten Befragungsabbruch und Item-Nonresponses kommt ebenso ein inkonsistenteres Antwortverhalten in Frage. Alternativ sind Heuristiken in Form eines vollständigen Ausblendens inhaltlich weniger relevanter (oder vergleichsweise unauffällig operationalisierter, da z. B. mit weniger Ausprägungen vorgegebener) Dimensionen erwartbar (Wason/Polonsky/Hyman 2002; für Befunde bei Choice- und Conjoint-Analysen Swait/Adamowicz 2001; Melles 2001; DeShazo/Fermo 2002). Vertreten wird bei Choice- und Conjoint-Analysen zudem auch die Gegenhypothese eines *konsistenteren* Antwortverhaltens bei mehr Dimensionen (Sauer 2009). Die dahinter stehende Annahme ist, dass den wenig-dimensionalen Vignetten urteilsrelevante Informationen fehlen, die daher von den Befragten selbst ‚konstruiert' werden müssen.[7] Gegenüber der expliziten Vorgabe durch den Forscher bedeutet die ‚Unterkomplexität' eine geringere inhaltliche Kontrolle über das Vignettenexperiment, was zumindest befragtenübergreifend eine geringere Präzision der Schätzungen erwarten lässt (DeShazo/Fermo 2002; Caussade et al. 2005: 632; Johnson 2006: 46f.). Ähnlich wird vermutet, dass unkontrollierte ‚Framing'-Effekte wahrscheinlicher werden (dazu z. B. Melles 2001: 186). Und schließlich gilt auch ein Informationsmangel als kognitiv belastend, weil es beispielsweise bei wenigen Merkmalsvorgaben schwieriger ist, Unterschiede in den Fallbeispielen zu erkennen und damit zwischen ihnen zu differenzieren (für dieses Argument bei Choice-Experimenten Hensher 2006). Als ein erster Beleg für einen solchen ‚information-underload' können die Befunde einer Wiederholungsbefragung gewertet werden, bei der Studierende zu drei Messzeitpunkten mit den jeweils selben Vignetten befragt wurden: Die Stabilität der Urteile erwies sich bei acht Dimensionen höher als bei fünf Dimensionen (Liebig/Meyermann/Schulze 2006).

Für alle Effekte ist jedenfalls unklar, ab welcher Dimensionszahl mit ihnen zu rechnen ist. Für die vorliegende Untersuchung wird daher mit fünf versus zwölf Di-

6 Neben der Anzahl der Dimensionen ist auch die Zahl der Ausprägungen pro Dimension relevant, weil damit die Größe des ‚Vignettenuniversums' festgelegt wird. Als Vignettenuniversum wird die Gesamtheit aller möglichen Varianten der Situations- bzw. Objektbeschreibungen bezeichnet.

7 In Vignettenstudien zur Einkommensgerechtigkeit könnte ein solches Informationsdefizit z. B. in der Berufserfahrung der Einkommensbezieher bestehen.

mensionen bewusst ein starker Kontrast gewählt. Die – gemessen an den vorliegenden Studien mit überwiegend fünf bis neun Dimensionen – überdurchschnittliche maximale Dimensionszahl von zwölf lässt ein Durchschlagen des ‚Überforderungseffektes' erwarten. Es ergeben sich zwei Unterhypothesen:

H_{1a}: *Bei zwölf Dimensionen sind Befragungsabbrüche häufiger als bei fünf Dimensionen.*

H_{1b}: *Das Urteilsverhalten ist bei zwölf Dimensionen inkonsistenter als bei fünf Dimensionen.*

Alternativ ist von einer vereinfachten Urteilsstrategie in Form einer Ausblendung einzelner Merkmale auszugehen (zu dieser ‚dimensional-reductions'-Strategie bei Choice-Anaylsen: Swait/Adamowicz 2001: 137):

H_{1c}: *Bei zwölf Dimensionen sind einzelne Vignettenvariablen weniger urteilsrelevant, zeigen also geringere Einflüsse auf die Urteile als bei fünf Dimensionen.*

3.2 Lern- und Ermüdungseffekte

In fast allen Vignettenstudien sollen die einzelnen Befragten mehrere Vignetten beurteilen. Gängig sind zehn bis 20 Vignetten, in einer Studie waren es ganze 95 Vignetten pro einzelnem Befragten (Beck/Opp 2001; Rossi et al. 1974). Die mehrfache Präsentation von Vignetten ermöglicht es, selbst bei geringen Befragtenzahlen noch ausreichend viele Urteilszahlen zur Hypothesentestung zu sammeln (Auspurg/Abraham/Hinz 2009). Zudem erlaubt sie, befragtenspezifische Urteils- und Entscheidungsregeln (sog. ‚within-subject'-Effekte) aufzudecken. Mit der wiederholten Bewertungsaufgabe sind allerdings Lerneffekte zu erwarten, die mit anderen Kennzeichen der Erhebungssituation in Wechselwirkung stehen. Sehr deutlich ist dies bei der Anzahl der Dimensionen. Bei einer höheren Dimensionszahl benötigen Lernprozesse länger, gleichzeitig könnten Ermüdungserscheinungen früher einsetzen. Lern- und Ermüdungseffekte sind wechselseitige Aspekte von Komplexität. Beim Lernen geht es um ein zunehmend konsistentes Antwortverhalten sowie um das Vermögen, mehr Dimensionen gleichzeitig in ein Urteil zu integrieren.[8] Ermü-

8 Eine im Befragungsverlauf zunehmende Beachtung von Dimensionen wird zudem damit begründet, dass die Probanden die in der Realität korrelierten Merkmale zu Beginn als redundant ansehen. Erst wenn sie nach einer ganzen Reihe von präsentierten Vignetten erkennen, dass sie im experimentellen Design unabhängig voneinander variieren, schenken sie ihnen mehr Aufmerksamkeit bzw. lassen sie separat in ihr Urteil einfließen (für Conjoint-Analysen Melles 2001: 118).

dungs- und Langeweile-Effekte schlagen sich umgekehrt in einer sinkenden Konsistenz und in einer Beachtung weniger Merkmale oder anderen vereinfachten Entscheidungsregeln nieder (für Choice-Analysen: Carson et al. 1994: 335f.).[9] Die Rolle und das Ausmaß von Lern- und Ermüdungseffekten sind für Vignettenstudien bislang unerforscht. Ebenso ist es eine noch völlig ungeklärte Frage, ab welcher Vignettenzahl mit einem Umkippen von Lern- in Ermüdungseffekte zu rechnen ist.

Als ein erster Orientierungspunkt können Erfahrungen aus den verwandten Choice-Experimenten herangezogen werden. Demnach nimmt die Urteilskonsistenz bis etwa zum zehnten Urteil zu, um danach wieder abzusinken (z. B. Bradley/Daly 1994: 180; Caussade et al. 2005: 631f.). Da selbst bei Vignettenstudien mit 50 oder mehr Vignetten bislang keine nennenswerten Probleme im Hinblick auf die Urteilsgüte berichtet werden (Jasso 2006), scheint bei der vorliegenden Fallzahl von maximal zehn Vignetten pro Befragten (dazu unten Abschnitt 4) eine Dominanz der Lerneffekte plausibel. Es ergeben sich die folgenden Annahmen:

H_{2a}: *Mit der Position der Vignetten steigt die Konsistenz des Antwortverhaltens und/oder die Anzahl berücksichtigter Dimensionen.*

H_{2b}: *Diese Lerneffekte treten stärker bei zwölf als bei fünf Dimensionen auf.*

3.3 Behandlung unlogischer Fälle

Bevor die tatsächlich zu bewertenden Vignetten zusammengestellt werden (also eine Auswahl aus dem Universum aller möglichen Kombinationen von Merkmalsausprägungen getroffen wird; dazu Beck/Opp 2001; Steiner/Atzmüller 2006; Dülmer 2007), ist es bisher gängige Praxis, ‚unlogische und unplausible Fälle' zu eliminieren. Es werden also solche Vignetten ausgeschlossen, die offensichtlich ungewöhnliche oder unsinnige Merkmalskombinationen enthalten. Ein Beispiel dafür wären erwerbstätige Personen ohne Schul- oder Berufsausbildung in einem Beruf, bei dem eine Ausbildung unabdingbar ist (etwa Richter, Hochschullehrer). Der Ausschluss solcher Fälle wird vor allem mit den zu erwartenden Folgen für das Antwortverhalten begründet. Offensichtlich unsinnige Fälle würden die Ernsthaftigkeit der Bewertungsaufgabe in Frage stellen und zu einem Anstieg der Item-Non-Response-Quote, oder gar zum völligen Befragungsabbruch (Faia 1980; Jasso 2006) führen.

9 Grafisch ist also ein umgekehrt u-förmiger Zusammenhang zwischen der Bearbeitungsabfolge der Vignetten und der Konsistenz bzw. Anzahl berücksichtiger Dimensionen zu erwarten.

Dieses Argument ist durchaus plausibel, doch sind die Kriterien, was als unlogisch oder unsinnig zu gelten hat, sehr vage. In vielen Faktoriellen Surveys geht es darum, möglichst unabhängig von den gängigen Normen, bestehenden Gesetzen und empirischen Beobachtungen Bewertungen vornehmen zu lassen, um so auch die kontrafaktischen Meinungen und Überzeugungen der Befragten zu erheben. Die Norm eines ‚logischen Falles' wird durch empirische Regelmäßigkeiten und damit zusammenhängenden Erwartungshaltungen geprägt. Faktorielle Surveys bieten jedoch die seltene Möglichkeit, die Probanden bewusst mit abweichenden Fällen zu konfrontierten – und gerade in der Reaktion auf solche ‚abweichende' Fälle kann ein Erkenntnisziel liegen. In dieser Hinsicht sind Eingriffe in die Merkmalskombinationen problematisch, engen sie doch die Variation der Situations- und Objektbeschreibungen *a priori* auf ein empirisch vorfindbares Maß ein (Beck/Opp 2001).

Solides methodisches Wissen besteht bislang ausschließlich im Hinblick auf die *statistischen* Folgen. Durch den gezielten Ausschluss einzelner Fälle wird die Orthogonalität der Dimensionen im Vignettenuniversum eingeschränkt, Multikollinearität wird also erzwungen (zu deren Konsequenzen für Schätzverfahren: Greene 2003: 56-59; Wooldridge 2003: 96-100). Die Relevanz des Ausschlusses von Fällen für die Balanciertheit und Unkorreliertheit von Vignettensamples ist inzwischen gut einschätzbar (Kuhfeld/Randall/Garratt 1994: 551; Dülmer 2007: 391f.; Steiner/Atzmüller 2006) und es liegen Algorithmen vor, welche die Einbußen an Effizienz gezielt minimieren (dazu Kuhfeld 2005). Aufgrund des andernfalls drohenden Effizienzverlustes lautet daher die eindeutige Empfehlung, diese Algorithmen auch einzusetzen.

Die Auswirkungen der unplausiblen oder unlogischen Fälle auf das *Antwortverhalten* sind dagegen weitaus strittiger, was vor allem durch fehlende einschlägige Untersuchungen bedingt ist.[10] Trifft die oben angesprochene Vermutung zu, dass durch unplausible Vignetten der grundsätzliche Glaube an den Wert der Befragung und damit den Nutzen eigener Mitarbeit beeinträchtigt wird, sind Befragungsabbrüche und invalide Antworten zu erwarten (Response-Sets oder flüchtige und inkonsistente Urteile). Es ergeben sich daher zunächst die folgenden Hypothesen:

H_{3a}: *Werden den Befragten unplausible Fälle vorgelegt, sind Befragungsabbrüche häufiger, als wenn dies nicht der Fall ist.*

H_{3b}: *Werden die Befragten mit unplausiblen Fällen konfrontiert, ist die Konsistenz ihres Antwortverhaltens geringer, als wenn dies nicht der Fall ist.*

10 Die zwischen dem Autorenteam Rossi/Alves (1980) und Faia (1980) ausgetragene Diskussion über die ‚Sinnigkeit' bzw. den Nutzen unplausibler Vignetten ist daher nach wie vor nicht mit empirischen Argumenten zu entscheiden.

