

Die Wirkung von Incentives
auf die Antwortqualität in Umfragen

Dissertation
zur Erlangung des Doktorgrades
der Sozialwissenschaftlichen Fakultät
der Georg-August-Universität Göttingen

vorgelegt von
André Dingelstedt
geboren in Halle (Saale)

Göttingen, 2015

Betreuungsausschuss

Erstbetreuer: Prof. Dr. Steffen-M. Kühnel

Weitere Betreuer: PD Dr. Micha Strack

Prof. Dr. Peter Preisendörfer

Weitere Mitglieder der Prüfungskommission:

Tag der mündlichen Prüfung: 24.11.2015

Inhaltsverzeichnis

1. Einleitung	1
2. Das Konzept der Datenqualität	4
2.1 Der Total Survey Error	4
2.1.1 Der „wahre“ Wert.....	9
2.1.2 Die Umsetzbarkeit des Konzepts des „wahren“-Werts in der Umfrageforschung.....	10
2.1.3 Nutzbarkeit des Total Survey Errors zur Definition von Qualität	13
2.2 Satisficing und Optimizing	15
2.2.1 Weak Satisficing und Strong Satisficing	19
2.2.1.1 Weak Satisficing.....	20
2.2.1.2 Strong Satisficing	23
2.2.2 Erinnerungsstrategien bei Häufigkeitsfragen	25
2.2.3 Die Nutzbarkeit des Satisficing-Konzepts zur Definition von Antwortqualität	27
2.3 Mikrozensusgesetz	28
2.3.1 Die wahrheitsgemäße Beantwortung.....	30
2.3.2 Die Vollständigkeit von Angaben	30
2.3.3 Das Einhalten gesetzter Fristen.....	32
2.4 Die Definition von Antwortqualität	33
3. Die Bedeutung von Incentives für die Umfrageforschung	34
3.1 Die Rücklaufquote	39
3.2 Die Rücklaufgeschwindigkeit.....	42
3.3 Die Zusammensetzung der Stichprobe	46
3.4 Die Qualität der Umfrage	47
4. Theorie	50
4.1 Cognitive Evaluation Theory	50
4.1.1 Forschungsstand zur Cognitive Evaluation Theory.....	56
4.1.2 Hypothesen zur Cognitive Evaluation Theory.....	58
4.2 Reziprozitätshypothese	59
4.2.1 Forschungsstand zur Reziprozitätshypothese	63
4.2.2 Hypothesen zur Reziprozität	64
4.3 Zusammenführung der Cognitive Evaluation Theory und der Reziprozitätshypothese	66
5. Aufbau der Studie	67
6. Operationalisierung	78
6.1 Die Messung der unabhängigen Variablen	78
6.1.1 Intrinsische Motivation	78

6.1.2 Reziprozität	83
6.1.3 Identifizierte Regulation (extrinsische Motivation)	85
6.2 Die Messung der abhängigen Variablen: die Indikatoren für Antwortqualität	85
6.2.1 Indikatoren für ein durchdachtes Bearbeiten eines Fragebogens.....	85
6.2.2 Indikatoren für ein (situational) wahrheitsgemäßes Bearbeiten des Fragebogens.....	92
6.2.3 Indikatoren für ein vollständiges Bearbeiten eines Fragebogens.....	95
6.2.4 Indikatoren für ein anweisungsfolgendes Bearbeiten des Fragebogens.....	100
7. Erste Analysen zur Bewertung der Messinstrumente der erklärenden Variablen.....	101
7.1 Indikatoren der intrinsischen Motivation.....	101
7.1.1 Der Zusammenhang zwischen den Messungen für intrinsische Motivation.....	101
7.1.2 Die Selbsteinstufung zur intrinsischen Motivation auf Basis der SIMS	103
7.1.3 Die Bearbeitungszeit als Indikator der intrinsischen Motivation	106
7.1.4 Die Wiederbefragungsbereitschaft als Indikator der intrinsischen Motivation.....	108
7.1.5 Die Worthäufigkeit als Indikator für die intrinsische Motivation.....	109
7.1.6 Die Auswahl eines angemessenen Indikators der intrinsischen Motivation	111
7.2 Indikatoren der Reziprozitätshypothese	114
7.3 Der Indikator für die identifizierte Regulation (extrinsische Motivation).....	117
8. Erste Analysen zu den abhängigen Variablen.....	120
8.1 Analyse der Indikatoren für ein durchdachtes Bearbeiten des Fragebogens:	120
8.1.1 Akquieszenz und Status Quo-Effekt.....	120
8.1.2 Die genutzte Erinnerungsstrategie zur Beantwortung von Häufigkeitsfragen.....	122
8.1.3 Indikatoren für ein durchdachtes Bearbeiten des Fragebogens: Konsistenz	124
8.1.4 Anzahl an Worten in offenen Fragen	129
8.1.5 Zusammenfassung der Indikatoren eines durchdachten Bearbeitens eines Fragebogens	133
8.2 Analyse der Indikatoren für ein wahrheitsgemäßes Bearbeiten des Fragebogens	135
8.2.1 Die Äußerung von Pseudo-Opinions (Falschangaben)	135
8.2.2 Soziale Erwünschtheit	138
8.2.3 Zusammenfassung der Indikatoren eines (situational) wahrheitsgemäßen Bearbeitens des Fragebogens	141
8.3 Analyse der Indikatoren für ein vollständiges Bearbeiten des Fragebogens	142
8.3.1 Das Überspringen von Fragebogenfragen	142
8.3.2 Das Ausweichverhalten bei Filterfragen.....	146
8.3.3 Zusammenfassung für die Indikatoren eines vollständigen Bearbeitens des Fragebogens.....	148
8.4 Indikatoren für ein anweisungsbefolgendes Bearbeiten des Fragebogens	149
8.5 Der Zusammenhang zwischen den ausgewählten Indikatoren der Antwortqualität	150

9. Hypothesenprüfung	151
9.1 Die Prüfung der Hypothesen zur Wirkung der intrinsischen Motivation und der Reziprozität auf die Antwortqualität	152
9.2 Prüfung der Hypothesen zur extrinsischen Motivation	166
9.3 Zusammenfassung der Ergebnisse	173
10. Diskussion der internen und externen Validität der Ergebnisse	176
11. Fazit	180
12. Literaturverzeichnis	184
Anhang	201

Abbildungsverzeichnis

Abbildung 1: Die ideale Position von monetären Incentives in postalischen Befragungen	2
Abbildung 2: Die Aufteilung der Fehlerkategorien nach Weisberg	5
Abbildung 3: Elemente des Total Survey Error zugeordnet in vier Befragungsphasen	13
Abbildung 4: Vom Optimizing zum Satisficing	20
Abbildung 5: Antwortstrategien bei Fragen zu Häufigkeiten	26
Abbildung 6: Das Waage-Modell nach Groves et al.	37
Abbildung 7: Response-Raten untergliedert nach Höhe des Incentive	39
Abbildung 8: Rücklaufgeschwindigkeit bei einer Umfrage mit unkonditionalem 1€-Incentive	42
Abbildung 9: Rücklaufgeschwindigkeit bei Vergabe oder Ankündigung einer 10 Sfr Telefonkarte	43
Abbildung 10: Die Rücklaufgeschwindigkeit bei einfacher und doppelter Vergabe von Incentives	44
Abbildung 11: Die Rücklaufgeschwindigkeit bei Beilegung eines monetären Incentives – Welle 1	45
Abbildung 12: Die Rücklaufgeschwindigkeit bei Beilegung eines monetären Incentives – Welle 2	45
Abbildung 13: Verschiedenen Arten der Motivationen nach Ryan & Deci	52
Abbildung 14: Die Wirkung von Belohnungen nach Deci et al. (1985)	55
Abbildung 15: Die Hypothesen bezüglich der Wirkung der intrinsischen Motivation und der Reziprozität auf die Antwortqualität	66
Abbildung 16: Die Hypothesen bezüglich der Wirkung der intrinsischen und der extrinsischen Motivation auf die Antwortqualität	67
Abbildung 17: Die drei Arten eines Experiments	68
Abbildung 18: Die Darstellung der Rücklaufquote unter Berücksichtigung der Einführung alternativer Rekrutierungsstrategien	74
Abbildung 19: Darstellung verschiedener Filterformate	90
Abbildung 20: Drei ausgewählte Antworten zu der offenen Frage: „Bitte beschreiben Sie, wo und wie Sie zum ersten Mal von der Droge LA-42 etwas mitbekommen haben.“	148
Abbildung 21: Die Hypothesen bezüglich der Wirkung der intrinsischen Motivation und der Reziprozität der Antwortqualität	152
Abbildung 22: Graphische Darstellung des Strukturgleichungsmodells bei restriktiver dellierung der Zusammenhänge der Faktoren auf die Indikatoren der Antwortqualität, Prüfung der Hypothesen 1a – 2c	153
Abbildung 23: Graphische Darstellung des Strukturgleichungsmodells bei Freigabe des restriktiven Zusammenhangs der verinnerlichten Reziprozitätsnorm, zur Prüfung der Hypothesen 1a – 2c	155
Abbildung 24: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe ohne Incentive, zur Prüfung der Hypothesen 1a – 2c	156
Abbildung 25: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe mit einem Incentive in Höhe von 5€, zur Prüfung der Hypothesen 1a – 2c	156
Abbildung 26: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe mit einem Incentive in Höhe von 20€, zur Prüfung der Hypothesen 1a – 2c	157
Abbildung 27: Die Hypothesen bezüglich der Wirkung der intrinsischen Motivation und der extrinsischen Motivation auf die Antwortqualität	166
Abbildung 28: Strukturgleichungsmodell für die Versuchsgruppe ohne Incentive, zur Prüfung der Hypothesen 3a und 3b	167
Abbildung 29: Strukturgleichungsmodell für die Versuchsgruppe mit einem Incentive in Höhe von 5€, zur Prüfung der Hypothesen 3a und 3b	167
Abbildung 30: Strukturgleichungsmodell für die Versuchsgruppe mit einem Incentive in Höhe von 20€, zur Prüfung der Hypothesen 3a und 3b	168