Faia (1980) erwartet zudem, dass die für die Unplausibilität ursächlichen Dimensionen in den Vordergrund geraten – die Befragten würden die Aufgabe in einen reinen ‚Intelligenztest' zur Entlarvung von ‚Anomalien' uminterpretieren. Gerade dies würde die Gültigkeit der Urteile beinträchtigen und verdient daher eine Überprüfung:

H_{3c}: *Nach einer Konfrontation mit unplausiblen Fällen beziehen die Befragten primär die für die Unplausibilität verantwortlichen Dimensionen in ihre Urteile ein, gewinnen diese somit relativ zu allen anderen Dimensionen an Bedeutung.*

Als alternative Begründung hierfür lässt sich ein Lerneffekt anführen: Die Befragten bemerken erst bei einer empirisch seltenen Kombination, dass die Merkmale unabhängig voneinander variieren und damit nicht redundant sind. Ähnlich könnte sich so eine sinkende Bereitschaft zu differenzierten Urteilen manifestieren: Dimensionen verlieren durch ein Umschwenken auf ein vereinfachtes, weniger Merkmale einbeziehendes und daher kognitiv weniger belastendes Antwortverhalten an Relevanz.

Die Diskussion dieser drei Problemstellungen verdeutlicht, dass komplexe Wechselwirkungen zwischen den methodischen Aspekten von Faktoriellen Surveys zu erwarten sind. Wir können hier schon aus Platzgründen nur die besonders nahe liegenden Zusammenhänge analysieren, im genannten DFG-Projekt wird derzeit ein weitaus größeres Spektrum methodischer Effekte untersucht.

4 Methodik und Datengrundlage

Die drei methodischen Probleme lassen sich nicht analytisch lösen, sondern erfordern eine empirische Herangehensweise. Ideal dazu ist ein Methodenexperiment, bei dem die Bedeutung von Designelementen für das Antwortverhalten durch ihre gezielte Variation beobachtbar wird. Wichtig ist, dass die methodischen Splits zufällig auf die Befragten verteilt werden und sie zudem nicht mit einzelnen Vignetten(decks) korreliert sind – wie bei jedem Experiment erlaubt erst diese Randomisierung, unbekannte Drittvariablen der Befragten zu neutralisieren und ungewünschte Konfundierungen mit den inhaltlichen Dimensionen der Vignetten zu vermeiden.[11] In den vorliegenden Experimenten wird die Komplexität der Vignetten über die Zahl der Dimensionen variiert: Etwa die Hälfte der Befragten bekommt durchgehend

11 Bei Choice-Experimenten werden derartige Studien unter dem Namen ‚Design of Design' geführt (z. B. Hensher 2004, 2006; Caussade et al. 2005). Im Prinzip handelt es sich um eine mehrfaktorielle Erweiterung des ‚split-ballot'-Designs: Es werden gleich *mehrere* Designelemente unabhängig voneinander variiert (dazu Sniderman/Grob 1996).

Vignetten mit fünf, die andere mit zwölf Dimensionen vorgelegt (es handelt sich also um ein reines ‚between-subject'-Design).[12] Zunächst wurden jedem Teilnehmer sieben Vignetten zugeteilt; aufgrund der geringen Abbruchquote wurde diese Zahl in einer zweiten (kleineren) Befragungswelle auf zehn erhöht.

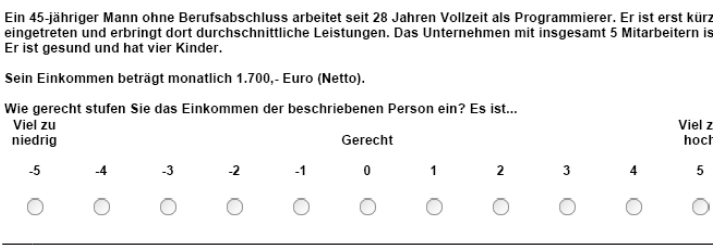
Als inhaltliche Fragestellung dient der besonders gut erforschte ‚Klassiker' von Vignettenstudien – die Erhebung von Einkommensgerechtigkeit (z. B. Alves/Rossi 1978; Jasso/Webster 1997, 1999; Jann 2003; Hermkens/Boerman 1989, Shepelak/Alwin 1986). Den Befragten werden jeweils fiktive Personen vorgestellt, die sich in einer Reihe von einkommensrelevanten Merkmalen unterscheiden, wie dem Geschlecht, Alter, Bildungsstand oder Beruf. Zusätzlich enthält jede Vignette das monatliche Netto-Einkommen der beschriebenen Person. Dieses soll dann auf einer elf-stufigen Ratingskala danach beurteilt werden, ob und in welchem Ausmaß es (un-)gerecht erscheint. Abbildung 1 zeigt eine Beispielvignette mit zwölf Dimensionen. Die Ausprägungen der Dimensionen sind darunter im Überblick aufgeführt.[13] Bei der Auswahl der Merkmale wurde darauf geachtet, dass ihre Relevanz für das Urteilsverhalten bereits belegt ist. Damit sollte sichergestellt werden, dass eine mögliche Nicht-Beachtung methodisch und nicht inhaltlich zu deuten ist (ähnlich für Choice-Experimente Hensher 2006: 16).

Um die zufällige Variation der experimentellen Splits und Vignetten mit verhältnismäßig wenig Aufwand umsetzen zu können, fiel die Wahl auf eine Online-Befragung. Ein weiterer Grund für diesen Befragungsmodus ist die gute Erfassbarkeit von Metadaten (z. B. Beantwortungszeiten), welche zusätzlichen Aufschluss über die Bearbeitungsstrategien versprechen. Bei Experimenten kommt es nicht auf eine repräsentative und zufällige Stichprobe der Probanden an, sondern es sind zumindest bei kleinen Stichproben homogene Experimentalgruppen vorteilhaft (da diese ein geringeres Risiko ungleich verteilter Drittvariablen bergen; z. B. Diekmann 2007: 337ff.). Ihre relativ große Homogenität und ihre gute Erreichbarkeit sprachen für die Wahl von Studierenden verschiedener Universitäten, die über E-Mail-Verteiler der Fachschaften kontaktiert und mit einem Link zur Befragung um ihre Teilnahme gebeten wurden.

12 Dies ist nicht ganz korrekt, denn als zweiter experimenteller Faktor wurde eine der Dimensionen, das Geschlecht der Vignettenpersonen, nur bei einem Teil der Befragten zwischen den Vignetten variiert (‚within'-Variation). Den anderen Befragten wurden stets nur Vignetten eines Geschlechts vorgelegt (‚between'-Variation), sie bewerteten also durchgehend jeweils nur Beschreibungen mit männlichen oder weiblichen Protagonisten, womit sich für sie die Anzahl variabler Dimensionen auf vier bzw. elf Merkmale reduziert. Der Hintergrund dieses Splits ist der, dass sich damit Effekte sozialer Erwünschtheit bzw. eines bewussten vs. unbewussten Urteilsverhaltens untersuchen lassen. Da dieser Faktor aber vollständig unabhängig variiert wurde, kann er an dieser Stelle und den nachfolgenden Analysen ausgeblendet werden – er verdient eine eigenständige Betrachtung.

13 Diese Aufstellung aller Dimensionen dient hier nur als Information für den Leser; den Befragten wurde diese Übersicht nicht vorgelegt.

Abbildung 1 Beispielvignette mit zwölf Dimensionen



Vignettendimensionen und Ausprägungen:

1) *Alter:* 25, 35, 45, 55 Jahre
2) *Geschlecht:* Mann, Frau
3) *Berufsabschluss:* ohne Berufsabschluss, mit abgeschlossener Berufsausbildung, mit Hochschulabschluss
4) *Beruf:* 10 Ausprägungen von Hilfsarbeiter/in bis Anwalt (Auswahl nach Dezentilen der Magnitude-Prestige-Skala)
5) *Einkommen:* 10 Ausprägungen von 250,- bis 15.000,- Euro Netto
6) *Berufserfahrung:* keine, 25%, 50%, 100% der potenziellen Erwerbszeit
7) *Betriebszugehörigkeit:* erst kürzlich eingetreten, schon seit langem im Unternehmen beschäftigt
8) *Leistung:* unterdurchschnittlich, durchschnittlich, überdurchschnittlich
9) *Betriebsgröße:* 5, 20, 200, 2.000 Mitarbeiter
10) *Wirtschaftliche Lage des Unternehmens:* vom Konkurs bedroht, ausgeglichene Bilanz, hohe Gewinne
11) *Gesundheitszustand:* gesund, 30% schwerbehindert
12) *Kinder:* 6 Ausprägungen von keine bis 5 Kinder.

Bei den Vignetten handelt es sich um eine fraktionalisierte Auswahl aus dem kompletten Universum für zwölf Dimensionen, wobei auf eine Orthogonalisierung aller Haupteffekte geachtet wurde (sog. ‚resolution III-Design', Kuhfeld/Randall/Garratt 1994: 546). Mit dieser Anforderung sind bei der vorliegenden Spezifikation von Dimensionen und Ausprägungen etwa 100 Vignetten für eine effiziente Stichprobe hinreichend.[14] Durch den Ausschluss logisch unmöglicher Kombinationen (wie Personen ohne Berufserfahrung, die schon lange im Betrieb arbeiten), reduzierte sich das Sample weiter zu insgesamt 93 unterschiedlichen Vignetten (empirisch seltene, aber gleichwohl mögliche Fälle wurden dagegen bewusst beibehalten – mehr dazu unten). Bei dem Split mit fünf Dimensionen wurde exakt dieselbe Vignettenstichprobe eingesetzt (es wurden einfach die überzähligen Dimensionen gelöscht). Zwar ließen sich für diese ‚sparsameren' Vignetten weitaus effizientere Designs bilden, gerade diese statistischen Effizienzwerte sollten aber konstant gehalten werden, um eine reine Abschätzung der *methodischen* Effekte zu ermöglichen. Nur unter Kontrolle der statistischen Effizienz lassen sich Unterschiede in den Signifikanzen von

14 Es wird eine D-Effizienz von 98,2 erreicht, wobei Werte über 90 als zufrieden stellend gelten (Kuhfeld 2005). Allerdings reduziert sich die Effizienz mit dem Ausschluss unlogischer Fälle wieder.

Regressionskoeffizienten tatsächlich auf das Antwortverhalten zurückführen.[15] Zugleich wird mit der Verwendung identischer Vignettensamples für die Splits mit fünf und zwölf Dimensionen einer Vermischung von inhaltlichen und Designeffekten vorgebeugt. Es lassen sich durch dieses Vorgehen auftretende Unterschiede im Antwortverhalten eindeutiger auf die differente Anzahl an Dimensionen zurückführen statt auf unterschiedliche inhaltliche Kombinationen der Vignettendimensionen.

Alle Teilnehmer wurden zufällig einem der beiden methodischen Splits sowie einem Subset an Vignetten zugewiesen. Pro Befragten wurde eine eigene Zufallsziehung von Vignetten (Ziehung ohne Zurücklegen) vorgenommen. Mit dieser randomisierten Setbildung sollte eine möglichst hohe Ausschöpfung der Stichprobe von 93 Vignetten gewährleistet werden. Zudem wurde eine befragtenspezifische, zufällige Reihenfolge der Vignetten gewählt, um Kontrast- und Reihenfolge-Effekte auszuschließen: ‚Extreme' Vignetten verteilen sich dann zufällig auf die Bearbeitungspositionen, womit über alle Befragten hinweg zu beobachtende Einflüsse der Reihenfolge eindeutiger als Lern- bzw. Ermüdungseffekte zu deuten sind. Die befragtenspezifische Zufallsauswahl von Vignetten hat zudem den Vorteil, dass sich automatisch weitere methodische Variationen zwischen den Befragten ergeben, etwa im Auftreten und der Häufigkeit von unplausiblen Fällen.[16]

Die Befragung fand im Zeitraum Dezember 2007 bis März 2008 statt. Die Vignetten wurden in einen Rahmenfragebogen integriert, in dem neben soziodemographischen Merkmalen politische und soziale Einstellungen über ‚klassische'

15 Schließlich ist für die Präzision der Schätzungen die statistische Effizienz der Vignettenstichprobe ähnlich wichtig wie die ‚kognitive Effizienz' der von den Befragten abgegebenen Urteile (für entsprechende Argumente in Bezug auf Choice- und Conjoint-Analysen Melles 2001: 109; Louviere 2001b).

16 Bei der Alternative einer bewussten bzw. fraktionalisierten Setbildung wären zwar Konfundierungen besser kontrollierbar, aber angesichts der geringen Setgröße von sieben bzw. zehn Vignetten auch nicht vermeidbar – gerade für die komplexere Variante mit zwölf Dimensionen wären der Preis unweigerlich starke Kontexteffekte der einzelnen Sets (selbst Haupteffekte sind innerhalb der einzelnen Sets untereinander korreliert). Aus diesen Gründen ist der Einsatz einer möglichst hohen Anzahl an unterschiedlichen Sets vorzuziehen, zumal angesichts des homogenen Samples und der hohen Befragtenzahl die Gefahr der Konfundierung von Vignetten- mit Befragtenmerkmalen gering erscheint (siehe für eine ausführliche Diskussion der Vor- und Nachteile unterschiedlicher Setbildungen Steiner/Atzmüller 2006). Hinzu kommt, dass fraktionalisierte Setbildungen einem der Analyseziele zuwiderlaufen: Sie arbeiten mit einer möglichst gleichmäßigen Verteilung von Extremfällen, was impliziert, dass auch unplausible Fälle sehr regelmäßig auf die Sets bzw. Befragten verteilt werden und daher die ‚between'-Varianz zu gering ausfallen dürfte, um ihren Einfluss verlässlich zu prüfen. Insgesamt lassen diese Abwägungen somit bei den vorliegenden Analysezielen eine randomisierte Setbildung als vorteilhaft erscheinen. Das mit ihr verbundene Risiko einer unbalancierten Verteilung von Vignetten auf die Splits mit fünf vs. zwölf Dimensionen (bzw. sieben vs. zehn Vignetten) ist angesichts der hohen Set- und Befragtenzahlen gering. Problematisch wären für die angestrebten Analysen insbesondere Unterschiede in den Korrelationsstrukturen. Diese stimmen jedoch in der Tat sehr gut zwischen den einzelnen Splits überein, wie die im Anhang aufgeführten Korrelationsmatrizen belegen (Tabellen A1 und A2).

Itemabfragen erhoben wurden. Den Befragungslink haben 558 Personen aufgerufen, für die Vignetten liegen 3.480 Urteile von insgesamt 460 Probanden vor.[17] Tabelle 1 zeigt die für die einzelnen experimentellen Varianten realisierten Fallzahlen.

Tabelle 1 Realisierte Fallzahlen für Vignettenurteile und Befragte[a]

	5 Dimensionen		12 Dimensionen		Summe	
	Vignetten	Befragte	Vignetten	Befragte	Vignetten	Befragte
Sieben Vignetten pro Befragten	1.213	176	1.109	162	2.322	338
Zehn Vignetten pro Befragten	574	59	584	63	1.158	122
Summe	1.787	235	1.693	225	3.480	460

[a] *Nur Befragte, die mindestens eine Vignette beantwortet haben.*

Bei der Datenauswertung ist die Mehrebenenstruktur zu beachten. Werden Befragten mehrere Vignetten vorgelegt, entsteht ein hierarchischer Datensatz (für eine anschauliche Darstellung Beck/Opp 2001). Auf der untersten Ebene stehen die Vignettenurteile, eine zweite Analyseebene bilden die Merkmale der Befragten. Da wir nur auf die Analyse von Vignettendimensionen (der ersten Ebene) abstellen und zudem ein homogenes Befragtensample verwenden, berücksichtigen wir die Datenstruktur lediglich durch die Schätzung von robusten Standardfehlern (Wooldridge 2003: 258ff., Wooldridge 2002; zur Modellwahl speziell bei Vignettenstudien: Jasso 2006; Auspurg/Abraham/Hinz 2009; Hox/Kreft/Hermkens 1991). Befragtenspezifische Schwankungen der Urteile und ihre mögliche Erklärung interessieren hier nicht. Die für die einzelnen Hypothesen eingesetzten Analysestrategien und Operationalisierungen werden im folgenden Abschnitt erläutert.

17 Aufgrund der verwendeten Samplingprozedur lassen sich keine Rücklaufquoten berichten. An dieser Stelle ist nochmals zu betonen, dass wir lediglich einen experimentellen Hypothesentest, nicht aber deskriptive Aussagen zu Gerechtigkeitseinstellungen anstreben. Dafür scheint der Verzicht auf eine Zufallsstichprobe unproblematisch. Mehrfachteilnahmen wurden so gut wie möglich ausgeschlossen.

5 Ergebnisse

5.1 Deskriptive Befunde

Bevor die Hypothesen mit multivariaten Analysen geprüft werden (Abschnitt 5.2), gibt ein Blick auf die deskriptiven Verteilungen und Rücklaufquoten erste Aufschlüsse über das Antwortverhalten. Insgesamt haben 124 der 558 Teilnehmer (22,2 %) die Umfrage nicht beendet. Die Abbrüche konzentrieren sich zu einem sehr großen Teil auf die Begrüßungsseite oder den Rahmenfragebogen vor den Vignetten; direkt im Vignettenteil haben lediglich 23 Befragte (4,1 % der Gesamt-Teilnehmerschaft) abgebrochen, im anschließenden Befragungsteil sind es weitere 19 Personen (3,4 %). Eine Differenzierung der Abbrüche nach experimentellen Splits erscheint angesichts dieser geringen Fallzahlen kaum sinnvoll. Festhalten lässt sich jedenfalls, dass selbst die umfangreiche Bewertungsaufgabe bei zwölf Dimensionen (der immerhin ca. die Hälfte der Befragten ausgesetzt war) und das Auftreten ungewöhnlicher Fälle (wie Anwälten ohne Hochschulabschluss) nicht zu auffallend hohen Abbruchquoten führen. Dies gilt ähnlich für Antwortverweigerungen: Lediglich 68 Vignetten, damit 1,9 % blieben unbeantwortet.[18]

Die vorangegangenen Ausführungen haben jedoch bereits gezeigt, dass sich eine mangelnde Kooperationsbereitschaft oder Überforderung ebenso in einem veränderten Antwortverhalten bei fortgesetzter Befragung niederschlagen kann – speziell dessen Verkennung wäre für die Ergebnisinterpretationen kritisch.[19] Einen ersten Hinweis auf mögliche Response-Sets liefern die Verteilungen der Vignettenurteile, wie sie in Tabelle 2 für die unterschiedlichen experimentellen Splits aufgeschlüsselt sind. Über alle Befragten hinweg (mittlere Spalte) als auch pro Befragten berechnet (letzte Spalte), wird eine etwas geringere Streuung (Standardabweichung) der Vignettenurteile, damit stärkere Konstanz des Antwortverhaltens bei zwölf gegenüber fünf Dimensionen offensichtlich. Allerdings verfehlt dieser Unterschied das Signifikanzniveau von fünf Prozent.[20]

18 Diese Quote an Missings entspricht etwa der von ‚herkömmlichen' Itemabfragen in der gleichen Erhebung. Die Befragten wurden direkt im Anschluss an die Vignetten gebeten, die Bedeutung der Vignettendimensionen für eine gerechte Entlohnung jeweils einzeln auf siebenstufigen Itemskalen einzustufen (von sollte ‚überhaupt keine Bedeutung' bis sollte ‚sehr große Bedeutung' spielen). Die Missings bewegen sich bei diesen Items zwischen 0,9 und 2,2 %, im Mittel sind es 1,2 % (vorherige Befragungsabbrüche nicht mitgezählt).

19 „If tasks are too long or too difficult or lack sufficient realism and credibility, data quality will suffer in the sense of not containing the information sought. Unfortunately, respondents generally answer the questions asked and seldom go out of their way to point out problems with tasks posed" (Carson et al. 1994: 355).

20 T-Test für die Mittelwertdifferenz der befragtenspezifischen Urteilsvarianz zwischen fünf und zwölf Dimensionen: $t = 1{,}48$; $p = 0{,}140$ bei zweiseitigem Test und Adaption für die verletzte Annahme der Varianzungleichheit (vorherige Prüfung mit Levene's Test).

Tabelle 2 Deskriptive Übersicht über die Vignettenurteile[a]

Experimentelle Variante	Anzahl	Mittelwert	S.D.	Mittlerer Mittelwert pro Befragten	Mittlere S.D. pro Befragten
5 Dimensionen, 7 Vignetten	1.213	5,21	3,10	5,20	2,94
5 Dimensionen, 10 Vignetten	574	5,44	3,21	5,45	3,12
12 Dimensionen, 7 Vignetten	1.109	5,51	2,96	5,51	2,87
12 Dimensionen, 10 Vignetten	584	5,36	2,98	5,35	2,86

[a] *Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.*

5.2 Multivariate Analysen

Erwartete Folgen einer zu hohen Komplexität sind ein inkonsistenteres Antwortverhalten (H_{1b}), statistisch eine geringere erklärte Varianz bzw. höhere Fehlervarianz, und ein Ausblenden einzelner Dimensionen (H_{1c}), was sich statistisch in geringeren Einflussstärken bzw. weniger signifikanten Effekten äußert. Zur Prüfung dieser beiden Annahmen dienen die in Tabelle 3 aufgeführten OLS-Regressionen, die wegen der hierarchischen Datenstruktur jeweils mit robusten Standardfehlern geschätzt sind. Um die methodischen Effekte besser von möglichen Drittvariableneffekten trennen zu können, werden die Regressionen für die ‚Zwölfer-Vignetten' ohne (Modell 2) und mit (Modell 3) Kontrolle der zusätzlichen Dimensionen präsentiert.[21]

Zunächst zur inhaltlichen ‚Lesart' der Ergebnisse: Bei der vorliegenden Kodierung der abhängigen Variablen bedeuten positive (negative) Koeffizientenwerte, dass das Einkommen als ungerechterweise zu hoch (niedrig) empfunden wird. Negative Effekte lassen sich somit als eine Erhöhung des als angemessen empfundenen Nettoeinkommens deuten. Nach allen drei Modellschätzungen wird beispielsweise Personen mit einem Berufsabschluss ein höheres Einkommen zugestanden als solchen ohne Abschluss. Für unser methodisches Forschungsinteresse ist aber interessanter, ob sich Unterschiede zwischen den Koeffizientenwerten der drei Modelle zeigen.

21 Prinzipiell sind die Vignettendimensionen bei fraktionalisierten Auswahlen unkorreliert. Sie geben also ihren reinen ‚Nettoeffekt' selbst dann wieder, wenn nicht auf Drittvariablen kontrolliert wird. Gerade hierin liegt ja eine wesentliche Stärke dieses Verfahrens. Einschränkung erfährt dies allerdings mit dem gezielten Ausschluss von Kombinationen, der unweigerlich zu Korrelationen führt. Dies betrifft auch das vorliegende Sample, von dem die logisch völlig unmöglichen Fälle ausgeschlossen wurden (wie z. B. Personen ohne Berufserfahrung, die schon lange in einem Betrieb arbeiten, vgl. Abschnitt 4). Eine Übersicht über die Korrelationen zwischen den einzelnen Dimensionen findet sich in Tabelle A3 im Anhang.