Tabellenverzeichnis

Tabelle 1: Die Komponenten des Non-Sampling Error nach Biemer & Lyberg	6
Tabelle 2: Beispiele für Antwortverhalten bei Weak und Strong Satisficing	20
Tabelle 3: Die Wirkung von Frames auf die Teilnahmebereitschaft	35
Tabelle 4: Aktuell angestrebter Studienabschluss	75
Tabelle 5: Die Aufteilung der befragten Studierenden nach Fakultäten	76
Tabelle 6: Finanzierungsquellen während des Studiums	77
Tabelle 7: Faktorladungen der explorativen Faktorenanalyse für die ausgewählten Indikatoren der intrinsischen Motivation	102
Tabelle 8: Die Korrelationen zwischen den drei extrahierten Faktoren	103
Tabelle 9: Korrelationen zwischen den SIMS-Items der intrinsischen Motivation	105
Tabelle 10: Mittelwerte für die stark zusammenhängenden Items der intrinsischen Motivation	105
Tabelle 11: Mittelwerte für die Bearbeitungszeit für den gesamten Fragebogen	106
Tabelle 12: Mittelwerte für die Bearbeitungsdauer, aufgegliedert nach den drei Fragebogenblöcken	107
Tabelle 13: Häufigkeitstabelle für die Wiederbefragungsbereitschaft	108
Tabelle 14: Häufigkeitstabelle für die Wiederbefragungsbereitschaft nach erneuter Rückfrage	109
Tabelle 15: Mittelwerte für die Anzahl der Worte	110
Tabelle 16: Mittelwerte für die Anzahl an Zeichen	110
Tabelle 17: Korrelationen zwischen den Items zur Messung der verinnerlichten Reziprozitätsnorm	115
Tabelle 18: Mittelwerte für den Mittelwertindex der verinnerlichten Reziprozitätsnorm, aufgegliedert nach den drei Versuchsgruppen	116
Tabelle 19: Mittelwerte für den Mittelwertindex der Bewertung des Versuchsleiters, aufgegliedert nach den drei Versuchsgruppen	117
Tabelle 20: Korrelationen zwischen den Items der extrinsischen Motivation	118
Tabelle 21: Mittelwerte für den Mittelwertindex der stark zusammenhängenden Items der extrinsischen Motivation, aufgegliedert nach den drei Versuchsgruppen	119
Tabelle 22: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen dem Status Quo-Effekt und der Incentivierung	122
Tabelle 23: Häufigkeitsverteilung von Erinnerungsstrategien zur Beantwortung einer Häufigkeitsfrage	123
Tabelle 24: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen der Erinnerungsstrategie und der Incentivierung	123
Tabelle 25: Häufigkeitstabelle über die zwei Konsistenzmessungen bei kurzer zeitlicher Abfolge	125
Tabelle 26: Häufigkeitsverteilung der Summe der Abweichungen aller Politikerbewertungen zu zwei Messzeitpunkten	127
Tabelle 27: Korrelationen zwischen den Messungen für Konsistenz bei kurzer und langer Zeitlicher Abfolge	129
Tabelle 28: Korrelationen der Worte zwischen den drei offenen Fragen	130
Tabelle 29: Korrelationen der Zeichen zwischen den drei offenen Fragen	130
Tabelle 30: Mittelwerte des Gesamtindex für die Anzahl der Worte, aufgegliedert nach den drei Versuchsgruppen	131
Tabelle 31: Mittelwerte der Anzahl der Worte in den Kommentaren aufgegliedert nach den drei Versuchsgruppen	132
Tabelle 32: Metrische Verteilungsinformationen zu der Anzahl der Worte in den Kommentaren	132
Tabelle 33: Korrelationen zwischen den Messungen der Pseudo-Opinions	136
Tabelle 34: Häufigkeiten der Pseudo-Opinions	137
Tabelle 35: Die Mittelwerte der Falschangaben, aufgegliedert nach den drei Versuchsgruppen	138
Tabelle 36: Mittelwerte der Fremd- und Selbsttäuschung, aufgegliedert auf die drei Versuchsgruppen	139
Tabelle 37: Die Mittelwerte der Antworten zu dem rekodierten Item „Ich bin mir oft unsicher in meinem Urteil“, aufgegliedert nach den drei Versuchsgruppen	140

Tabelle 38: Korrelationen zwischen den Pseudo-Opinions und Dimensionen der sozialen Erwünschtheit	141
Tabelle 39: Die Häufigkeit an Übersprüngen ohne Begründung über einen Kommentar	145
Tabelle 40: Mittelwerte zu den nicht begründeten Übersprüngen, aufgegliedert nach den drei Versuchsgruppen	146
Tabelle 41: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen dem Filterverhalten und der Incentivierung	147
Tabelle 42: Faktorladungen der explorativen Faktorenanalyse für die ausgewählten Indikatoren der Antwortqualität	150
Tabelle 43: Unstandardisierte Regressionsgewichte des Strukturgleichungsmodells, aufgegliedert nach den drei Versuchsgruppen (Hypothesen 1a – 2c)	159
Tabelle 44: Unstandardisierte Regressionsgewichte des Strukturgleichungsmodells, aufgegliedert nach den drei Versuchsgruppen (Hypothesen 3a – 3b)	169
Tabelle 45: Übersicht über die Ergebnisse der Hypothesenprüfungen	175

1. Einleitung

Die standardisierte Befragung ist in der sozialwissenschaftlichen Forschung ein anerkanntes und häufig genutztes Erhebungsverfahren, um Einblicke in die Einstellungen von Bevölkerungsgruppen zu erlangen. In den letzten Jahrzehnten konnte jedoch ein deutlicher Rückgang der Teilnahmebereitschaft an Umfragen festgestellt werden. Tourangeau (2007) führt dies auf einen einfachen Mechanismus zurück:

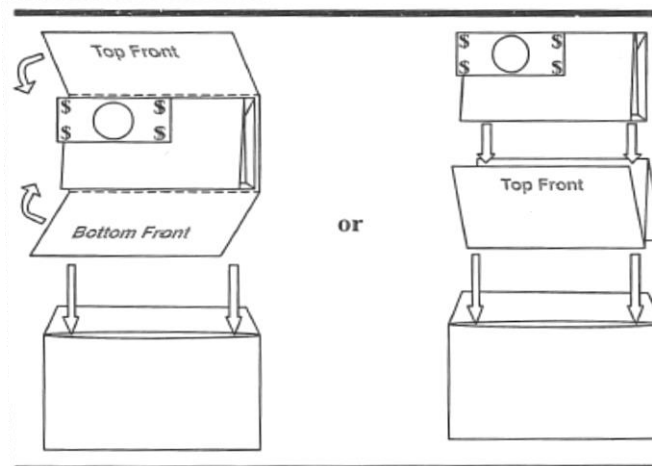
„people are busier (or at least they *feel* busier), and they've adopted strategies for fending off unwanted intrusions. Although surveys probably constitute a minor portion of the impositions of contemporary life, the defensive measures people now habitually take serve to filter our survey requests along with other intrusions. Beyond that, it is harder to see surveys in the idealistic light in which the founders of survey research (and presumably the general public) saw them 70 years ago. Civic engagement probably is declining generally, and, in any case, survey participation no longer seems the altruistic gesture it once did" (Tourangeau (2007), S. 252).

Aus dieser Problematik heraus etablierte sich der Gedanke, den potentiell Befragten¹ Anreize bei Teilnahme in Aussicht zu stellen oder vor Beginn der Befragung Geschenke zuzusenden. Diese Anreize bzw. Geschenke - auch Incentives genannt - sind zumeist monetärer Natur. Die daraus resultierenden Steigerungen der Teilnahmebereitschaft waren sehr deutlich (vgl. Church (1993); Singer (1998)), so dass sich Incentives als probates Mittel zur Teilnahmesteigerung etablierten. Um die Wirkung von Incentives zu optimieren, wurden bis heute vielzählige Leitfäden und Ratgeber veröffentlicht (vgl. Pforr (2015); Dillman (2009); Stadtmüller & Porst (2005)). In der Tailored Design Method geben Dillman et al. (2009) beispielhaft Vorschläge

¹ Im Folgenden wird aus Gründen der sprachlichen Vereinfachung nur die männliche Form verwendet. Es sind jedoch stets Personen männlichen und weiblichen Geschlechts gleichermaßen gemeint.

dafür, wie die Geldscheine bei postalischen Befragungen in die Umschläge eingefügt werden sollten, damit sie sofort von den Empfängern wahrgenommen werden:

Abb. 1: Die ideale Position von monetären Incentives in postalischen Befragungen



Quelle: Dillman et al. (2009), S. 265.

Seit den 80er Jahren wird auch untersucht, inwiefern Incentives sich auf das Befragtenverhalten während der Bearbeitung des Fragebogens auswirken (vgl. Hansen (1980); Berk et al. (1987); James & Bolstein (1990)). Hierbei kann die Vermutung geäußert werden, dass die Befragten primär am Incentive interessiert sind und die Fragebögen daher nicht sorgfältig bearbeiten. Um dieser Frage nachzugehen, wurden Studien durchgeführt, wobei tendenziell keine negativen Effekte auf die Datenqualität aufgedeckt werden konnten (Pffor (2015); Boulianne (2008)). Es muss hierbei jedoch kritisch darauf hingewiesen werden, dass die Studien zumeist keine klare Definition des Begriffs der Datenqualität aufweisen und die Indikatoren demzufolge ohne theoretische Absicherung verwendet werden. Darüber hinaus fehlen im Forschungsfeld empirisch abgesicherte Theorien zur Erklärung der Wirkung von Incentives auf die Datenqualität in Befragungen. Eine theoretische Absicherung erscheint umso wichtiger, da in aktuellen Studien häufiger negative Befunde zur Antwortqualität aufgrund der Incentivierung berichtet werden (Barge & Gehlbach (2012)). Barge & Gehlbach warnen daher vor einem

unüberlegten Einsatz von Incentives: „What remains unclear, however, is whether unintended and perhaps negative consequences may result from using incentives. If it turns out that incentives can degrade item-level data quality under certain situations, many institutions may need to rethink their data collection plans“ (Barge & Gehlbach (2012), S. 26).

Ziel der vorliegenden Arbeit ist daher auf Grundlage theoretischer Konzepte – unter Verwendung eines Experiments – die Frage zu klären, ob und inwiefern Incentives systematisch auf die Antwortqualität wirken. Hierfür wird im zweiten Kapitel dieser Studie zuerst der Begriff der Antwortqualität hergeleitet und definiert. Im dritten Kapitel wird dann der aktuelle Forschungsstand zu der Wirkung von Incentives vorgestellt und zusammengefasst. Im vierten Kapitel werden zwei theoretische Ansätze vorgestellt, welche in dieser Studie als Grundlage zur Erklärung der Wirkung von Incentives auf die Antwortqualität dienen. Dies ist zum einen die Cognitive Evaluation Theory von Deci & Ryan (1985) und zum anderen die Reziprozitätshypothese nach Gouldner (1960). Aus diesen Ansätzen werden für die späteren Analysen Kausalhypothesen abgeleitet. Vorher wird im fünften Kapitel das Studiendesign beschrieben und dann erste empirische Informationen über die Stichprobenszusammensetzung vorgestellt. Darauf folgen im sechsten Kapitel die Operationalisierungen der zugeordneten Messkonzepte, welche im siebten und achten Kapitel empirisch auf Anwendbarkeit geprüft werden. Im neunten Kapitel werden die abgeleiteten Hypothesen zur Wirkung von Incentives auf die Antwortqualität mithilfe von Strukturgleichungsmodellen geprüft und im zehnten Kapitel die Validität diskutiert. Das elfte Kapitel stellt das Fazit dieser Arbeit dar.

2. Das Konzept der Datenqualität

Zu Beginn der Studie wurde eine Literaturrecherche durchgeführt, mit dem Ziel, auf bestehende Definitionen zur Antwortqualität zurückzugreifen. Die Recherche ergab jedoch, dass im Großteil der Studien auf eine theoretische Herleitung und Erläuterung des genutzten Begriffes der Qualität verzichtet wird. In wenigen Studien, welche die Prüfung von Qualität systematisch angehen, wird zwar prinzipiell versucht einzelne Indikatoren theoretisch zu fundieren (vgl. Medway & Tourangeau (2015); Medway (2012)), eine Definition und theoretische Ableitung des Begriffs der Antwortqualität bleibt jedoch aus. Als theoretische Herleitungen werden üblicherweise bei der Beschreibung und Erklärung von Umfragequalität das Konzept des Total Survey Errors (Biemer & Lyberg (2003); Weisberg (2005)) und/oder das Satisficing-Konzept von Krosnick (1991) genutzt. Aus diesen beiden Konzepten soll eine Definition abgeleitet werden, welche noch um einen normativen Ansatz (aus dem Mikroenzusgesetz) erweitert wird.