Tabelle 3 OLS-Regressionen der Vignettenurteile[a] (robuste Standardfehler in Klammern; sign. Unterschiede der Koeffizienten zwischen Modell 1 und 2 hervorgehoben)[b]

	Modell 1 5 Dimensionen	Modell 2 12 Dimensionen	Modell 3 12 Dimensionen
Weibliche Vignettenperson	-0,057 (0,122)	-0,136 (0,115)	-0,105 (0,113)
Alter [Jahre]	-0,021*** (0,005)	-0,029*** (0,005)	-0,020*** (0,005)
Abschluss (Ref.: kein Abschluss)			
– Berufsabschluss	-0,654*** (0,133)	-0,472*** (0,131)	-0,429*** (0,129)
– Hochschulabschluss	-1,126*** (0,129)	-0,623*** (0,126)	-0,830*** (0,130)
Berufprestige [10 MPS-Score]	-0,157*** (0,011)	-0,097*** (0,012)	-0,106*** (0,012)
Nettoeinkommen [100,- Euro]	0,060*** (0,002)	0,055*** (0,002)	0,058*** (0,002)
Berufserfahrung [Prozent der potenziellen Erwerbszeit]			0,066 (0,048)
Schon seit langem im Betrieb beschäftigt (Ref.: erst seit kurzem)			-0,645 (0,131)***
Leistung (Ref.: unterdurchschnittlich)			
– durchschnittlich			-0,813*** (0,129)
– überdurchschnittlich			-0,788*** (0,138)
Anzahl Mitarbeiter [100]			0,028*** (0,006)
Betriebssituation (Ref.: vom Konkurs bedroht)			
– ausgeglichene Bilanz			-0,037 (0,130)
– hohe Gewinne			-0,292** (0,122)
Zu 30% schwerbehindert (Ref.: gesund)			0,049 (0,114)
Anzahl Kinder			-0,152*** (0,029)
Konstante	6,465*** (0,280)	6,274*** (0,236)	6,820*** (0,272)
Beobachtungen:			
– Vignetten	1.787	1.693	1.693
– Befragte	235	225	225
R^2	0,47	0,45	0,49

[a] Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.
[b] Prüfung mittels Interaktionstermen zwischen den Vignettendimensionen und der Dimensionszahl in einem gepoolten Modell, Signifikanzniveau von fünf Prozent.
**** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Dies ist im Hinblick auf die Vorzeichen nicht der Fall, jedoch zeigen die Vignettenmerkmale bei den komplexeren zwölfdimensionalen Varianten oftmals einen betragsmäßig schwächeren Einfluss. Ein Chow-Test bestätigt signifikante Differenzen zwischen den Modellen 1 und 2 ($F = 4{,}04$ bei $df = 7$ und 459; $p = 0{,}000$).[22] Einzeln geprüft erweisen sich die Einflussstärken des Hochschulabschlusses und des Prestiges als signifikant verschieden. Da gerade die Einflüsse dieser beiden Variablen bei Kontrolle für die weiteren Dimensionen stabil bleiben (die Koeffizienten unterschieden sich nur marginal zwischen Modell 2 und 3), ist dieser Unterschied nicht durch Drittvariableneffekte bedingt, sondern er deutet vielmehr darauf hin, dass mit höherer Komplexität tatsächlich Dimensionen tendenziell ausgeblendet werden.[23] Die Anteile erklärter Varianz (R^2-Werte), welche als Maß für die Konsistenz des Antwortverhaltens herangezogen werden können, unterscheiden sich dagegen nicht substantiell zwischen den Modellen.[24]

Insgesamt wird unsere erste Hypothese (H_{1a}, H_{1b}, H_{1c}) damit nur in dem Teilaspekt H_{1c} bestätigt.[25] Die Anzeichen für eine kognitive Überforderung sind – trotz der hohen Dimensionszahl – gering. Solange nicht Wege gefunden werden, die Komplexität von Vignetten zu kontrollieren, sollten absolute Effektstärken dennoch vorsichtig interpretiert werden (für sich genommen und beim Vergleich von Studien). An dieser Stelle ist auf eine weitere, bei anderen Autoren zu findende, problematische Interpretation zu verweisen: Oftmals werden ‚hohe' R^2-Werte als Beleg dafür gewertet, dass es gelungen sei, alle für die Befragten relevanten Merkmale in die Vignetten aufzunehmen (es bleibt kaum mehr etwas unerklärt, somit

22 Technisch besteht dieser Test darin, ein gepooltes Modell zu schätzen, in das zusätzlich eine Dummyvariable für die zu prüfende Designvariante (hier die Anzahl der Dimensionen) sowie Interaktionsterme aller Vignettendimensionen mit dieser Designvariante aufgenommen werden. Geprüft wird dann, ob die Aufnahme dieser Variablen *insgesamt* zu einer signifikanten Modellverbesserung führt; im vorliegenden Falle einer OLS-Regression, ob es zu einem signifikanten Anstieg der erklärten Varianz kommt (für Details: Wooldridge 2003: 238f.).

23 Der Vergleich zwischen fünf- und zwölfdimensionalen Vignetten ist statistisch nicht trivial. Mit einer höheren Variablenzahl steigt automatisch die Wahrscheinlichkeit von Korrelationen der Variablen untereinander oder von Konfundierungen mit Wechselwirkungen. Aufgrund der hohen Anzahl an möglichen Wechselwirkungen (bei der Variante mit zwölf zum Teil kategorialen Dimensionen liegen allein mehr als 70 mögliche Interaktionen erster Ordnung vor) sind diese nicht alle modellierbar (mitunter wird dies bereits durch die Stichprobenbildung verhindert). Damit ist nicht gänzlich auszuschließen, dass die Effekte im Falle der höher-dimensionalen Vignetten leicht verzerrt geschätzt werden (‚omitted-variable-bias'; wir danken P. Steiner für diesen wertvollen Hinweis). Es sollten daher künftig nochmals Replikationen mit anderen Vignettenstichproben durchgeführt werden. Darauf wird in der Schlussbetrachtung (Abschnitt 6) zurückgekommen.

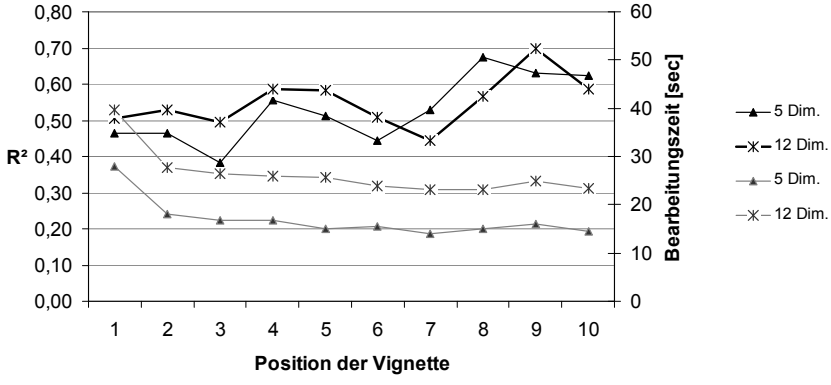
24 Der Vergleich von R^2-Werten zwischen Modellen ist nicht unproblematisch (Wooldridge 2003). Im vorliegenden Fall scheinen die Voraussetzungen jedoch erfüllt: Die Fallzahlen sind vergleichbar und ebenso bestehen nur minimale Unterschiede in der Varianz der abhängigen Variablen.

25 Wobei sich die These H_{1a} (häufigere Befragungsabbrüche bei höher-dimensionalen Vignetten) aufgrund der geringen Abbruchquoten nicht *statistisch* prüfen lässt.

seien alle urteilsrelevanten Informationen berücksichtigt; z. B. Beck/Opp 2001: 302). Wie unsere Ergebnisse zeigen, kann dies ein Trugschluss sein, denn die hinzukommenden Merkmale in Modell 3 erweisen sich fast ausnahmslos als signifikant, ohne dass es zu einem bedeutenden Anstieg der Varianzaufklärung käme. Eine hohe Modellanpassung ist somit zwar ein Maß für ein in sich konsistentes Urteilsverhalten, damit aber noch nicht unbedingt ein Indikator dafür, dass *alle* inhaltlich relevanten Dimensionen berücksichtigt sind.[26]

Zu beachten ist ferner, dass unseren Befragten mit maximal zehn Vignetten vergleichsweise wenige Urteile abverlangt wurden. Möglicherweise fallen kognitive Überforderungen und Ermüdungen erst bei weitaus höheren Vignettenzahlen ins Gewicht, oder schwächen sich umgekehrt mit zunehmender Übung ab. Damit sind die Hypothesen 2a und 2b angesprochen, die eine mit der Beantwortungssequenz zunehmende Konsistenz des Antwortverhaltens postulieren, speziell bei den komplexeren ‚Zwölfer'-Vignetten. Zur Überprüfung stellen wir Regressionsschätzungen getrennt für die einzelnen Bearbeitungspositionen der Vignetten an.

Abbildung 2 R^2-Werte (dicker gedruckte, obere Linien) und Bearbeitungszeiten pro Vignette (schwächere, untere Linien) in Abhängigkeit von der Position der Vignette und Anzahl ihrer Dimensionen



26 Womit auch Aussagen wie die folgende eine Relativierung finden: „The factorial survey method makes it possible to assess the number and identity of the characteristics a person uses in reaching a judgement." (Jasso 2006: 342)

In Abbildung 2 sind die resultierenden R^2-Werte für die beiden Designvarianten (fünf- vs. zwölfdimensional) gegen die Positionen der Vignetten abgetragen (dunklere, obere Linien). Da diese ebenfalls Aufschluss über Lern- bzw. Ermüdungseffekte geben, sind zugleich die mittleren Bearbeitungszeiten pro Vignette[27] (untere bzw. hellere Linien) dargestellt.

Was die Varianzaufklärung bzw. R^2-Werte betrifft, ist im Bearbeitungsverlauf ein leichter Anstieg zu erkennen. Die durchschnittliche Bearbeitungszeit pro Vignette sinkt dagegen insbesondere nach der ersten Vignette sprunghaft und mit abnehmender Rate weiter bis zur siebten Vignette. Zusammen genommen deutet dies auf einen Lerneffekt hin: Die Befragten können die Vignetten in zunehmend kürzerer Zeit beantworten, ohne dass es zu Einbußen ihrer Antwortkonsistenz käme. Entgegen unserer Erwartungen (H_{2b}) gilt dies nicht verstärkt für die komplexeren Vignetten: Die Linien für die fünf- und zwölfdimensionalen Vignetten verlaufen jeweils parallel zueinander, was bedeutet, dass die Lerneffekte für beide Versionen etwa gleich stark ausfallen. Für die vermutete Wechselwirkung zwischen Komplexitäts- und Lerneffekten findet sich also kein Beleg. Um die Interpretation als einen Lerneffekt abzusichern, ist zusätzlich noch zu prüfen, ob die steigende (oder zumindest gleich bleibende) Konsistenz nicht einer verstärkten Ausblendung von Dimensionen, also einer vereinfachten Entscheidungsheuristik, geschuldet ist. Um dies auszuschließen, wurden separate Regressionen mit dem ersten, zweiten und letzten Drittel der Vignetten berechnet. Die hier aus Platzgründen nicht dargestellten Modellschätzungen unterscheiden sich nicht signifikant voneinander,[28] d. h. die Anzahl einflussreicher Dimensionen, ihre Effektstärken und allgemein das Antwortmuster bleiben in der Bearbeitungssequenz stabil. Trotz der hohen Komplexität von zwölf Dimensionen führen also bereits die ersten Vignettenurteile zu sehr reliablen Urteilen – was bedeutet, dass sie nicht als ‚Übungsfälle' betrachtet

27 Die verwendete Online-Programmierung erlaubt es, die Bearbeitungszeit pro Vignette auf die Sekunde genau zu messen; exakter handelt es sich um die Zeit, die zwischen dem Abschicken der jeweiligen Vignettenseite und der Beendigung der vorherigen Seite verstrichen ist. Für derartige Zeitmessungen ist die bei Online-Befragungen geringe Kontrolle über das Setting nachteilig: Pausen der Befragten werden unweigerlich mit zur Bearbeitungszeit gerechnet. Aus diesem Grunde wurde jeweils das obere Fünf-Prozent-Perzentil der Antwortzeiten aus den Berechnungen ausgeschlossen (zur grundsätzlichen Empfehlung einer Bereinigung um ‚outliers' bei Befragungszeiten: Urban/Mayerl 2007; Mayerl/Selke/Urban 2005).

28 Entsprechende Chow-Tests fallen nicht signifikant aus. Einzeln betrachtet nehmen die Dimensionen mit den Vignettenpositionen in ihren Effektstärken tendenziell zu (wiederum Vergleich des ersten mit den beiden anderen Dritteln an Vignetten), kommt es also zu einer immer stärkeren Beachtung der Dimensionen, was den Lerneffekt eher noch untermauert. Allerdings wird die Signifikanzschwelle von fünf Prozent keinesfalls erreicht. Ebenfalls finden sich keine signifikanten Modellunterschiede, wenn die Berechnungen getrennt für die beiden Splits mit fünf und zwölf Dimensionen wiederholt werden.

werden müssen, die aus den Ergebnisanalysen ausgeschlossen werden sollten (zu einer entsprechenden Empfehlung bei Choice-Analysen: Caussade et al. 2005: 632, Anm. 6). An dieser Stelle ist aber darauf hinzuweisen, dass sich diese Aussagen nicht über die hier vorliegende, geringe Vignettenzahl und das sehr alters- und bildungshomogene Befragtensample hinaus verallgemeinern lassen. Ob bei höheren Vignettenanzahlen und/oder anderen Befragtengruppen nicht doch Ermüdungserscheinungen durchschlagen, bleibt künftigen Untersuchungen vorbehalten.

Die Wirkung unplausibler Fälle kann ähnlich untersucht werden. Dafür ist zunächst eine Festlegung erforderlich, was überhaupt als ‚unplausibel' zu gelten hat (vgl. Abschnitt 3.3) – aufgrund der hier bestehenden fließenden Übergänge und sicherlich auch subjektiv differierenden Einschätzungen keine triviale Aufgabe. Um möglichst objektiv vorzugehen, ziehen wir die Rückmeldungen aus ca. 60 mündlichen Pretestinterviews heran, die mit demselben Vignettensample im Herbst 2007 im Rahmen eines Forschungs-Projektseminars an der Universität Konstanz durchgeführt wurden. Für diese Interviews wurden bewusst sehr heterogene Personen der Allgemeinbevölkerung ausgewählt. Besonders häufig monierten die Befragten die Unsinnigkeit von Kombinationen der beiden Dimensionen ‚Beruf' und ‚Ausbildungsabschluss': Speziell ‚Anwälte ohne Ausbildung und Hochschulabschluss' sorgten oftmals geradezu für Verärgerung.[29] Wir werten entsprechend alle Vignetten als unplausibel, bei denen die geschilderten Personen nicht über einen Ausbildungs- oder Hochschulabschluss verfügen, der für ihren Beruf in Deutschland eigentlich Voraussetzung oder zumindest sehr üblich wäre.[30] Dies trifft auf insgesamt 22,5 % (N=800) unserer Vignetten zu.

Wir ziehen zudem eine alternative Operationalisierung auf Basis der Dimensionen ‚Einkommen' und ‚Beruf' heran. Speziell für Vignetten mit einem *berufsspezifisch* ungewöhnlichen Einkommen finden sich ebenfalls mehrere Pretestkommentare, welche ihre Ernsthaftigkeit in Zweifel ziehen (z. B. sorgten Vollzeit arbeitende, leitende Manager mit einem monatlichen Gehalt von nur 250 Euro netto für starke Irritationen, oder Friseure mit 15.000 Euro netto). Überdies besteht für diese beiden Dimensionen wiederum die Möglichkeit einer weitgehend objektiven Definiti-

29 Es fielen Äußerungen wie „man fühle sich auf die Palme gebracht"; „wer denkt sich so einen Unsinn aus". Befragt wurden Personen unterschiedlichen Alters und Bildungsgrades, die Rekrutierung und Durchführung der Interviews geschah durch die Seminarteilnehmer – ihnen sei an dieser Stelle unser Dank ausgesprochen.

30 Konkret sind dies: Anwälte ohne Hochschulabschluss oder nur mit Berufsabschluss; Verwaltungsfachkräfte, Elektroingenieure, Sozialarbeiter, Lokführer und leitende Manager ohne Ausbildung. Bei den übrigen (allesamt auch nicht gesetzlich geschützten) Berufsangaben (Friseure, Pförtner, Programmierer und ungelernte Arbeiter) scheinen dagegen Tätigkeiten ohne Ausbildungsabschluss plausibler.

on. Unplausible Vignetten lassen sich durch einen Abgleich der Vignetteneinkommen mit den realen Nettoeinkommen der jeweiligen Berufsgruppen identifizieren. Dazu bestimmen wir zunächst anhand der Stichprobe des Sozio-ökonomischen Panels (SOEP) von 2007 die mittleren tatsächlichen Nettoeinkommen der zehn in den Vignetten verwendeten Berufsgruppen und berechnen dann für jede Vignette die absolute Differenz zwischen ihrem ‚virtuellen' Vignetteneinkommen und dem tatsächlichen mittleren Berufseinkommen nach dem SOEP. Atypisch sind dann Fälle mit einer betragsmäßig besonders großen Differenz nach oben oder unten; konkret werten wir alle Vignetten mit einer absoluten Abweichung von mindestens 3.000 Euro als unplausibel (das sind 23,9 % aller Vignetten).[31]

Zu wählen ist noch ein geeignetes Analyseverfahren. Der zunächst nahe liegende Abgleich von Regressionsschätzungen auf Basis von plausiblen versus unplausiblen Fällen wird durch die unterschiedliche Varianz der unabhängigen Variablen in diesen beiden Gruppen beeinträchtigt: Bei den unplausiblen Fällen liegt *per se* eine andere Korrelation und Varianz der sie definierenden Dimensionen vor. Zudem würden die deutlichen Diskrepanzen in den Fallzahlen den Vergleich erschweren (die unplausiblen Fälle machen jeweils nur eine Minderheit aus). Theoretisch zu erwarten ist aber ohnehin ein Effekt, der sich nicht nur auf die unplausiblen Vignetten selbst bezieht, sondern ebenso auf die nachfolgenden: Wir gehen schließlich davon aus, dass die Konfrontation mit unrealistischen Fällen den *grundsätzlichen* Glauben an die Ernsthaftigkeit der Befragung und damit die *generelle* Kooperationsbereitschaft schmälert – d. h. genau genommen rechnen wir *ab* dem Auftreten unplausibler Vignetten mit einem weniger konsistenten (H_{3b}) oder stärker vereinfachten (d. h. weniger, bzw. allein die unplausiblen Dimensionen einbeziehenden) Urteilsverhalten (H_{3c}).

Dies legt es nahe, die Antwortmuster *bis* und *ab* dem Auftreten eines ersten unplausiblen Falls miteinander zu vergleichen. Wann das erste Mal eine unplausible Vignette erscheint, und wie viele unplausible Vignetten es pro Befragten sind, variiert aufgrund der zufälligen Deckzusammenstellung zwischen den Befragten. Insbesondere die zufällige Reihenfolgeposition pro Befragten erlaubt es, die Wirkung der

31 Neben dieser Definition über das Quartil haben wir Kontrollrechnungen mit dem Quintil und Dezentil durchgeführt – also mit einer noch stärkeren Eingrenzung unplausibler Fälle. Diese bestätigen unsere Befunde vollends und werden daher nicht separat dargestellt. Zur Berechnung der Netto-Berufseinkommen wurde die generierte Einkommensvariable des SOEP 2007 verwendet (für nähere Informationen zum SOEP: Wagner/Frick/Schupp 2007).

unplausiblen Vignetten von Lerneffekten zu trennen.[32] Tabelle 4 beinhaltet Regressionsschätzungen einerseits für die Definition der Unplausibilität über die Ausbildung (Modelle 1 und 2), andererseits für ihre Operationalisierung über das Einkommen (Modelle 3 und 4). Es werden jeweils Schätzungen für die Urteile vor (Modelle 1 und 3) sowie ab dem Auftreten eines ersten ‚unplausiblen' Falles (Modelle 2 und 4) gegenübergestellt. Ein Chow-Test weist die Modellunterschiede bei beiden Operationalisierungen als signifikant aus ($F = 2{,}06$; $p = 0{,}046$ bei $df = 7$ und 459; bzw. $F = 71{,}37$; $p = 0{,}000$ bei $df = 7$ und 459). Bei der Ausbildungs-Operationalisierung geht dies neben einem Niveaueffekt (die Konstante ist in Modell 2 geringer als in Modell 1) auf die Ausbildungsdimension zurück. Der Einfluss des Berufsabschlusses unterscheidet sich zu einem Zehn-, der Hochschulabschluss zu einem Fünf-Prozent-Niveau signifikant (Prüfung durch Schätzung eines gemeinsamen Modells mit entsprechenden Interaktionstermen).[33] Dagegen erweist sich bei der Operationalisierung über das Einkommen (Modelle 3 und 4) allein der Einfluss dieser Dimension als signifikant verschieden. Eine Analogie besteht auch in der Richtung der Unterschiede: Die Effekte fallen jeweils für die Modelle ab dem Auftreten eines unplausiblen Falles (also in Modell 2 statt 1 und 4 statt 3) schwächer aus. Dies entspricht nicht

32 Hierzu wird in den Modellen zudem für die Bearbeitungsposition kontrolliert. Bei beiden Operationalisierungen treten unplausible Vignetten zu etwa 60 % auf einer der ersten drei Rangpositionen das erste Mal auf, und es haben jeweils ca. ein Drittel der Befragten eine bis maximal zwei unplausible Vignetten erhalten. Maximal sind es fünf unplausible Vignetten pro einzelnem Befragten. Das Bestehen eines Lerneffektes wird überdies durch die entsprechenden Analysen im vorherigen Abschnitt entkräftet, schließlich haben sich die dort geprüften Unterschiede der Koeffizienten nach Bearbeitungsposition der Vignetten nicht als signifikant erwiesen. Um Fehlschlüsse zu vermeiden, sind dennoch ein paar weitere Überlegungen zur Vergleichbarkeit der beiden Gruppen erforderlich. Das Problem der unterschiedlichen Varianz der Dimensionen ist noch nicht völlig ausgeräumt – die Antworten ab dem Auftreten unplausibler Fälle beziehen sich unweigerlich auf stärker variierende und geringer korrelierte Vignettendimensionen, schließlich sind bei diesen Vignetten (im Gegensatz zur Vergleichsgruppe ohne unplausible Fälle) *alle* Merkmalskombinationen zulässig und tritt somit beispielsweise das berufsspezifische Einkommen in einer höheren Bandbreite auf. Diese stärkere Varianz und Unkorreliertheit bewirken aber *per se* eine höhere Schätzpräzision, damit ‚Power' von Signifikanztests. Finden sich größere Einflussstärken unplausibler Dimensionen, stellen diese daher noch nicht unbedingt ein Beleg für die von Faia (1980) vermutete Fokussierung der Befragten auf diese Merkmale dar, sie können ebenso allein mathematisch-statistisch bedingt sein. Und selbst wenn sich eine höhere kognitive Aufmerksamkeit der Befragten feststellen ließe, wäre diese dann noch nicht notwendigerweise der ‚Irrealität' der Fälle geschuldet, sie könnte ebenso Anzeichen eines ‚number-of-levels'- oder ‚range'-Effektes sein. Hierunter werden höhere kognitionspsychologische Aufmerksamkeiten der Befragten für stärker (und in größeren Spannweiten) variierende Merkmale gefasst, unabhängig von deren Inhalten. Diese Effekte, die zumindest für die verwandten Conjoint-Verfahren bereits gut belegt sind (z. B. Ohler et al. 2000; Louviere 2001a; Wittink/Krishnamurthi/Nutter 1982; Wittink/Krishnamurthi/Reibstein 1989; Perrey 1996), bieten also alternative Interpretationen für die Wirkung unplausibler Fälle. Auf diese Vermischung weisen ähnlich bereits Creyer/Ross 1988 hin, mitunter durch empirische Befunde gestützt (dazu auch Klein 2002: 15).

33 Es wurden Interaktionsterme zwischen allen Vignettendimensionen und einer Dummy-Variable gebildet, welche die Vignette einordnet (vor vs. ab unplausiblem Fall).

unserer Annahme, lässt sich aber gleichwohl als ein schlüssiger Befund deuten: Die Befragten beziehen ab der Konfrontation mit einer unplausiblen Kombination die ursächlichen Dimensionen weniger in ihr Urteil ein – man könnte auch sagen, sie nehmen diese weniger ‚ernst'.[34] Sollte sich dieser überraschende Befund in künftigen Untersuchungen (die aufgrund des zugegebenermaßen ‚ad hoc'-Charakters unserer Interpretation angeraten scheinen) replizieren, kann die Empfehlung nur lauten, sparsam mit solchen Fällen umzugehen; oder zumindest wie hier mit Zufallsreihenfolgen der Vignetten zu arbeiten, um Konfundierungen der Auswirkungen unplausibler Fälle (die eben erst ab ihrem Erscheinen auf den hinteren Bearbeitungspositionen auftreten können) mit inhaltlichen Effekten zu vermeiden.