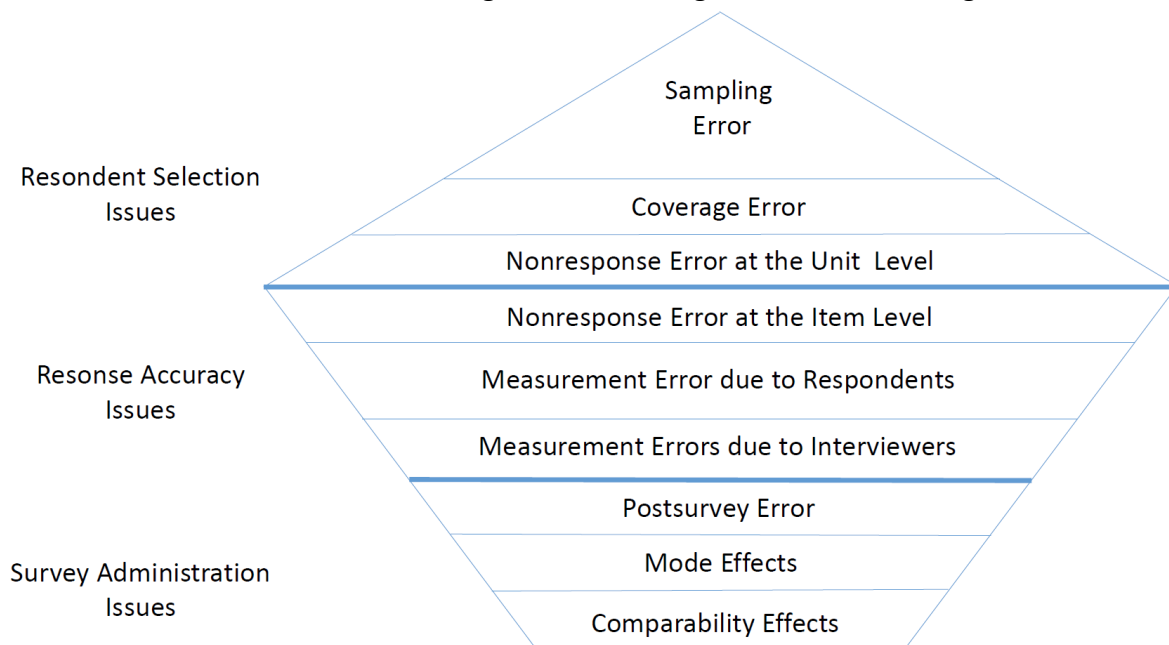
2.1 Der Total Survey Error

Die Idee des Total Survey Errors (dt. = „Totale Umfragefehler“) besagt, dass verschiedene Fehlerarten im Prozess der Stichprobenziehung, Datenerhebung und –auswertung sowie des Erhebungsdesigns einen Einfluss auf die Qualität von Daten haben: „The total survey error approach emphasizes the several possible sources of survey error, along with constraints that affect the minimization of those errors and various effects that are inherent to surveys“ (Weisberg 2005, S. 325). Das daraus folgende Ziel liegt darin, diese Fehlerquellen genau zu untersuchen und zu klassifizieren. Groves & Lyberg (2010) sehen die konzeptionellen Anfänge der Fehlerklassifikation für Umfragen in den 40er Jahren mit der Veröffentlichung von Deming (1944) in der *American Sociological Review*. In diesem Artikel stellt Deming 13 Faktoren vor²,

² Die 13 Faktoren können dem Anhang entnommen werden, S. 205.

mit deren Hilfe die Brauchbarkeit einer Umfrage eingeschätzt werden kann. Seitdem haben sich viele Autoren aus verschiedenen Disziplinen mit dem Thema beschäftigt, ohne sich jedoch auf ein einheitliches Konzept des Total Survey Errors zu einigen. Die aktuelle Situation lässt sich am besten so beschreiben, dass verschiedene Klassifikationskonzepte vorliegen, welche je nach Autor schwach bzw. stark variieren können (vgl. Költringer (1993); Biemer & Lyberg (2003); Weisberg (2005)).³ Eine der Gemeinsamkeiten liegt in der groben Aufteilung der Umfragefehler in Sampling und Nonsampling Errors. Der Sampling Error umfasst nach Biemer & Lyberg (2003) die statistischen Abweichungen, welche im Rahmen der Stichprobenziehung auftreten können. Weisbergs Klassifikation (2005) umfasst erweiternd auch den Coverage Error und den Nonresponse Error at the Unit Level. Diese werden von Biemer & Lyberg (2003) zwar auch bedacht, aber der Kategorie des Nonsampling Errors zugeordnet.

Abb. 2: Die Aufteilung der Fehlerkategorien nach Weisberg



Quelle: eigene Darstellung, nach Weisberg (2005), S. 19.

³ Aufgrund der ausführlichen Ausarbeitungen und Darstellungen wird im Folgenden nur auf die umfassenden Konzeptionen von Biemer & Lyberg (2003) und Weisberg (2005) eingegangen.

Unter den Nonsampling Errors summiert sich, je nach Konzeption, eine Vielzahl an Fehlerquellen, welche tendenziell unabhängig von der Stichprobenziehung sind. Biemer & Lyberg (2003) begründen die Trennung mit der Kontrollierbarkeit der Fehler: „(...) nonsampling errors can be viewed as mistakes or unintentional errors that can be made at any stage of the survey process. Despite our best efforts to avoid them, nonsampling errors are inevitable particularly in large-scale data collections. Sampling errors, on the other hand, are intentional errors in the sense that we can control their magnitude by adjusting the size of the sample. With a sampling size of 1, sampling error is at its maximum, and as we increase the sample size to the population size (...), sampling error becomes smaller and smaller“ (Biemer & Lyberg (2003), S. 37). Die bisher ausführlichste Kategorisierung von *Nonsampling Errors* ist in dem Werk von Biemer & Lyberg (2003) zu finden. Die von ihnen klassifizierten verschiedenen Fehlerquellen werden daher folgend tabellarisch aufgezeigt und erläutert:

Tab. 1: Die Komponenten des Nonsampling Error nach Biemer & Lyberg

<u>Sources of Error</u>	-	<u>Types of Error</u>
<i>Specification Error</i>		Concepts Objectives Data elements
<i>Frame Error</i>		Omissions Erroneous inclusions Duplications
<i>Nonresponse Error</i>		Whole unit Within unit Item Incomplete Information
<i>Measurement Error</i>		Information system Setting Mode of data collection Respondent Interview Instrument
<i>Processing Error</i>		Editing Data entry Coding Weighting Tabulation

Quelle: eigene Darstellung, nach Biemer & Lyberg ((2003), S. 39.

Unter *Specification Error* verstehen Biemer & Lyberg (2003) Fehlerquellen, welche aufgrund von Unstimmigkeiten bei der Konzeption einer Befragung auftreten können: „Specification error occurs when the concept implied by the survey question and the concept that should be measured in the survey differ. This occurs, the wrong parameter is being estimated in the survey, and thus inference based on the estimate may be erroneous. Specification error is often caused by poor communication between the researchers, data analyst, or survey sponsor and the questionnaire designer“ (Biemer & Lyberg (2003), S. 38). Der *Frame Error* hingegen umfasst die Fehler, die bei der Konstruktion eines Stichprobenplans auftreten können: „There are a number of errors that can occur when the frame is constructed. Population elements may be omitted or duplicated an unknown number of times. There may be elements on the frame that should not be included (e.g., businesses that are not farm in a farm survey)“ (Biemer & Lyberg (2003), S. 40 f.). Die Kategorie *Nonresponse Error* beschäftigt sich mit Ausfällen im Rahmen der Erhebung bzw. der Befragung. Biemer und Lyberg (2003) unterscheiden hier zwischen unit nonresponse und item nonresponse: „A *unit nonresponse* occurs when a sampling unit (household, farm, establishment, etc.) does not respond to any part of the questionnaire. (...) *Item nonresponse* occurs when the questionnaire is only partially completed (i.e., some items are skipped or left blank that should have been answered)“ (Biemer & Lyberg (2003), S. 41). Die vierte Kategorie ist der *Measurement Error*: „The key components of measurement error are the respondent, the interviewer, and the survey questionnaire. Respondents may either deliberately or unintentionally provide incorrect information. Interviewers can cause errors in a number of ways. They may falsify data, inappropriately influence responses, record responses incorrectly, or otherwise fail to comply with the survey procedures. The questionnaire can be a major source of error if it is poorly designed. Ambiguous questions,

confusing instructions, and easily misunderstood terms are examples of questionnaire problems that can lead to measurement error“ (Biemer & Lyberg (2003), S. 41). Der *Processing Error* umfasst alle Fehler, welche im Rahmen der Datenbearbeitung auftreten: „(...) errors that arise during the data processing stage, including errors in the editing of data, data entry, coding, the assignment of survey weights, and the tabulation of survey data.“ (Biemer & Lyberg (2003), S. 43).

Für eine Definition der Antwortqualität erscheinen prinzipiell der oben dargestellte Non-response Error und der Measurement Error (auch: Messfehler) als relevant, da diese auf das Befragtenverhalten zurückzuführen sind. Die allgemeine Bezeichnung als *Measurement Error* ist in den verschiedenen Konzepten des Total Survey Error jedoch nicht einheitlich definiert. Nach Weisberg (2005) kann Measurement Error wie folgt beschrieben werden: „Respondent-related error occurs to the extent that the respondents are not providing the answers they should, given the researcher’s intentions“ (Weisberg (2005), S. 72). Biemer et al. (1991) hingegen sehen nicht die Abweichung von der Intention des Fragebogenentwicklers als relevante Abweichung, sondern folgen primär dem Konzept des „wahren“ Wertes. Sie sprechen hierbei von *Observational Errors*: „*Observational errors* are deviations of the answers of respondents from their true values on the measure (...), these are *measurement errors*“ (Biemer et al. (1991), S. 2). Diese Messfehler werden dabei nicht als unabhängig vom Messinstrument angesehen: „There are also effects on the quality of respondents’ answers from the wording of the question or flow of the questionnaire, which are labeled *instrument error*“ (Biemer et al. (1991), S. 3).

2.1.1 Der „wahre“ Wert

Eine häufig getroffene Grundannahme der verschiedenen Konzepte zum Total Survey Error ist die Existenz eines „wahren“ Wertes. Der Begriff wird hierbei aus der klassischen Testtheorie übernommen, und kann wie folgt definiert werden: „Der **wahre Wert (Tau, τ)** einer **Person (v)** ist als **Mittelwert** über unendlich oft **wiederholte unabhängige Messungen (t)** der **beobachteten Werte (x)** einer **Person (v)** definiert. Es handelt sich folglich um den **Erwartungswert E (X)** der intraindividuellen Verteilung der beobachteten Werte x einer Person“ (Bühner (2011), S. 43; Hervorhebungen im Original). Die Umsetzung der Idee des „wahren Wertes“ ist in empirischen Umfragen jedoch nicht gerade trivial. „Für die Bestimmung von wahren Werten in Persönlichkeits- und Leistungsvariablen sind Wiederholungen des Messvorgangs mit demselben Messinstrument aber problematisch. Es könnten Erinnerungseinflüsse auftreten und die in den oben aufgeführten Axiomen geforderte Zufälligkeit der Fehlergrößen verletzen. Eine wiederholte Anwendung desselben Messinstruments scheidet somit aus“ (Moosbrugger & Kelava (2012), S. 106). Darüber hinaus wird der „wahre“ Wert, wie in den Konzeptionen des Total Survey Error, zumeist nicht ohne Fehler gemessen. Hieraus folgt die Idee des Messfehlers, welche auch im Total Survey Error angewendet wird: „Der **Messfehler (Epsilon, ϵ)** einer **Person (v)** zu einem **Zeitpunkt (t)** setzt sich aus der Differenz zwischen **beobachtetem Messwert (x)** zum Zeitpunkt t und konstantem, über Zeitpunkte hinweg nicht variierenden, **wahren Wert (Tau, τ)** zusammen. Der Messfehler repräsentiert dabei alle unkontrollierten und unsystematischen Störeinflüsse bei der Messung“ (Bühner (2011), S. 46; Hervorhebungen im Original). Diese sollen sich „herausmitteln“ und werden dadurch kontrollierbar. Problematischer ist der systematische Messfehler: „Das heißt **systematische Messfehler verzerren den wahren Wert** entweder nach oben oder unten, können vom wahren Wert jedoch nicht ohne weiteres getrennt werden“ (Bühner, S. 48, Hervorhebungen im Original). Als Beispiel für einen

systematischen Messfehler kann die soziale Erwünschtheit genannt werden (vgl. Esser (1991), Hartmann (1991), Lischewski (2015)).

Bei Übertragung des Konzepts auf die Umfrageforschung stellt sich die Frage, inwiefern die Idee eines raum-zeitlich unabhängigen „wahren“ Wertes in der empirischen Umfrageforschung aufrecht zu erhalten ist, da hierfür zeitlich stabile Einstellungen und Bewertungen unterstellt werden müssen. Es soll daher die Anwendbarkeit für die Umfrageforschung kurz diskutiert werden.

2.1.2 Die Umsetzbarkeit des Konzepts des „wahren“-Werts in der Umfrageforschung

In der Einstellungsforschung gibt es einige Befunde, welche die Existenz eines „wahren“ Wertes zweifelhaft erscheinen lassen. Hierfür wurden von verschiedenen Autoren (z.B. Converse (1964); Zaller & Feldman (1992)) widersprechende Befunde aufgezeigt bzw. alternative Konzepte entwickelt, die im Folgenden kurz vorgestellt werden:

Converse (1964) zeigte in seiner Studie auf, dass nur ein geringer Anteil der Bevölkerung eine gefestigte politische Einstellung hat: „The substantive conclusion imposed by these technical maneuvers is simply that large portions of the electorate do not have meaningful beliefs, even on issues that have formed the basis for intense political controversy among elites for substantial periods of time“ (Converse (1964), S. 245). Mit diesem Schluss schränkt Converse den Anwendungsbereich des Konzepts des „wahren“ Wertes stark ein. Es gibt zwar stabile politische Einstellungen, welche aber nur bei wenigen Personen vorhanden sind. Aus dieser Aufteilung heraus entwickelte Converse (1964) die „Black-White-These“. Achen (1975) kam, abgrenzend zu Converse, zu dem Schluss, dass es gar nicht so sehr die mangelnde Stabilität der Antwort sei, welche für die geringen Zusammenhänge im Zeitverlauf verantwortlich ist, sondern vielmehr die (mangelhafte) Frageformulierung. Dabei hebt er hervor: „Measurement error is

primarily a fault of the instruments, not of the respondents.” (Achen (1975, S. 1229). Unter Kontrolle des zufälligen Messfehlers konnte er in den Daten von Converse doch stabile politische Einstellungen finden.

Zaller & Feldman (1992) brachten eine alternative Sichtweise in die Diskussion ein: „Most citizens, we argue, simply do not possess preformed attitudes at the level of specificity demanded in surveys. Rather, they carry around in their heads a mix of only partially consistent ideas and considerations. When questioned, they call to mind a sample of these ideas, including an oversample of ideas made salient by the questionnaire and other recent events, and use them to choose among the options offered. But their choices do not, in most cases, reflect anything that can be describes as true attitudes; rather, they reflect the thoughts that are most accessible in memory at the moment of response” (Zaller & Feldman (1992), S. 580). Es wird damit aufgezeigt, dass eine Einstellung nicht zwingend als stabiler Wert zu betrachten ist, sondern auch eine situationale Kontextabhängigkeit aufweisen kann. Weitergehend schreiben sie: „Our claim is that even when people exhibit high levels of response instability, the opinions they express may still be based on real considerations. Even when these considerations turn out to be transitory, the opinion statements they generate are not, for that reason, necessarily lacking in authenticity” (Zaller & Feldman (1992), S. 612). Die Idee findet sich auch formalisiert im Framingansatz nach Esser (vgl. Esser (1990; 1999)), welcher die Definition der Situation explizit berücksichtigt. Aus diesem Ansatz heraus formuliert er kritisch: „Da der Befragte bei der Entscheidung zur Antwort jeweils immer alle Situationsmerkmale als „Problem“ wahrnimmt, und da die Konzentration auf den „wahren Wert“ eine vom Sozialforscher aus zu seinen Zwecken bewertete externe Vorgabe ist, dann kann man in der Tat davon spre-

chen, daß die Annahmen der klassischen Testtheorie und die darauf aufbauenden Implikationen der Methodologie der Umfrageforschung ein sehr einseitiges Bild des Befragten gezeichnet haben“ (Esser (1986), S. 333).

Der letzten Argumentation folgend kann bei Faktfragen noch am ehesten von einem „wahren“ Wert ausgegangen werden und dann ist eine unverzerrte Messung nur mit enormen Aufwand zu erreichen. Zur Prüfung des „wahren“ Wertes werden sog. Validierungsfragen herangezogen, das heißt es werden den Befragten Faktfragen gestellt, wobei der Wahrheitsgehalt der Antworten von den Forschern anhand von Datenbeständen überprüft werden kann. Hierbei ist jedoch auf das Problem von Erinnerungseffekten zu verweisen. Diese können dazu führen, dass eine falsche Angabe gemacht wird, welche aber, z.B. aufgrund einer fehlerhaften Erinnerung oder Verdrängung, als wahr geglaubt wird. Liegen bewusste Erinnerungslücken vor, so kann es für die Befragten sehr hilfreich sein, ihre Antworten mit Hilfe von Wissensspeichern (z.B. Dokumente, Tagebücher) zu fundieren. Schwerwiegender erscheinen jedoch falsche Erinnerungen, welche auf einer „Selbsttäuschung“ basieren. Als Beispiel soll eine Person skizziert werden, welche die eigene Teilhabe an den Verbrechen des zweiten Weltkrieges verleugnet bzw. verdrängt. Diese „Selbsttäuschung“ kann hierbei ein persönlichkeitschützender Mechanismus sein, welcher unbewusst gegenwärtige Verhaltensweisen und Einstellungen bedingt (vgl. Rosenthal (1999)). Hier liegt es nun an den Forschern zu entscheiden, ob eine ggf. richtig geglaubte Falschantwort, auf Grundlage einer Selbstlüge, noch „wahr“ ist.⁴ Es ist hervorzuheben, dass es für den Forscher bei der Messung von Einstellungen (wenn überhaupt) nur schwer ersichtlich ist, ob eine Falschantwort vorliegt und wenn ja, inwiefern diese von der

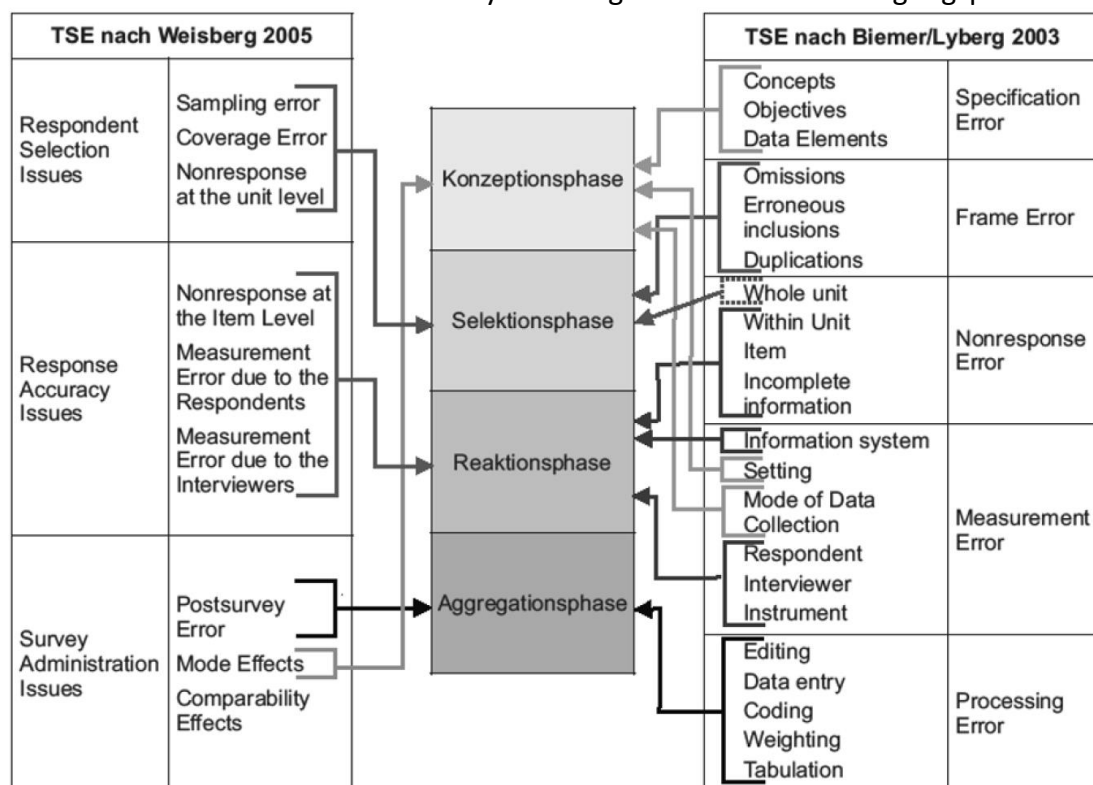
⁴ Hier stellt sich die Frage, für welche „wahren“ Informationen sich die Forscher letztlich interessieren. Ist der Forschungsgegenstand die aktuelle Lebenswelt, inklusive der subjektiven Wahrnehmungen oder objektive Sachverhalte, welche ggf. keine direkte Relevanz (mehr) haben.

Persönlichkeit als „wahr“ definiert wird. Die Idee des „wahren“ Wertes scheint damit für die Umfrageforschung an starke Grundannahmen gekoppelt zu sein, welche nicht pauschal akzeptiert werden können.

2.1.3 Nutzbarkeit des Total Survey Errors zur Definition von Qualität

Bachleitner et al. (2010) versuchen die prominenten Ansätze von Weisberg (2005) und Biemer & Lyberg (2003) in vier Phasen zusammenzuführen und damit ein einheitliches Konzept zum Total Survey Error zu etablieren:

Abb. 3: Elemente des Total Survey Error zugeordnet in vier Befragungsphasen



Quelle: Bachleitner et al. (2010), S. 156.

Durch die oben dargestellten Phasen und Klassifikationsvorschläge wird deutlich, dass Qualität von Umfragen in verschiedene Bereiche aufgliedert werden kann und nicht allgemein von der Datenqualität gesprochen werden kann. Aufgrund dessen ist eine genaue Definition

des unterstellten Qualitätsbegriffs sehr wichtig, da sich daraus verschiedene Indikatoren für die Messung der Qualität ableiten lassen. Das Konzept des Total Survey Errors leistet selbst keine eigene inhaltliche Definition des Begriffes „Qualität“. Es kann nicht einmal abgeleitet werden, dass eine Qualitätssteigerung die Abwesenheit von Fehlerarten bedeutet, da beispielhaft eine Erhöhung der Responsequote zu einer Erhöhung des Measurement Errors führen kann (vgl. Groves & Lyberg (2010), S. 871 f.).