Unsere Erwartung hinsichtlich der Antwortkonsistenz werden ebenfalls kaum erfüllt – mit dem Auftreten unplausibler Vignetten kommt es lediglich bei der Definition über das Einkommen zu einer leichten Abnahme der Konsistenz, gemessen am R^2. Bei der Operationalisierung über die Ausbildung ist dagegen sogar ein leichter Anstieg dieses Wertes zu verzeichnen. Die Konsistenz der Urteile wird hier durch die Abstraktion von den unrealistischen Dimensionen also sogar erhöht, was nochmals auf die mit Vorsicht zu behandelnde Interpretation des Wertes als einen Indikator für ‚erschöpfende Urteilsregeln' verweist (zu entsprechenden Interpretationen Beck/Opp 2001: 302; Jasso 2006: 416).

Bevor wir zu einem Fazit kommen, wollen wir die Einflüsse auf die Antwortkonsistenz und die Bearbeitungszeit noch in zwei abschließenden, multivariaten Modellen zusammenfassen. Als Maß für die Konsistenz verwenden wir die unerklärt gebliebene Varianz, genauer gesagt die quadrierten Residuen und schätzen OLS-Regressionen mit den Designvariablen als unabhängigen Variablen. Ein negativer Effekt bedeutet dann eine geringere Fehlervarianz bzw. höhere Konsistenz des Antwortverhaltens. Wie Tabelle 5 zeigt, finden sich lediglich zwei derartige (und nur zum Zehn-Prozent-Niveau) signifikante Effekte (Modell 1): Ab dem Auftreten von unplausiblen Fällen verringert sich die Fehlervarianz bzw. erhöht sich die Antwortkonsistenz (was wie oben gezeigt einem weniger Dimensionen einbeziehenden, damit vereinfachten Antwortverhalten geschuldet ist), und ebenso ist die Antwortkonsistenz umso höher, je mehr Zeit sich die Befragten für das Beantworten der einzelnen Vignetten nehmen.[35]

34 Der Annahme von Faia (1980) ist dies ebenso diametral entgegengesetzt wie einem ‚number-of-levels-' oder ‚range'-Effekt.

35 Um von grundsätzlichen Unterschieden in der individuellen ‚Basisgeschwindigkeit' (der vom Frageinhalt unabhängigen Grundgeschwindigkeit der Befragten) zu abstrahieren, wurden die Bearbeitungszeiten der Vignetten bei den hier vorliegenden Analysen jeweils pro Befragten mit seiner Antwortzeit bei ‚herkömmlichen' Itembatterien gewichtet (sog. ‚Latenzzeiten'). Die Geschwindigkeiten wurden dabei vorab um ‚Ausreißer' (oberes Fünf-Prozent-Perzentil) bereinigt. Zur Empfehlung eines ähnlichen Vorgehens siehe Mayerl/Selke/Urban 2005; Urban/Mayerl 2007.

Tabelle 4 OLS-Regressionen der Vignettenurteile[a] in Abhängigkeit des Auftretens unplausibler Fälle (robuste Standardfehler in Klammern; sign. Unterschiede in den Koeffizienten zwischen Modell 1 und 2 bzw. 3 und 4 hervorgehoben)[b]

	Definition über Ausbildung		Definition über Einkommen	
	Modell 1 Vor unpl. Vignetten	Modell 2 Ab unpl. Vignetten	Modell 3 Vor unpl. Vignetten	Modell 4 Ab unpl. Vignetten
Weibliche Vignettenperson	-0,244*	-0,023	-0,178	-0,072
	(0,130)	(0,104)	(0,117)	(0,100)
Alter [Jahre]	-0,030***	-0,025 ***	-0,030 ***	-0,022 ***
	(0,006)	(0,004)	(0,005)	(0,004)
Abschluss (Ref.: kein Abschluss)				
– Berufsabschluss	-0,972***	-0,571 ***	-0,671 ***	-0,670 ***
	(0,202)	(0,105)	(0,145)	(0,107)
– Hochschulabschluss	-1,386***	-0,784 ***	-1,101 ***	-0,962 ***
	(0,182)	(0,117)	(0,134)	(0,109)
Berufsprestige [10 MPS-Score]	-0,135 ***	-0,114 ***	-0,126 ***	-0,149 ***
	(0,015)	(0,010)	(0,013)	(0,010)
Nettoeinkommen [100,- Euro]	0,058 ***	0,057 ***	0,161 ***	0,055 ***
	(0,002)	(0,002)	(0,005)	(0,001)
Position der Vignette[c]	0,004	0,063 ***	-0,029	0,071 ***
	(0,032)	(0,021)	(0,029)	(0,021)
Konstante	7,152 ***	5,813 ***	5,021 ***	6,124 ***
	(0,353)	(0,241)	(0,277)	(0,252)
Beobachtungen:				
– Vignetten	1.301	2.179	1.197	2.283
– Befragte	355	400	344	409
R^2	0,44	0,48	0,56	0,52

[a] Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.

[b] Prüfung mittels Interaktionstermen zwischen den Vignettendimensionen und der Dimensionszahl in einem gepoolten Modell, Signifikanzniveau von fünf Prozent.

[c] Hier als lineare Variable ausgewiesen, da ausschließlich Kontrollfunktion. Bei einer alternativen Modellierung als Dummy-Variablen bleiben die Ergebnisse stabil.

**** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Tabelle 5 OLS-Regressionen der Quadrierten Residuen[a] und Bearbeitungszeiten[b] pro Vignette (Robuste Standardfehler in Klammern)

	Modell 1 Quadrierte Residuen[a]	Modell 2 Bearbeitungszeit pro Vignette[b]
Position Vignette	-0,106	-0,064***
	(0,197)	(0,004)
Position Vignette, quadriert	0,006	0,005***
	(0,018)	(0,000)
Zwölf Dimensionen (Ref.: fünf)	-0,123	0,119***
	(0,300)	(0,008)
Ab unplausiblem Fall[c]	-0,564*	-0,004
	(0,306)	(0,008)
Bearbeitungszeit pro Vignette[a]	-1,461*	
	(0,865)	
Konstante	6,102***	0,371***
	(0,554)	(0,010)
Beobachtungen:		
– Vignetten	3.095	3.095
– Befragte	416	416
R^2	0,00	0,26

[a] *Residuen einer OLS-Regression des Vignettenurteils auf die ersten fünf Vignettendimensionen.*
[b] *Bearbeitungszeit in Sekunden, pro Befragten mit der Bearbeitungszeit einer entsprechenden Itembatterie gewichtet. Bei beiden Bearbeitungszeiten wurde das obere Fünf-Prozent-Perzentil ausgeschlossen. Hieraus resultieren die geringeren Fallzahlen in diesen Modellen.*
[c] *Hier definiert über den Ausbildungsabschluss.*
*** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Die in Modell 2 betrachtete Bearbeitungszeit sinkt dagegen mit der Position der Vignetten, d. h. den Befragten gelingen zunehmend zeiteffiziente Urteile. Wie an dem positiven Effekt der quadrierten Bearbeitungszeit ersichtlich ist, handelt es sich um einen Effekt mit abnehmender Rate (grafisch einen ‚u-förmigen' Effekt).[36] Zudem bestätigt sich nochmals die ‚zeitraubendere' Bearbeitung von Vignetten mit einer höheren Dimensionszahl. Unplausible Fälle führen zumindest bei der hier ver-

36 Der Wendepunkt wäre nach der vorliegenden Modellschätzung bei der zehnten Vignette erreicht. Da für darüber hinausgehende Vignetten keine Beobachtungen vorliegen, ist dieser Befund aber nochmals mit umfangreicheren Vignettendecks pro Befragten zu validieren.

wendeten Operationalisierung über die Ausbildung nicht zu einem *zeitlich* flüchtigeren Antwortverhalten.[37]

6 Zusammenfassung und Schlussfolgerungen

Der Faktorielle Survey hat sich in der soziologischen Norm- und Einstellungsforschung inzwischen als Erhebungsmethode sehr gut durchgesetzt. Aber auch in anderen soziologischen Forschungszusammenhängen (wie z. B. der Diskriminierungsforschung) wird das Verfahren in den letzten Jahren vermehrt als eine ideale und innovative Methodik entdeckt. In Diskrepanz zu dieser guten Etablierung steht die geringe Erforschung des Verfahrens selbst. Es fehlen anwendungsbezogene Kriterien für die Konzeption Faktorieller Surveys, was ihre Durchführung erschwert (Beck/Opp 2001: 283f.). Ferner bestehen substantielle Zweifel an ihrer (internen) Validität fort. Invalide Urteile könnten etwa aus einer kognitiven Überforderung und/oder der Anwendung vereinfachter Entscheidungsstrategien (Heuristiken) resultieren. Ohne gezielte Methodenstudien ist kaum zu entscheiden, inwieweit die mit Vignetten gewonnenen Ergebnisse belastbar oder als methodische Artefakte zu interpretieren sind.

Der vorliegende Beitrag zielt daher auf eine erste Untersuchung der Stabilität und Konsistenz des Antwortverhaltens in Abhängigkeit von Designmerkmalen, konkret der Komplexität, Reihenfolge und Plausibilität von Vignetten. Mittels einer experimentellen Onlinebefragung von 460 Studierenden wurden systematische Analysen zum Einfluss der Anzahl der Dimensionen, Lerneffekten und der Wirkung von unplausiblen Fällen vorgenommen. Als ein erster übergreifender Befund lässt sich festhalten, dass diese drei methodischen Aspekte für die Antwortmuster und damit die Interpretationen der inhaltlichen Ergebnisse durchaus relevant sind.

So zeigen unsere Analysen, dass die Komplexität von Vignetten, jedenfalls gemessen an ihrer Dimensionszahl, die Urteile signifikant beeinflusst. Die Effekte einzelner Merkmale schwächen sich mit der Anzahl der Dimensionen ab – was zunächst bedeutet, dass die Interpretationen der absoluten Effektstärken nicht überstrapaziert werden sollten. Der Einfluss der einzelnen Merkmale scheint mitunter eine Funktion ihrer ‚Einzelständigkeit' zu sein, was zumindest beim Vergleich von

37 Bei einer Operationalisierung über das Einkommen findet sich allerdings eine zum Fünf-Prozent-Niveau signifikante Verringerung der Beantwortungszeit (um durchschnittlich 0,016 Sekunden). Diese differenten Befunde je nach Operationalisierung unplausibler Fälle fordern zu weiteren Untersuchungen heraus.

unterschiedlichen Studien zu berücksichtigen ist. Anders gesagt: aussagekräftige Vergleiche von Effekten in verschiedenen Erhebungen bedürfen ähnlich komplexer Vignetten. Die bei höherer Komplexität zu beobachtenden Heuristiken führen zumindest dann zu Artefakten, wenn sie keine Entsprechung mehr zu realen Urteilen aufweisen (ähnlich Swait/Adamowicz 2001: 147). Nicht-signifikante Einflüsse sind möglicherweise nochmals mit weniger-dimensionalen Fallbeispielen zu validieren. Die gute Botschaft lautet aber, dass die Auswirkungen auf das Antwortverhalten insgesamt gering sind und die Befragten selbst die hier vorgelegte, hohe Komplexität von zwölf Dimensionen insgesamt noch gut zu bewerkstelligen scheinen. Aufgrund der zu vermutenden Wechselwirkungen mit anderen Designvariablen (wie der Anzahl an Ausprägungen) und den Merkmalen der Befragten (kognitive Leistungsfähigkeit) sind allerdings vertiefende Untersuchungen angebracht.