Nach der Vorstellung der Grundidee des Total Survey Error wird nun der für diese Studie angesprochene Qualitätsbereich benannt und aufgegliedert. Hierfür muss die Forschungsfrage kurz wiederholend zusammengefasst werden: Welchen Einfluss weist eine Incentivierung auf die Antwortqualität während der Bearbeitung eines Fragebogens auf? Folglich ist die relevante Fehlerkategorie im Bereich des Nonsampling Error zu finden. Es wird hierbei unterstellt, dass die Teilnahme an einer Befragung prinzipiell ein (mehr oder weniger) bewusster Akt ist und damit, wenn auch nicht komplett, im Rahmen kognitiver Steuerungsmechanismen liegt. Da die Klassifikationen nach Weisberg (2005) und Biemer & Lyberg (2003) aktuell am umfangreichsten sind, werden diese herangezogen und über die Zusammenfassung von Bachleitner et al. (2010) genutzt. Der in der Forschungsfrage angesprochene Qualitätsbereich findet sich in der Reaktionsphase wieder und umfasst damit das Verhalten der Befragten während der Teilnahme an einer Umfrage. Diese Phase schließt daher den Beantwortungsprozess und die daraus resultierende Antwortqualität mit ein.

2.2 Satisficing und Optimizing

In der empirischen Umfrageforschung wird u.a. das Konzept des Satisficing⁵ genutzt, um Antwortverhalten zu erklären. Die Grundannahme ist hierbei, dass zur Antwortfindung die Teilnehmer mental aufwändige Prozesse durchlaufen müssen. Wird von den Befragten im Rahmen der Beantwortung ein hoher mentaler Aufwand geleistet, so wird dann üblicherweise von Optimizing gesprochen, d.h. die Befragten haben die Antwort vollständig durchdacht. Die permanente Aufrechterhaltung eines hohen kognitiven Anstrengungsgrades wird jedoch als unwahrscheinlich angesehen und führt zum Gegenstück des Optimizing: dem Satisficing. Krosnick (1991) definiert Satisficing wie folgt: „Rather than continuing to expend the mental effort necessary to generate optimal answers to question after question, respondents are likely to compromise their standards and expend less energy instead“ (Krosnick (1991), S. 215). Die Konsequenzen sind hierbei vielfältig: „It may involve selecting the first reasonable response, thus avoiding the need to read or listen to the rest of the list. It may involve simply agreeing with assertions. It may manifest itself in the form of a lack of differentiation in rating questions (i.e. the respondent gives the same answer to each item in a list) or a tendency to respond “don’t know” or responses that are the result of the mental equivalent of coin flipping“ (Krosnick (2000), S. 5). Krosnick entwickelte, unter Zuhilfenahme des psychologischen Konzeptes zur Erklärung des Antwortprozesses (nach Tourangeau & Rasinski (1988))⁶ eine in den Sozialwissenschaften oft genutzte Grundlage zur Erklärung von Antwortverhalten in Umfragen (vgl.

⁵ Der Begriff des Satisficing geht zurück auf Herbert Simon (1947/1957), welcher im Rahmen wirtschaftswissenschaftlicher Managementplanung den Versuch unternahm die Arbeitsleistung mithilfe des Rational Choice Ansatzes zu erklären.

⁶ Die erste Darstellung des kognitiven Antwortprozesses findet sich bei Tourangeau (1984), wobei diese noch nicht so differenziert und ausgearbeitet ist. Daher wird im Folgenden auf den gemeinsamen Artikel von Tourangeau & Rasinski (1988) verwiesen.

Krosnick & Alwin (1987)).⁷ Nach dem Modell von Tourangeau & Rasinski (1988) kann der Antwortprozess in vier Phasen untergliedert werden: 1) Comprehension, 2) Retrieval, 3) Judgement und 4) Response. Comprehension umfasst alles, was unter dem Verständnis einer gestellten Frage subsumiert werden kann (z.B. unbekannte oder mehrdeutige Worte, Satzkonstruktion). Nach dem Comprehension folgt das Retrieval. Hierbei wird der Rückgriff auf memorisierte Wissensbestände verstanden, welche je nach Stimulus der Frage benötigt werden. In der dritten Phase folgt mit dem Judgement die Bewertung und Abwägung der zur Antwortfindung benötigten Informationen. In der letzten Phase, der Response, wird die Antwort in das Antwortformat des Fragebogens übertragen und damit für den Forscher dokumentiert. Werden alle vier Phasen des Antwortprozesses ordnungsgemäß durchlaufen, so liegt nach Krosnick (1991) Optimizing vor. Bei einer Abweichung vom idealtypischen Antwortprozess (z.B. durch Ankreuzen beliebiger Antwortkategorien unabhängig vom Inhalt) liegt Satisficing vor.⁸

Nach Krosnick (1991) ist ein Auftreten von Satisficing von drei Komponenten abhängig: der Aufgabenschwierigkeit, der Fähigkeit zur Aufgabenbewältigung und der Motivation während der Bearbeitung. Diese drei Komponenten werden von ihm in einer Gleichung zusammengefasst, wobei das Ergebnis als die Wahrscheinlichkeit für Satisficing definiert⁹ wird:

$$P(\text{Satisficing}) = \frac{a_1(\text{Task Difficulty})}{a_2(\text{Ability}) \times a_3(\text{Motivation})} \cdot^{10}$$

⁷ Die theoretische Verbindung der beiden Konzepte wurde durch die CASM-Forschung (Cognitive Aspects of Survey Methodology) vorbereitet und gestützt. Für vertiefende Informationen siehe Jabine et al. (1984).

⁸ In dieser Verknüpfung der Idee des Satisficing mit dem Modell des kognitiven Antwortprozesses liegt der Mehrwert gegenüber der Konzeption von Zaller & Feldman (1992). Dies liegt darin begründet, dass bei Zaller & Feldman der Prozess der Antwortgenerierung nicht erläutert wird und folglich keine Erklärungen über ein Antwortverhalten abgeleitet werden können.

⁹ Für die Darstellung der Gleichung siehe Krosnick (1991; 2000).

¹⁰ In einem Gespräch mit Krosnick (24.08.2015) wurde deutlich, dass diese Gleichung nicht empirisch abgesichert ist und er ihr aus diesem Grund noch kritisch gegenübersteht.

Diese drei Komponenten können wiederum in Kernbestandteile aufgegliedert werden. So besteht die Aufgabenschwierigkeit aus mehreren Facetten, welche auf die verschiedenen Aspekte eines Beantwortens von Fragen wirken können:

„Interpretation: The difficulty of interpreting the question will be affected by the number of words in the question, the familiarity of the words used and the extent to which any of the words may have multiple possible meanings.

Retrieval: It is more difficult to retrieve information relating to previous states rather than the current state. It is more difficult to retrieve information relating to multiple objects (“How many times did you do any of X, Y or Z?”, as opposed to “How many times did you do X?”) or multiple evaluative dimensions (“Rate each product for quality and size”).

Judgement: Absolute judgements can be less demanding and subjective than relative ones. The task forming a judgement can be easier if it can be decomposed into stages.

Response selection: Reporting the judgement involves selecting a response. This is generally easier if response categories have verbal rather than numeric labels, and if the words used in the labels are familiar and unambiguous.

Interviewer pace: In the case of interviewer-administered surveys, the difficulty of the respondent’s task can be also be affected by the speed at which the interviewer asks the questions and the time allowed for answers.

Distraction: Socio-environmental factors, largely outside the control of the researcher, can also impact upon task difficulty. In particular, responding can be more difficult if the respondent is distracted, for example by the presence of other people, or by voices or noise” (Krosnick (2000), S. 6f.).

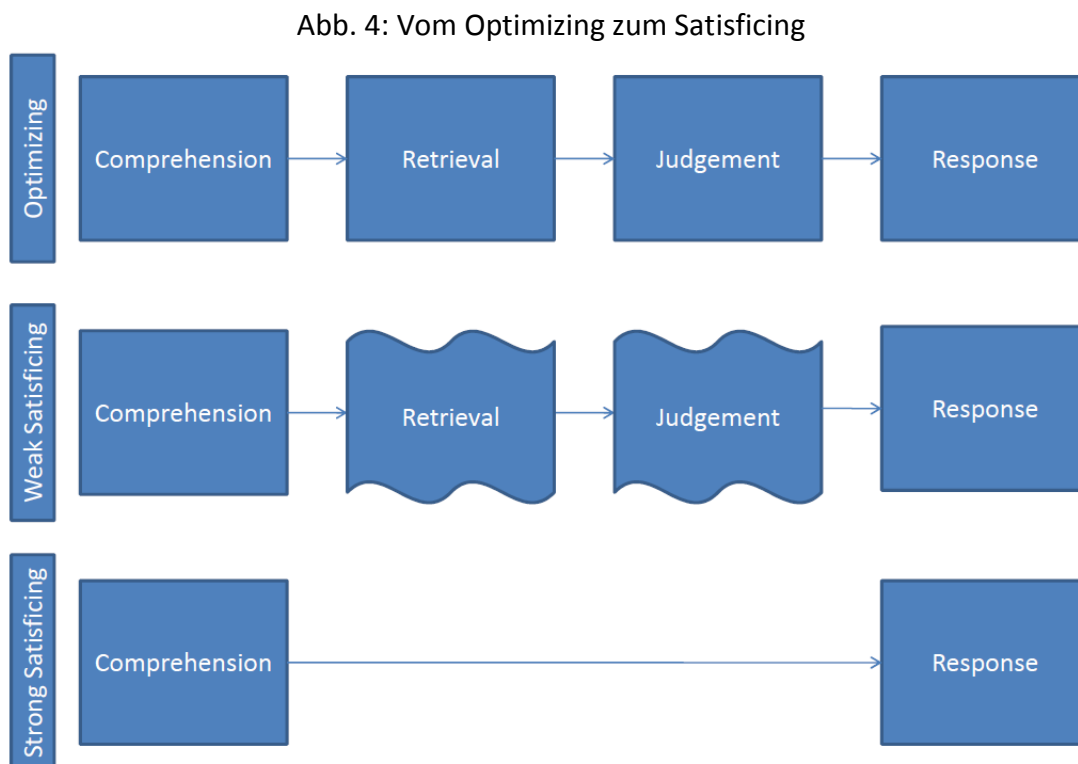
Die Fähigkeit der Befragten zur Bewältigung von Aufgaben lässt sich nach Krosnick wie folgt erfassen: „The ability of a person to perform the role of survey respondent adequately will depend upon their cognitive skills. It will also be affected by the extent to which he or she has

previously thought about the topic of the question and the extent to which he or she has a relevant pre-consolidated judgement stored in memory" (Krosnick (2000), S. 7). Zuletzt wird auch die Motivation in Bestandteile aufgegliedert: „Many factors can affect the motivation of a respondent. These include the need for cognition, accountability, the importance of the topic to the respondent personally, the respondent's belief about the overall importance of the survey, the behavior of the interviewer, the number of prior questions, and so on" (Krosnick (2000), S. 7).

Da die Aufgabenschwierigkeit sowie die Motivation mit jeder Frage wechseln können, ist die Wahrscheinlichkeit für ein Satisficing abhängig vom jeweils gestellten Item. Die Berechnung des Wahrscheinlichkeitswertes ist dabei nicht unkritisch zu sehen: Nach vertiefter Durchsicht der Literatur zum Konzept des Satisficing bleibt unklar, aus welchen theoretischen Konzepten die oben beschriebenen Komponentenbestandteile abgeleitet werden. Es überrascht daher nicht, dass für den Aufbau der oben dargestellten Wahrscheinlichkeitsgleichung keine begründete Erklärung gegeben ist. Sieht man einmal davon ab, ist auch festzustellen, dass keine Informationen bezüglich des benötigten Schätzverfahrens zur Berechnung der Parameter (a_1 , a_2 und a_3) gegeben werden. Aus den genannten Gründen erscheint es nicht verwunderlich, dass der oben dargestellte Wahrscheinlichkeitsindex in empirischen Studien nur sehr selten herangezogen wird (Holbrook et al. (2003)) und Satisficing stattdessen entweder als summativer Index über die Häufigkeit des Auftretens der einzelnen Negativfolgen berechnet wird (Medway & Tourangeau (2015); Barge & Gehlbach (2012)), oder die Negativfolgen (z.B. Akquieszenzmessungen) selbst direkt in die Analysen eingehen und damit stellvertretend für Satisficing stehen (Medway (2012); Krosnick et al. (1996)).

2.2.1 Weak Satisficing und Strong Satisficing

Beim Satisficing werden von Krosnick (1991) zwei Stufen unterschieden: „weak“ und „strong“ Satisficing. Bei einem „weak“ Satisficing wird die zweite und dritte Phase des kognitiven Antwortprozesses nur unvollständig oder verzerrt durchschritten. Die gegebene Antwort kann sich folglich von einer Antwort, welche bei einem vollständigen Durchlaufen des kognitiven Antwortprozesses gewählt worden wäre, unterscheiden. Bei einem „strong“ Satisficing wird die zweite und dritte Phase komplett übersprungen und aus aktuellen kognitiven Verankerungen eine Spontanantwort generiert (vgl. Krosnick (2000), S. 6).¹¹ Zur besseren Verdeutlichung werden die Phasen des kognitiven Antwortprozesses (nach Tourangeau & Rasinski (1988)) graphisch dargestellt, wobei die oben dargestellten Anomalien im Ablauf (aufgrund von Satisficing) berücksichtigt werden.



Quelle: eigene Darstellung, nach Krosnick (2000), S. 6.

¹¹ Es wird hiermit noch erwähnt, dass die Nutzung einer Satisficing-Strategie keine bewusste Entscheidung sein muss. Silber (2015) verweist z.B. auf die Möglichkeit eines kulturell bedingten Satisficing-Frames.

Ein schwaches und ein starkes Satisficing kann, gemäß Krosnick (1991) über verschiedene Verhaltensbefunde im Antwortprozess festgestellt werden¹²:

Tab. 2: Beispiele für Antwortverhalten bei Weak und Strong Satisficing

Weak Satisficing	Strong Satisficing
Primacy und Recency	Status Quo-Effekt
Akquieszenz	Nicht-Differenzierung in Antwortskalen
	Wahl von "Weiß nicht" / "Keine Angabe"
	Mental Coin Flip

Quelle: eigene Darstellung, nach Krosnick (1991), S. 215.

2.2.1.1 Weak Satisficing

a) Primacy und Recency

Bei einem Primacy Effekt wird angenommen, dass die Teilnehmer einer schriftlichen Befragung bei einer Liste von Antwortoptionen die Antworten wählen, die weiter oben in der Antwortskala aufgeführt sind. Dies liegt zum einen im Aufwand der Abwägung aller Antwortmöglichkeiten begründet und zum anderen in der Zufriedenheit mit der erstbesten passenden Antwort. „Thus, weak satisficing seems likely to produce primacy effects under conditions of visual presentation“ (Krosnick (1991), S. 216). Bei mündlichen Befragungen kann bei einem schwachen Satisficing gleichermaßen ein Primacy oder Recency Effekt erwartet werden. Ein Recency Effekt bedeutet, dass die unteren bzw. letztgenannten Antwortkategorien bevorzugt gewählt werden, da diese besser im Gedächtnis verankert werden: „(...) respondents are able to devote the most processing time to the final items read; these items remain in short-term

¹² In der Darstellung von Krosnick (1991) scheint das weak und strong Satisficing in einem kategorialen Verhältnis zu stehen. Dies begründet sich daraus, dass Satisficing Item-abhängig ist und damit nur einer der beiden Kategorien zugeordnet werden kann. Demgemäß kann die befragte Person entweder ein Optimizing, oder ein weak, bzw. strong Satisficing im Verhalten aufweisen.

memory after interviewers pause to let respondents answer“ (Krosnick (1991), S. 217). Auch hierbei wird der kognitive Antwortprozess nur unvollständig oder verzerrt durchschritten. Es soll hierbei aber darauf hingewiesen werden, dass nicht jede Zustimmung der ersten oder letzten Antwortkategorien eine Folge eines Primacy- oder Recency-Effekts sein muss. Krosnick (1991) argumentiert hierbei, dass dies allerdings aufgrund des hohen mentalen Aufwands wahrscheinlich ist: „However, it is conceivable that some respondents listen to a list of response alternatives without evaluating any of them. Once the list is read, these individuals may begin their thinking by recalling the first alternative and thinking about that one. Then, they may progress through the list, one by one, from beginning to end. Given that fatigue should investigate weak satisficing relatively quickly, a primacy effect would be expected“ (Krosnick (1991), S. 217).

b) Akquieszenz

Generell entwickelte sich der Begriff der Akquieszenz in der Umfrageforschung aus der Feststellung, dass einige Befragte eine Neigung zu einem deutlich vermehrten Zustimmungsverhalten bei Einstellungsitems aufweisen (vgl. Cronbach, (1942; 1946)). Zur Erklärung dieses Phänomens wurden viele verschiedene Definitions- und Erklärungsansätze entwickelt. So kann z.B. unter Akquieszenz eine inhaltsunabhängige Zustimmungstendenz verstanden werden. Mit Blick auf Satisficing kann Akquieszenz je nach Erklärungsansatz dem weak oder strong Satisficing zugeordnet werden¹³:

¹³ Es gibt neben den zwei folgenden Einflüssen von Satisficing-Strategien noch einen weiteren Ansatz zur Erklärung von Akquieszenz. Dieser basiert auf der Annahme das kulturelle Einflüsse und/oder psychologische Dispositionen eine Zustimmungstendenz fördern: „One explanation for acquiescence response bias occurs partly due to social norms to be polite. Consistent with this, acquiescence response bias is stronger in among cultures that put a high value on politeness and deference.“ (Holbrook (2008a), S. 3)

1) Der erste Erklärungsansatz bezieht sich auf ein weak Satisficing:

Es wird hierbei unterstellt, das Befragte bei Zustimmungsfragen üblicherweise darüber nachdenken, ob eine Zustimmung zu einem Sachverhalt angemessen ist. „If respondents fail to generate any such reasons they would presumably say ‘disagree’, and if respondents succeed in generating enough such statements they would presumably say ‘agree’. Because most assertions offered in survey questions are probably reasonable, it seems extremely likely that many respondents using this decision rule will succeed in generating enough reasons to justify saying ‘agree’ most of the time“ (Krosnick (1991), S. 218).

2) Ein zweiter Erklärungsansatz bezieht sich auf ein strong Satisficing:

Dieser Erklärungsansatz basiert auf der Annahme einer Unterwürfigkeitshaltung der Befragten gegenüber der forschenden Person. Eine Zustimmung erfolgt dann dadurch, dass von den Befragten unterstellt wird, dass Forscher nur sinnvolle und korrekte Angaben, bzw. Abfragen vorgeben und damit eine Zustimmung ebenfalls sinnvoll und gewünscht ist (vgl. Lenski & Legett (1960)). Dies tritt vor allem auf, wenn die Ausgestaltung des Erhebungsinstruments die Zielgruppe überfordert: „[...] acquiescence is particularly problematic when the domain of content to be measured is abstract, ambiguous, or unfamiliar to respondents“ (Armer & Baldigo (1973), S. 186). Hier liegt folglich ein starkes Satisficing vor, da die Phasen des Retrieval und des Judgement im kognitiven Antwortprozess komplett übersprungen werden.

2.2.1.2 Strong Satisficing

a) Status Quo-Effekt

In Befragungen werden des öfteren Fragen zu aktuellen politischen oder gesellschaftlichen Veränderungen gestellt. Werden die Fragen so formuliert, dass sich die Befragten zwischen einem Status Quo (z.B.: Es ist alles genau richtig wie es gerade ist) und einer Veränderung entscheiden müssen (z.B.: In unserer Gesellschaft muss mehr für die Rechte der Frauen getan werden), wird erwartet, dass bei einem starken Satisficing die Wahl zugunsten des Status Quo ausfällt: „In response to these sorts of questions, the easiest answer to give on the basis of little thought is ‘keep things as they are’“ (Krosnick (1991), S. 218). Aber nicht jede Wahl des Status Quo bedeutet auch ein Satisficing: „Some of these individuals probably arrive at this response after executing an effortful cognitive process that constitutes optimizing. However, many of them may give this answer instead without any retrieval or judgement, simply because it appears to be a reasonable answer“ (Krosnick (1991), S. 219).

b) Nicht-Differenzierung in Antwortskalen

In Item-Batterien, bei stets gleicher Polung der Antwortkategorien wird erwartet, dass die Befragten als Folge von Satisficing weniger in den Antworten variieren: „Doing so may sometimes be the result of a careful consideration of the merits of the objects, but this response strategy could also be the result of strong satisficing. Satisficing respondents could, for example, simply select a point on the response scale that appears to be reasonable for the first object, and then rate all the remaining objects at that point“ (Krosnick (1991), S. 219). An dieser Stelle soll erneut darauf hingewiesen werden, dass das oben beschriebene Antwortverhalten ein Indiz für Satisficing darstellt und der Verhaltensbefund einer Nicht-Differenzierung in Rating Skalen auch das Ergebnis bei Optimizing sein kann.

c) Wahl von „Weiß nicht“ / „Keine Angabe“

Die Wahl der Antwortkategorie „Weiß nicht“ oder „Keine Angabe“ kann nach Krosnick (1991) ebenfalls ein Resultat des strong Satisficing sein: „Regardless of the format of a question, respondents can always provide an answer that appears reasonable by telling the interviewer that they ‘don’t know’ what their opinion is. Doing so requires not retrieval or judgement, so it would constitute a form of strong satisficing“ (Krosnick (1991), S. 219). Jedoch muss erwähnt werden, dass die Wahl von „Weiß nicht“ oder „Keine Angabe“ nicht zwangsläufig ein Resultat von Satisficing bedeuten muss. Dies liegt darin begründet, dass die beiden Kategorien auch eine inhaltliche Komponente aufweisen und daher aufgrund tatsächlicher Unwissenheit oder inhaltlich begründeter Abwägungsprozesse gewählt werden können.

d) Mental Coin Flip

Der Mental Coin Flip zählt ebenfalls zum strong Satisficing. „That is, these respondents may simply choose randomly from among the response alternatives offered by a closed ended question“ (Krosnick (1991), S. 220). Hierbei muss jedoch angemerkt werden, dass der Begriff „randomly“ einen, dem Satisficing widersprechenden starken mentalen Aufwand für die Befragten impliziert. Dies begründet sich darin, dass der Befragte bei der Beantwortung von Fragen mental einen Pseudo-Zufallsprozess generieren muss und erst daraufhin Antwortkategorien wählen kann. Es erscheint damit fragwürdig, ob ein solcher mentaler Aufwand für die Befragten wirklich geringer ist als ein vollständiges Durchlaufen des kognitiven Antwortprozesses.