Methodenstudien zu den verwandten Conjoint- und Choice-Analysen sprechen für ein komplexes Verhältnis von Lern- und Ermüdungseffekten. Um dieses vollständig abzubilden, ist die hier verwendete Fallzahl von maximal zehn Vignetten pro Befragten zu gering. Ein interessanter Befund ist aber schon einmal, dass bis zu unserer letzten, zehnten Vignette Lerneffekte dominieren, welche sich primär in einer zunehmenden Antwortgeschwindigkeit bei gleich bleibender Konsistenz (R^2) äußern. Die mit den ersten Vignetten gewonnenen Urteile sind in unserer Stichprobe reliabel, sie werden inhaltlich also durch die nachfolgenden bestätigt. Dies spricht für die grundsätzliche Verwertbarkeit dieser ‚ungeübten' ersten Urteile und damit die Validität von Studien, die mit einem reinen ‚between-subject'-Design arbeiten (nur eine Vignette pro Befragten). Ab welcher Anzahl an Vignetten die quantitativen Zugewinne an Urteilen mit merklichen Einbußen ihrer Datenqualität bezahlt werden, ist dagegen erst mit umfangreicheren Vignettendecks zu klären. Zudem sollte geprüft werden, ob diese Befunde auch einer anderen Komplexität von Vignetten und einem heterogeneren Befragtensample Stand halten.[38]

Von den einen als Stärke des Verfahrens gelobt, sehen Kritiker gerade durch empirisch seltene, daher besonders ‚virtuelle' Vignetten artifizielle Urteile herbeigeführt. Eine Skepsis, die nach unseren Analysen durchaus angebracht ist. Unplausible Merkmalskombinationen scheinen zwar zu keinen drastischen Befragungsabbrüchen oder Antwortverweigerungen zu führen (die Quote an Abbrüchen und Non-Responses ist insgesamt sehr gering), aber sie provozieren eine geringere

38 Anzunehmen ist, dass die Komplexität für die Befragten auch mit dem inhaltlichen Thema variiert, genauer gesagt mit ihrer Vertrautheit mit dem zu beurteilenden Gegenstand. Dies wurde zumindest für Choice-Experimente vereinzelt bereits untersucht (in der gesundheitsökonomischen Panel-Studie von Bryan et al. 2000 ist allerdings *kein* Effekt der Erfahrung auf die Reliabilität der Antworten festzustellen).

Berücksichtigung (bis möglicherweise vollständige Ausblendung) der für die Unplausibilität ursächlichen Dimensionen. Dieser Befund ist bei inhaltlichen Ergebnisinterpretationen zu beachten: Fehlende oder geringe Signifikanzen können statt einer genuinen Irrelevanz für das Urteilsverhalten ebenso anzeigen, dass die Dimensionen allein *in Folge ihrer Irrealität* weniger ernst genommen werden. Sollte sich dieses Ergebnis in weiteren Untersuchungen bestätigen, kann die praktische Empfehlung nur lauten, auf unplausible Fälle zu verzichten, oder zumindest sparsam mit ihnen umzugehen. Dank computerbasierter Verfahren lassen sich die durch ihren Ausschluss hervorgerufenen Einbußen an Effizienz der Vignettenstichproben (Balanciertheit und Unkorreliertheit der Dimensionen) auf ein vertretbares Maß reduzieren.

Unsere Analysen haben zudem gezeigt, dass es zur Feststellung methodischer Effekte multipler Kriterien bedarf. Vereinfachte Entscheidungsregeln tragen tendenziell zu einer hohen Messgüte der Ergebnisse bei (bewertet an der Varianzaufklärung), die Abweichung von den tatsächlichen Einstellungen und Urteilsregeln der Befragten kann gleichwohl groß sein.[39] Die zentrale Schlussfolgerung ist hier, dass die R^2-Werte allein als ein Indikator für die Konsistenz der Urteile zu gebrauchen sind, nicht aber als ein Maß dafür, inwieweit es gelungen ist, alle urteilsrelevanten Dimensionen ausfindig zu machen. Die Befragten scheinen sich bei einer drohenden Überforderung eher auf ein weiterhin konsistentes, aber gerade darum weniger detailliertes Urteilsverhalten zu konzentrieren. Faktorielle Surveys sind demnach primär ein geeignetes Verfahren für die Feststellung der Signifikanz *einzelner* Merkmale (etwa für entsprechende Hypothesentests), und weniger für die Aufdeckung inhaltlich *erschöpfender* Urteilsregeln. Anders ausgedrückt lassen sich mit ihnen Aussagen über den Einfluss der berücksichtigten Dimensionen treffen, nicht aber über die zusätzliche (Ir-)Relevanz weiterer Merkmale. Dies verweist nochmals auf die hohe Bedeutung einer sorgfältigen Auswahl der Dimensionen.

Aufgrund der Vielzahl weiterer methodischer Problemlagen und der zu erwartenden Wechselwirkungen mit anderen Designmerkmalen (wie z. B. der Anzahl an Ausprägungen und deren Variation und Bandbreite) ist mit den hier vorgelegten Untersuchungen erst ein Anfang gemacht. Empfehlenswert erscheinen zunächst Replikationen mit anderen Vignettenstichproben, etwa mit einem fraktionalisierten Design mit zusätzlicher Konfundierung aller Interaktionen erster Ordnung und/oder mit fraktionalisierten statt randomisierten Setbildungen. Dies erscheint ange-

39 Bei Berücksichtigung lediglich eines (oder weniger) Merkmale ist eine hohe Antwortkonsistenz schließlich keine kognitive ‚Herausforderung' – als Indikator für eine valide Messung ist sie gerade darum nicht hinreichend.

bracht, weil durch die Auswahl und Zusammenstellung von Vignetten unweigerlich die im Vignettenuniversum gegebene, vollständige Orthogonalität aller Dimensionen und ihrer Interaktionen verloren geht. In diesem Zusammenhang ist nicht *gänzlich* auszuschließen, dass sich die dadurch hervorgerufene Verringerung der statistischen Effizienz unterschiedlich auf die hier gegenübergestellten Gruppen mit weniger oder mehr Vignettendimensionen bzw. auf die Gruppen von Vignetten vor und ab dem Auftreten unplausibler Fälle verteilt, was ihre Vergleichbarkeit etwas beeinträchtigen könnte. Zudem sind weitere, hier aus Platzgründen nicht angesprochene methodische Aspekte von Interesse, wie beispielsweise die Wahl möglichst geeigneter Präsentationsformen (Fließtext oder tabellarische Darstellung) und der Einsatz von unterschiedlichen Antwortskalen. Darüber hinaus wären andere statistische Auswertungsverfahren zu erproben. Der gängigen Praxis folgend wurden die Vignettenurteile als metrisch behandelt, genau genommen weisen sie lediglich ordinales Skalenniveau auf. Die Wahl von OLS-Schätzungen wird zwar allgemein durch ihre hohe Robustheit und bessere Interpretierbarkeit gerechtfertigt (Winship/Mare 1984); speziell für Vignettenstudien wurde bislang aber noch zu wenig ausgelotet, welche Analysegewinne sich mit adäquateren – besser mit der Datenstruktur korrespondierenden – (Mehr-)Ebenenverfahren erzielen lassen.[40]

Mit den Daten des DFG-geförderten Projekts ‚Der Faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen' können eine Vielzahl der benannten Aspekte untersucht werden. Nach den hier präsentierten, ersten Befunden verdienen sie in methodischer wie inhaltlicher Hinsicht stärkere Beachtung.

Literatur

Alves, W. M. und P. H. Rossi, 1978: Who should get what? Fairness judgments of the distribution of earnings. American Journal of Sociology 84: 541-564.

Auspurg, K., M. Abraham und Th. Hinz, 2009: Wenig Fälle, viele Informationen: Die Methodik des faktoriellen Surveys als Paarbefragung. S. 179-210 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Auspurg, K. und M. Abraham, 2007: Die Umzugsentscheidung von Paaren als Verhandlungsproblem. Eine quasiexperimentelle Überprüfung des Bargaining-Modells. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 271-293.

40 Aufgrund der mehrfachen Bewertungsaufgabe sind zumindest bei den hier verwendeten, geschlossenen Antwortskalen Zensierungen der Urteile zu befürchten (wurden bereits extreme Urteile abgegeben, lässt sich das Antwortverhalten nicht mehr hinreichend abstufen). Dies kann ebenfalls zu verzerrten Ergebnissen führen und legt den Einsatz von einschlägigen Regressionsverfahren (z. B. Tobit-Modellen) nahe. Auch hier sind Wechselwirkungen mit dem Auftreten von unplausiblen Fällen zu erwarten, da diese vermehrt zu extremen Antworten motivieren dürften.

Beck, M. und K.-D. Opp, 2001: Der faktorielle Survey und die Messung von Normen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 283-306.
Berk, R. A. und P. H. Rossi, 1977: Prison reform and state elites. Cambridge: Ballinger.
Buskens, V. und J. Weesie, 2000: An experiment on the effects of embeddedness in trust situations. Buying a used car. Rationality and Society 12: 227-253.
Bradley, M. und A. Daly, 1994: Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21: 167-184.
Bryan, S., L. Gold, R. Sheldon und M. Buxton, 2000: Preference measurement using conjoint methods. An empirical investigation of reliability. Health Economics 9: 385-395.
Carroll, D. J. und P. E. Green, 1995: Psychometric methods in marketing research. Part I, Conjoint analysis. Journal of Marketing Research 32: 358-391.
Carson, R., J. J. Louviere, D. A. Anderson, P. Arabie, D. Bunch, D. A. Hensher, R. M. Johnsons, W. F. Kuhfeld, D. Steinberg, J. Swait und H. Timmerman, 1994: Experimental analysis of choice. Marketing Letters 5: 351-368.
Caussade, S., J. de D. Ortúzar, L. I. Rizzi und D. A. Hensher, 2005: Assessing the influence of design dimensions on stated choice experiment estimates. Transportation research part B: Methodological 39: 621-640.
Creyer, E. und W. T. Ross, 1988: The effect of range-frequency manipulations on conjoint importance weight stability. Advances in Consumer Research 15: 505-509.
DeShazo, J. R. und G. Fermo, 2002: Designing choice sets for stated preference methods. The effects of complexity on choice consistency. Journal of Environmental Economics and Management 44: 123-143.
Diefenbach, H. und K.-D. Opp, 2007: When and why do people think there should be a divorce? An application of the factorial survey. Rationality and Society 19: 485-517.
Diekmann, A., 2007. Empirische Sozialforschung. Reinbek bei Hamburg: Rowohlt.
Dülmer, H., 2001: Bildung und der Einfluss von Argumenten auf das moralische Urteil. Eine empirische Analyse zur moralischen Entwicklungstheorie Kohlbergs. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 1-27.
Dülmer, H. und M. Klein, 2003: Die Messung gesellschaftlicher Wertorientierungen via Conjoint- und Vignettenanalyse: Ein Ansatz zur adäquaten Operationalisierung von Ingelharts materialistischen und postmaterialistischen Wertorientierungen. Unveröffentlichter Abschlussbericht an die Fritz-Thyssen Stiftung.
Dülmer, H., 2007: Experimental plans in factorial surveys. Random or quota design? Sociological Methods & Research 35: 382-409.
Eifler, S., 2007: Evaluating the validity of self-reported deviant behavior using vignette analyses. Quality & Quantity 41: 303-318.
Faia, M., 1980: The vagaries of the vignette world. A comment on Alves and Rossi. American Journal of Sociology 85: 951-954.
Garrett, K., 1982: Child abuse: problems of definition. S. 177-204 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Greene, W. H., 2003: Econometric Analysis. Prentice Hall: New York.
Groß, J. und C. Börensen, 2009: Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung. S. 149-178 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.
Hechter, M., J. Ranger-Moore, G. Jasso und C. Horne, 1999: Do values matter? An analysis of advance directives for medical treatment. European Sociological Review 15: 405-430.
Hechter, M., H. Kim und J. Baer, 2005: Prediction versus explanation in the measurement of values. European Sociological Review 21: 91-108.
Hembroff, L. A., 1987: The seriousness of acts and social contexts. A test of Black's theory of the behavior of law. American Journal of Sociology 93: 322-347.