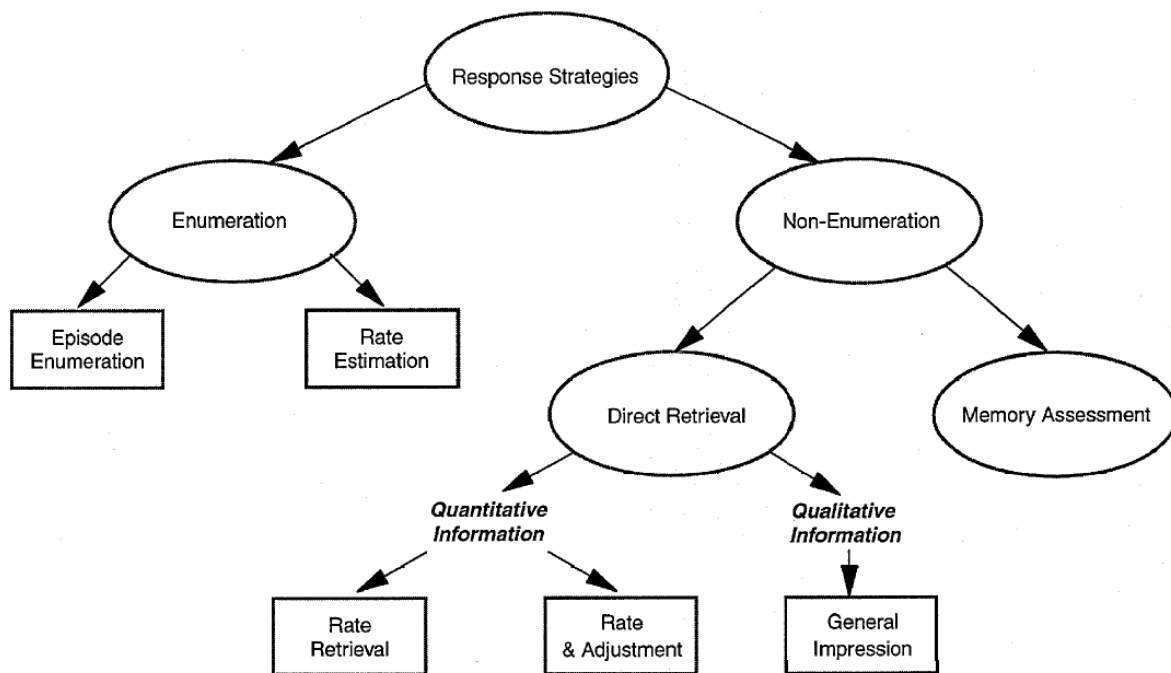
Die obige Auflistung von Konsequenzen eines Satisficing (nach Krosnick (1991) ist hierbei nicht abschließend zu betrachten. So kann beispielhaft auch ein verkürzter Erinnerungsprozess als Resultat von Satisficing verstanden werden. Dies soll folgend mithilfe der Darstellung verschiedener Erinnerungsstrategien für die Beantwortung von Häufigkeitsfragen verdeutlicht werden.

2.2.2 Erinnerungsstrategien bei Häufigkeitsfragen

Aus der Konzeption von Krosnick (1991) kann abgeleitet werden, dass ein durchdachtes Bearbeiten im Sinne des Optimizing ggf. erst dann gelingen kann, wenn zur Antwortfindung benötigte passive Wissensbestände aktiviert und damit erinnert werden (Retrieval). Dies kann jedoch, je nach Verankerungsgrad der notwendigen Informationen mit einem erhöhten kognitiven Aufwand verbunden sein. Die Befragten können zur Vermeidung eines solchen Aufwands Erinnerungsstrategien nutzen, welche nur einen verkürzten oder oberflächlichen Erinnerungsprozess zur Folge haben. Das Verhalten bei einem erhöhten Erinnerungsaufwand kann daher auch über den Satisficing-Ansatz beschrieben werden.

Conrad et al. (1998) verdeutlichen dies, indem sie schematisch verschiedene Strategien zur Beantwortung von Häufigkeitsfragen vorstellen und (auch in Bezug auf den kognitiven Aufwand) gegeneinander abgrenzen. Die folgenden Strategien werden hierbei von Conrad et al. (1998) unterschieden:

Abb. 5: Antwortstrategien bei Fragen zu Häufigkeiten



Quelle: Conrad et al. (1998), S. 361.

„At the highest level we distinguish between enumeration strategies and all other strategies. This reflects our belief that there is a fundamental difference between using remembered episodes and using generic, event-type information as the basis of a frequency report. Two of the strategies we explored appear under the “Enumeration“ heading: episode enumeration and rate estimation“ (Conrad et al. (1998), S. 360). Unter *episode enumeration* fällt das Erinnern an deutlich abgegrenzte Situationen, wobei die Erinnerungswerte bei Abruf der Häufigkeit einfach aufsummiert werden. Die Autoren gehen hierbei davon aus, dass diese Strategie vor allem bei seltenen Ereignissen, welche gut unterscheidbar sind, gewählt wird. Sind die Ereignisse sehr häufig oder ähnlich (und damit verwechselbar) wird den Autoren zufolge eine andere Strategie bevorzugt: „The second strategy, rate estimation, appears under *Enumeration* (...) because the relationship between reaction times and frequency reports implies that respondents retrieve individual episodes from a sample portion of the reference period and

then extrapolate to the entire period. The strategy seems to be preferred for events of moderate regularity and similarity, and it produces large estimates. These factors, in combination, may lead respondents to stop enumerating before they have retrieved all episodes“ (Conrad et al. (1998), S. 360). Dem gegenübergestellt sind die Strategien bei einem *Direct Retrieval*: „*Direct Retrieval* strategies operate on information that is encoded in respondents' memory before they hear the question. This stored information can be *Quantitative* or *Qualitative*. The kind of quantitative information that is stored is rate information, for example the knowledge that “I purchase gas several times a week“. Both the rate retrieval and rate adjust strategies seem to rely on retrieved knowledge of this type, and both are applied to regular and similar events. High regularity is a prerequisite for the availability of rate information and high similarity discourages episode enumeration (Menon, 1993)“ (Conrad et al. (1998), S. 362). Darüber hinaus kann auch aus einem qualitativen (allgemeinen) Eindruck heraus eine Häufigkeit abgeleitet werden. Diese Ableitung aus einer „General Impression“ benötigt dafür zwei mentale Übersetzungsschritte: „(...) a stored impression must be retrieved and, once it is retrieved , the impression must be converted into an actual number“ (Conrad et al. (1998), S. 362).

Bei einem Satisficing kann demgemäß erwartet werden, dass Häufigkeiten auf Basis von Schätzungen überwiegen, da a) die Erinnerung an jedes einzelne Ereignis sehr aufwändig sein kann und b) beim Vorliegen qualitativer Eindrücke diese erst in eine Häufigkeit umgewandelt werden müssen, um überhaupt gezählt werden zu können.

2.2.3 Die Nutzbarkeit des Satisficing-Konzepts zur Definition von Antwortqualität

Neben der Kritik an der Operationalisierung von Satisficing bleibt das Grundkonzept überzeugend, da aus ihm vielfältige Erklärungen für das Antwortverhalten abgeleitet werden können.

Dabei lassen sich auch Spezialfälle und Besonderheiten eines durchdachten Handelns erklären. So ist beispielhaft denkbar, dass nach reiflicher und reichlicher Überlegung der Befragten bewusst eine Falschantwort gegeben wird. Hierfür muss das Modell des kognitiven Antwortprozesses nur leicht angepasst werden. Tourangeau & Rasinski (1988) schlagen diesbezüglich eine Modifikation des Modells vor, indem aus der vierten Phase eine weitere Phase extrahiert wird: das Editing. Hierunter wird das bewusste Anpassen der Wahl einer Antwortkategorie zugunsten einer gewünschten Selbstdarstellung verstanden. Editing tritt den Autoren zufolge aus zwei verschiedenen Gründen auf: „the answer is checked for consistency with prior answers or for social desirability“ (Tourangeau & Rasinski (1988), S. 300). Eine falsche Antwort muss folglich nicht per se das Ergebnis eines unvollständigen Durchlaufens des Antwortprozesses sein, wenn eine Anpassung der gewählten Antwortkategorie auch als Mittel zur Herbeiführung von Konsistenz im Fragebogenverlauf verstanden wird.¹⁴ Die Begründung für ein Streben nach Konsistenz kann dabei neben einem intrinsischen Bedürfnis auch in der sozialen Erwünschtheit gesehen werden. Dies ist beispielhaft gegeben, wenn Befragte gezielt für die Forscher Konsistenzen generieren, da sie unterstellen, dass diese für Analysen wichtig und daher gewünscht sind.

2.3 Mikrozensusgesetz

Als weiteres Definitionsfundament der Antwortqualität wird das Gesetz zur *Durchführung einer Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt sowie die Wohnsituation der Haushalte* (Mikrozensusgesetz 2005 - MZG 2005) als staatlich autorisierte Bearbeitungsnorm herangezogen. Hierbei haben die ausgewählten Befragten des Mikrozensus, neben

¹⁴ Hierbei muss jedoch die Idee des „wahren“ Wertes aufgegeben werden. Es ist dabei dann sehr interessant dies mit dem Antwortgenerierungsprozess nach Zaller & Feldman (1992) zu verknüpfen, da nach deren Ansatz Antworten erst „on the spot“, also auf Abfrage (situationsbezogen) generiert werden.

einem kurzen freiwilligen Antwortteil, die gesetzliche Verpflichtung den Großteil der Fragen zu beantworten. Dies ist in § 7 Absatz 1 festgelegt:

(1) Für die Erhebungen besteht Auskunftspflicht, soweit in Absatz 4 nichts anderes bestimmt ist.¹⁵

Die Ausgestaltung der Auskunftspflicht ist für die Erhebungen im Rahmen des Mikrozensus nun wiederum in § 15 im Gesetz über die Statistik für Bundeszwecke geregelt und erstreckt sich über sechs Absätze. Relevant für die Erstellung einer Definition für Antwortqualität ist hierbei vor allem Absatz 3:

(3) Die Antwort ist wahrheitsgemäß, vollständig und innerhalb der von den statistischen Ämtern des Bundes und der Länder gesetzten Fristen zu erteilen. Die Antwort ist erteilt, wenn die ordnungsgemäß ausgefüllten Erhebungsvordrucke

- 1. bei Übermittlung in schriftlicher Form der Erhebungsstelle zugegangen sind,*
- 2. bei Übermittlung in elektronischer Form von der für den Empfang bestimmten Einrichtung in für die Erhebungsstelle bearbeitbarer Weise aufgezeichnet worden sind.*

Die drei oben angeführten Kategorien „wahrheitsgemäß“, „vollständig“ und „innerhalb der gesetzten Fristen“ sollen nun genauer – im Hinblick auf die Nutzbarkeit als Definitionsmerkmale für Antwortqualität – betrachtet werden.

¹⁵ Absatz 4 lautet: Die Auskünfte über das Erhebungsmerkmal Wohn- und Lebensgemeinschaft nach § 4 Abs. 1 Nr. 1, das Erhebungsmerkmal vermögenswirksame Leistungen und angelegter Gesamtbetrag nach § 4 Abs. 2 Nr. 2 sowie die Erhebungsmerkmale nach § 4 Abs. 1 Nr. 2 Buchstabe b und Nr. 14, Abs. 2 Nr. 1 und 3, Abs. 5 und die Hilfsmerkmale nach § 5 Abs. 1 Nr. 2 sind freiwillig.

2.3.1 Die wahrheitsgemäße Beantwortung

Im Rahmen der Befragung des Mikrozensus werden primär Fragen zu empirisch validierbaren Sachverhalten gestellt. Bei solchen Messungen kann daher noch am ehesten von einem „wahren“ Wert ausgegangen werden (z.B. Schulabschluss). Dies bedeutet jedoch nicht, dass die gegebenen Antworten auch alle „wahr“ sind. Dies begründet sich z.B. über Erinnerungseffekte, so dass eine falsche Angabe gemacht wird, welche aber auch, z.B. aufgrund einer fehlerhaften Erinnerung oder Verdrängung, als wahr geglaubt wird. Daneben sind aber auch bewusste Falschangaben möglich, was die Antwortqualität ebenfalls senkt. Die Problematik zum Konzept des „wahren“ Wertes für die Umfrageforschung wurde bereits in Kapitel 2.1.2 dargestellt und soll daher nicht noch einmal ausführlich dargestellt werden. In sozialwissenschaftlichen Befragungen werden – im Gegensatz zum Mikrozensus – zumeist Einstellungen und Bewertungen erfasst, ist die Forderung nach einem „wahrer“ Antwort nur schwer aufrecht zu erhalten. Daher erscheint es angebracht, den Ansätzen von Zaller & Feldman (1992) und Esser (1986) folgend, dann eher von einer situational wahrheitsgemäßen Beantwortung zu sprechen.