Hermkens, P. L. J. und F. A. Boerman, 1989: Consensus with respect to the fairness of incomes. Differences between social groups. Social Justice Research 3: 201-215.
Hensher, D. A., 2004: How do respondents handle stated choice experiments? Information processing strategies under varying information load. Working paper 04-14. University of Sydney: Institute of Transport Studies.
Hensher, D. A., 2006: Revealing differences in willingness to pay due to the dimensionality of stated choice designs. An initial assessment. Environmental & Resource Economics 34: 7-44.
Horne, C., 2003: The internal enforcement of norms. European Sociological Review 19: 335-343.
Hox, J. J., I. Kreft und P. Hermkens, 1991: The analysis of factorial surveys. Sociological Methods & Research 19: 493-510.
Jann, B., 2003: Lohngerechtigkeit und Geschlechterdiskriminierung. Experimentelle Evidenz. Unveröffentlichtes Manuskript an der Eidgenössischen Technischen Hochschule Zürich.
Jasso, G., 1988: Whom Shall We Welcome? Elite judgments of the criteria for the selection of immigrants. American Sociological Review 53: 919-932.
Jasso, G., 1994: Assessing individual and group differences in the sense of justice. Framework and application to gender differences in the justice of earnings. Social Science Research 23: 368-406.
Jasso, G. und M. Webster, 1997: Double standards in just earnings for male and female workers. Social Psychology Quarterly 60: 66-78.
Jasso, G. und K.-D. Opp, 1997: Probing the character of norms. A factorial survey analysis of the norms of political action. American Sociological Review 62: 947-964.
Jasso, G. und M. Webster, 1999: Assessing the gender gap in just earnings and its underlying mechanisms. Social Psychology Quarterly 62: 367-380.
Jasso, G., 2006: Factorial survey methods for studying beliefs and judgments. Sociological Methods and Research 34: 334-423.
John S., C. und N. A. Bates, 1990: Racial composition and neighborhood evaluation. Social Science Research 19: 47-61.
Johnson, R. F., 2006: Comment on "Revealing differences in willingness to pay due to the dimensionality of stated choice designs. An initial assessment". Environmental & Resource Economics 34: 45-50.
Klein, M., 2002: Die Conjoint-Analyse. Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. ZA-Information 50: 7-45. http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/archiv/za-information/ (16.3.2009).
Kuhfeld, W. F., T. D. Randall und M. Garratt, 1994: Efficient experimental design with marketing research applications. Journal of Marketing Research 31: 545-557.
Kuhfeld, W. F., 2005: Marketing research methods in SAS. Experimental design, choice, conjoint and graphical techniques. Cary: SAS Institute.
Liebig, S. und S. Mau, 2002: Einstellungen zur sozialen Mindestsicherung. Ein Vorschlag zur differenzierten Erfassung normativer Urteile. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 109-134.
Liebig, S. und S. Mau, 2005: Wann ist ein Steuersystem gerecht? Zeitschrift für Soziologie 34: 468-491.
Liebig, S., A. Meyermann und A. Schulze, 2006: Temporal stability of justice evaluations. Paper presented at the 11th conference of the international society for justice research. Berlin: Humboldt Universität.
Louviere, J. J., 2001a: What if consumer experiments impact variances as well as means? Response variablility as a behavioral phenomen. Journal of Consumer Research 28: 506-511.
Louviere, J. J., 2001b: Choice experiments. An overview of concepts and issues. S. 13-36 in: Bennett, J. und R. Blamey (Hg.): The choice modelling approach to environmental valuation. Cheltenham/Northhampton: Edward Elgar.
Mayerl, J., P. Selke und D. Urban, 2005: Analyzing cognitive processes in CATI-Surveys with response latencies. An empirical evaluation of the consequences of using different

baseline speed measures. Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart: SISS 2/2005.
Melles, Th., 2001: Framing-Effekte in der Conjoint-Analyse. Ein Beispiel für Probleme der Merkmalsdefinition. Aachen: Shaker.
Meudell, M. B., 1982: Household and social standing. Dynamic and static dimensions. S. 69-94 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Miller, J. L., P. H. Rossi und J. E. Simpson, 1986: Perceptions of justice. Race and gender differences in judgments of appropriate prison sentences. Law & Society Review 20: 313-334.
Nisic, N. und K. Auspurg, 2009: Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich – Validität, Grenzen und Möglichkeiten beider Ansätze. S. 211-246 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.
Nock, S. L., 1982: Family social standing. Consensus on characteristics. S. 95-118 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Ohler, T., A. Le, J. Louviere und J. Swait, 2000: Attribute range effects in binary response tasks. Marketing Letters 11: 249-260.
Orme, B., 2006: Getting started with conjoint analysis. Strategies for product design and pricing research. Madison/Wisconsin: Research Publishers LLC.
O`Toole, R., S. W. Webster, A. W. O`Toole und B. Lucal, 1999: Teachers´ recognition and reporting of child abuse. A factorial survey. Child Abuse & Neglect 23: 1083-1101.
Perrey, J., 1996: Erhebungsdesign-Effekte bei der Conjoint-Analyse. Marketing – Zeitschrift für Forschung und Praxis 18: 105-116.
Rooks, G., W. Raub, R. Selten, und F. Tazelaar, 2000: How inter-firm co-operation depends on social embeddedness: A vignette study. Acta Sociologica 43: 123-137.
Rossi, P. H., W. A. Sampson, C. E. Bose, G. Jasso und J. Passel, 1974: Measuring household social standing. Social Science Research 3: 169-190.
Rossi, P. H., 1979: Vignette analysis. Uncovering the normative structure of complex judgments. S. 176-186 in: R. K. Merton, J. S. Coleman und P. H. Rossi (Hg.): Qualitative and Quantitative Social Research. Papers in honour of Paul F. Lazarsfeld. New York: Free Press.
Rossi, P. H. und W. M. Alves, 1980: Rejoinder to Faia. The American Journal of Sociology 85: 954-955.
Rossi, P. H. und A. B. Anderson, 1982: The factorial survey approach. An introduction. S. 15-67 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgments. The factorial survey approach. Beverly Hills u.a.: Sage.
Rossi, P. H. und S. L. Nock, 1982: Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Sauer, C., 2009: Methodische Probleme von Conjoint- und Vignettenanalysen – Literaturreview. Arbeitsbericht Nummer 1 des Projekts „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen". Bielefeld/Konstanz: Universität Bielefeld/Universität Konstanz.
Seyde, C., 2005: Beiträge und Sanktionen in Kollektivgutsituationen. Ein faktorieller Survey. Arbeitsbericht 42 des Instituts für Soziologie. Leipzig: Universität Leipzig.
Shepelak, N. J. und D. F. Alwin, 1986: Beliefs about Inequality and Perceptions of Distributive Justice. American Sociological Review 51: 30-46.
Shlay, A. B., H. Tran, M. Weinraub und M. Harmon, 2005: Teasing apart the child care conundrum. A factorial survey analysis of perceptions of child care quality, fair market price and willingness to pay by low-income, African American parents. Early Childhood Research Quarterly 20: 393-416.
Smith, T. W., 1986: A study of non-response and negative values on the factorial vignettes on welfare. GSS Methodological Report 44. Chicago: NORC.
Sniderman, P. M. und D. B. Grob, 1996: Innovations in experimental design in attitude surveys. Annual Review of Sociology 22: 377-399.

Steiner, P. M. und C. Atzmüller, 2006: Experimentelle Vignettendesigns in faktoriellen Surveys. Kölner Zeitschrift für Soziologie und Sozialpsychologie 58: 117-146.
Struck, O., A. Krause und C. Pfeifer, 2008: Entlassungen: Gerechtigkeitsempfinden und Folgewirkungen. Theoretische Konzepte und empirische Ergebnisse. Kölner Zeitschrift für Soziologie und Sozialpsychologie 60: 102-122.
Swait, J. und W. Adamowicz, 2001: The influence of task complexity on consumer choice. A latent class model of decision strategy switching. Journal of Consumer Research 28: 135-148.
Urban, D. und J. Mayerl, 2007: Antwortlatenzzeiten in der survey-basierten Verhaltensforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 692-713.
Wagner, G. G., J. R. Frick und J. Schupp, 2007: The German Socio-Economic Panel Study (SOEP) – Evolution, scope and enhancements. Schmollers Jahrbuch (Journal of Applied Social Science Studies) 127 (1): 139-169.
Wason, K. D., M. J. Polonsky und M. R. Hyman, 2002: Designing vignette studies in marketing. Australasian Marketing Journal 10: 41-58.
Will, J. A., 1993: The dimensions of poverty. Public perceptions of the deserving poor. Social Science Research 22: 312-332.
Winship, C. und R. D. Mare, 1984: Regression models with ordinal variables. American Sociological Review 49: 512-525.
Wittink, D. R., L. Krishnamurthi und J. B. Nutter, 1982: Comparing derived importance weights across attributes. Journal of Consumer Research 8: 471-474.
Wittink, D. R., L. Krishnamurthi und D. J. Reibstein, 1989: The effect of differences in the number of attribute levels on conjoint results. Marketing Letters 1: 113-123.
Wooldridge, J. M., 2002: Econometric analysis of cross section and panel data. Cambridge/Mass.: MIT Press.
Wooldridge, J. M., 2003: Introductory econometrics. A modern approach. Mahson, Ohio: Thomson.
Zimbardo, P. G., 1988: Psychologie. Berlin u. a.: Springer.

Korrespondenzadresse: Katrin Auspurg
Universität Konstanz
Fach D40
78457 Konstanz
katrin.auspurg@uni-konstanz.de

Statistischer Anhang

Tabelle A1 Übersicht über die Korrelationen der Dimensionen in den Splits mit 5 und 12 Dimensionen[a]

	Split mit 5 Dimensionen				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	-0,056	1,000			
Abschluss	-0,019	0,016	1,000		
Berufsprestige	0,017	0,096	-0,023	1,000	
Einkommen	-0,027	0,008	0,042	0,148	1,000
	Split mit 12 Dimensionen				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	0,034	1,000			
Abschluss	-0,007	-0,022	1,000		
Berufsprestige	0,032	0,049	-0,031	1,000	
Einkommen	0,030	-0,031	0,073	0,172	1,000

[a] *Korrelationskoeffizient nach Pearson bzw. bei der ordinalen Dimension ‚Abschluss' Rang-Korrelationskoeffizient nach Spearman.*

Tabelle A2 Übersicht über die Korrelationen der Dimensionen in den Splits mit 7 und 10 Vignetten pro Befragten[a]

	Split mit 7 Vignetten				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	-0,033	1,000			
Abschluss	0,003	0,002	1,000		
Berufsprestige	0,039	0,070	-0,031	1,000	
Einkommen	-0,015	-0,016	0,055	0,182	1,000
	Split mit 10 Vignetten				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	0,032	1,000			
Abschluss	-0,050	-0,011	1,000		
Berufsprestige	-0,006	0,080	-0,016	1,000	
Einkommen	0,038	-0,004	0,059	0,115	1,000

[a] *Korrelationskoeffizient nach Pearson bzw. bei der ordinalen Dimension ‚Abschluss' Rang-Korrelationskoeffizient nach Spearman.*

Tabelle A3 Übersicht über die Korrelationen[a] aller zwölf Vignettendimensionen (nur Split mit 12 Dimensionen)

	Geschlecht	Alter	Abschluss	Berufs-prestige	Einkommen	Berufs-erfahrung	Betriebs-zugehörigkeit	Leistung	Betriebs-größe	Wirtsch. Lage	Gesundheit	Kinder
Geschlecht	1,000											
Alter	0,034	1,000										
Abschluss	-0,007	-0,022	1,000									
Berufsprestige	0,032	0,049	-0,031	1,000								
Einkommen	0,030	-0,031	0,073	0,172	1,000							
Berufserfahrg.	0,018	0,187	-0,107	0,109	0,045	1,000						
Betriebszug.	0,024	0,279	-0,101	0,093	0,117	0,416	1,000					
Leistung	-0,015	0,051	-0,025	-0,015	0,001	0,027	0,013	1,000				
Betriebsgröße	-0,036	0,072	0,156	-0,092	-0,153	-0,012	0,212	-0,083	1,000			
Wirtsch. Lage	0,049	0,059	-0,031	0,091	-0,032	0,042	0,059	0,012	0,075	1,000		
Gesundheit	0,027	0,015	0,013	0,086	-0,165	-0,013	-0,041	-0,072	-0,014	-0,023	1,000	
Kinder	0,018	0,101	0,023	-0,051	0,001	0,088	-0,064	-0,005	0,063	0,023	-0,087	1,000

[a] *Bei metrischen und binären Variablen Korrelationskoeffizient nach Pearson; bei ordinalen Variablen (Abschluss, Leistung und wirtschaftliche Lage) Rang-Korrelationskoeffizient nach Spearman.*