2.3.2 Die Vollständigkeit von Angaben

Der Begriff der Vollständigkeit bezieht sich im Rahmen der Befragung auf Item-Nonresponse. Item-Nonresponse kann untergliedert werden in MCAR (Missing Completely At Random), MAR (Missing at Random) und MNAR (Missing Not At Random) (vgl. Rubin (1976); Schnell (2012)). Die Reichweite der einzelnen Typologien von Item-Nonresponse sind hierbei jedoch recht unterschiedlich: Bei MCAR reduziert der völlig zufällig fehlende Wert die Stichprobengröße, aber wirkt sich ansonsten nicht auf die statistischen Analysen aus. Bei MAR kann Item-Nonresponse durch im Datensatz enthaltene Variablen erklärt werden, d.h. somit hängt die

Wahrscheinlichkeit des Fehlens von Antworten von beobachteten Variablen im Datensatz ab. Bei MNAR sind die zur Erklärung benötigten beobachteten Variablen nicht gegeben und folglich kann eine zugrunde liegende Ausfallsystematik nicht erfasst bzw. beschrieben werden. „Bei den allermeisten Surveys der empirischen Sozialforschung dürften die Ausfallprozesse weder MCAR noch MNAR sein, sondern MAR. Dies ist zwar einerseits erfreulich, da diese Art von Ausfällen prinzipiell korrigierbar ist, andererseits wird aber deutlich, dass eine Korrektur auch notwendig ist“ (Schnell (2012), S. 173).¹⁶ Als Lösungsstrategien werden zumeist zwei Verfahren genutzt: Sample-Selection-Modelle oder die multiple Imputation.¹⁷ Beide Verfahren sind an strenge Annahmen geknüpft und garantieren keine „besseren“ Ergebnisse. Aus diesem Grund ist eine vollständige, im Sinne von keine Antwort auslassende Bearbeitung des Fragebogens für die Auswertung von Fragebogendaten für die forschenden Personen sehr wichtig. Es kann weitergehend unterschieden werden, ob eine Frage wirklich übersprungen wurde oder eine Kategorie wie „Weiss nicht“ bzw. „Keine Angabe“ gewählt wurde. Diese Unterscheidung ist relevant, da bei der Wahl einer Missing-Kategorie noch immer eine (implizite) Bearbeitungsanweisung befolgt wird und damit im weiteren Sinne in den Vorgaben der Forscher gearbeitet wird. Die Ursachen für das Nichtbeantworten von Fragen werden dann zumeist über zwei Wege erklärt: über einen kognitionspsychologischen Ansatz und einen Rational Choice Ansatz (vgl. Rässler & Riphahn 2006). Der kognitionspsychologische Ansatz basiert auf einem kognitiven Antwortprozess (z.B. nach Tourangeau & Rasinski (1988)). Bei diesem kann in jeder der vier Phasen eine Störung erfolgen, z.B. aufgrund eines fehlenden Verständnisses (Comprehension), welche wiederum Item-Nonresponse begünstigt. Beim Rational Choice Ansatz wird hingegen vor allem das bewusste Abwägen des Nutzens und der Kosten (z.B. Zeit,

¹⁶ Kritisch zu sehen, nur wenn man Ausfallprozess kennt kann er modelliert werden.

¹⁷ Die multiple Imputation wird bei dem Ausfallprozess MAR eingesetzt, während Sample Selection-Modelle bei MNAR genutzt werden können.

Ressourcen) einer Beantwortung als Grund für Item-Nonresponse genannt und weist damit einen globaleren Charakter auf.

Neben dieser negativen Färbung von Item-Nonresponse kann aber auch ein positiver Blickwinkel eingenommen werden, nämlich dann, wenn die befragte Person eine Antwort verweigert, da sie sich nicht in den Antwortkategorien wiederfindet und eine Falschangabe vermeiden möchte. In einem solchen Fall liegt der Grund für Item-Nonresponse in der Fragebogenkonstruktion, da keine erschöpfenden Antwortvorgaben umgesetzt wurden und auch Antwortkategorien wie „Weiß nicht“ oder „Keine Angabe“ fehlten.¹⁸ Letztlich muss aber festgestellt werden, dass nur schwer festgestellt werden kann, ob wirklich Meinungslosigkeit, kognitives Unvermögen, Aufwandsminimierung oder eine vom Forscher vorgegebene unvollständige Antwortbandbreite Grund für die Antwortverweigerung ist.

2.3.3 Das Einhalten gesetzter Fristen

Die letzte definitorische Komponente betrifft das rechtzeitige Bearbeiten und Rücksenden von Fragebögen, da der Erhalt eines Fragebogens zumeist an eine vermerkte Frist gekoppelt ist. Diese Frist dient den Forschern dazu, den Zeitplan im Untersuchungsablauf einzuhalten und – neben dem Ziel die Daten ab einem bestimmten Zeitpunkt analysieren zu können – auch die Kosten für die Erhebungsphase zu kalkulieren. Dillman et al. (2009) haben mit der Total Design Method (später: Tailored Design Method) ein umfassendes Instrumentarium erstellt, worin verschiedene Maßnahmen zur Steigerung der Rücklaufquote vorgeschlagen werden. Die Ein-

¹⁸ Es ist zu beachten, dass die Antwortmöglichkeiten „weiß nicht“ oder „Keine Angabe“ auch eine inhaltliche Ebene widerspiegeln und daher eigentlich nicht gewählt werden können. Eine Antwort, welche nicht zu den gegebenen Antwortkategorien passt ist etwas anderes als Unwissen oder Antwortverweigerung.

haltung von Fristen erscheint auf den ersten Blick als definitorisches Merkmal von Antwortqualität vielleicht ungewöhnlich, unter Berücksichtigung des Aspekts des Befolgens von Arbeitsanweisungen kann es jedoch als plausibler Bestandteil verstanden werden. In diesem Kontext sollte den Befragten die Frist erläutert und deutlich erkennbar aufgezeigt werden.

2.4 Die Definition von Antwortqualität

Nach Darlegung der zwei theoretischen Konzepte und den normativen Forderungen des Mikrozensusgesetzes werden nun alle relevanten Aspekte in einer Definition zusammengefügt werden. Mit Hilfe des Total Survey Error wurde eine Einordnung in die Fehlerkategorie der Reaktionsphase vorgenommen und daraus folgend Konsequenzen für das Befragtenverhalten übernommen. Im Rahmen der Satisficing-Konzeption von Krosnick (1991) soll die Komponente des durchdachten Antwortens im Sinne des Optimizing übernommen werden und aus dem Mikrozensusgesetz die Verpflichtung zu einer wahrheitsgemäßen, vollständigen und anweisungsbefolgenden Beantwortung herangezogen.

Die Definition lautet:

Eine hohe Antwortqualität von Befragten ist gegeben, wenn das Beantworten der gestellten Fragen *durchdacht, vollständig und (situational) wahrheitsgemäß ist, ggf. unter Rückgriff auf externe Wissensbestände, bei Beachtung von Arbeitsanweisungen.*¹⁹

Die Antwortqualität wird hierbei, wie auch das Satisficing als Reaktion auf Gegebenheiten verstanden und erhält damit einen variablen Charakter. So kann die Antwortqualität zwischen

¹⁹ Es kann argumentiert werden, dass nicht vier sondern nur zwei Komponenten die Antwortqualität bestimmen. Dies kann dadurch gestützt werden, dass ein durchdachtes und wahrheitsgemäßes Beantworten eine inhaltliche Ebene der Antwortqualität anspricht und das vollständige Bearbeiten und Befolgen von Arbeitsanweisungen den formalen Aspekt umfasst.

Befragungen oder auch innerhalb einer Befragung variieren. Unter Gegebenheiten wird beispielhaft die Erhebungssituation, das Erhebungsinstrument oder die Befindlichkeit verstanden.

Diese Definition ist die Grundlage für die Analyse der Wirkung von Incentives auf die Antwortqualität. Das folgende dritte Kapitel behandelt die Definition von Incentives und gibt einen Überblick über den aktuellen Forschungsstand, mit dem Ziel die Forschungslücke aufzuzeigen, welche mit dieser Studie geschlossen werden soll.

3. Die Bedeutung von Incentives für die Umfrageforschung

Da im Rahmen dieser Studie geprüft werden soll, inwiefern die Incentivierung auf die Antwortqualität wirkt, soll zuerst geklärt werden, was unter einem Incentive verstanden werden kann. Hierfür werden zwei Erläuterungen vorgestellt:

„The reasons for providing this payment or gift may be to encourage people to cooperate with the survey, or to thank them for taking part. These payments or gifts are typically referred to as “incentives”, suggesting perhaps that the emphasis is on encouragement.” (Laurie & Lynn (2008), S. i)

Die erste Beschreibung stellt vor allem den motivierenden Charakter von Incentives in den Vordergrund. Ein Incentive kann hierbei jedoch von den Befragten auf verschiedene Weise interpretiert werden: Die Befragten können das Incentive als „Bezahlung“ oder als „Danke schön“ wahrnehmen. Diese Unterscheidung kann insofern relevant werden, als dass sie, gemäß dem Framing-Ansatz von Esser (1999), bei den Befragten eine unterschiedliche Wahrnehmung der Befragungssituation bedeuten kann. Singer & van Hoewick (1998) haben sich

mit der Fragestellung zur Bedeutung von Frames auf Wirkung von Incentives beschäftigt und kommen zu folgenden Ergebnissen:

Tab. 3: Die Wirkung von Frames auf die Teilnahmebereitschaft

<i>Experimental Condition</i>	(1)	(2)	<i>(N)</i>
	<i>Agreement to Be Interviewed</i>	<i>Response Rate</i>	
	%	%	<i>N</i>
Pen as a token of appreciation	30.4	76.0	125
Pen as payment for time	33.6	75.2	125
\$10 check as token of appreciation	51.2	79.2	125
\$10 check as payment for time	64.0	81.6	125
\$10 check as payment, promised	44.8	84.0	125

Quelle: Singer & van Hoewick (1998), S. 14.²⁰

Die Betitelung eines 10\$ Checks als „Bezahlung“ bewirkte eine signifikant höhere Teilnahmebereitschaft als die Bezeichnung „Dankeschön“ bzw. „kleine Aufmerksamkeit“.²¹ Dies stützt die Vermutung, dass die Definition der Situation (der Frame) einen Einfluss auf die Wirkung von Incentives aufweisen kann.

Der mehrdeutige Charakter von Incentives wird auch in der Definition von Schnell (2012) ersichtlich:

„Als Befragungsanreize oder „Incentives“ werden kleine Geschenke (Kugelschreiber, Briefmarken etc.), seltener Dienstleistungen (Kinderbetreuung für die Dauer der Befragung, medizinische Untersuchungen) und vor allem finanzielle Entlohnungen bezeichnet“ (Schnell 2012, S. 181 f.).

²⁰ Die Response Rate beschreibt den Rücklauf der Antwortschreiben. Auf diesen sollten die Probanden angegeben, ob sie an einer Befragung teilnehmen möchten.

²¹ Ob und inwiefern eine Verschiedenheit eines Frames für die Befragungssituation auch eine Verschiedenheit für die Konsequenzen bezüglich der Bearbeitung an einer Umfrage aufweist ist bisher aufgrund der wenigen einschlägigen Experimente unklar.

Diese eher globale Definition umfasst jedoch nicht alle Charakteristika von Incentives. Der Zusammenschluss *The Council of Professional Associations on Federal Statistics* (auch: COPAFS), mit Sitz in den USA versuchte den Begriff der Incentives umfassender zu definieren, scheiterte jedoch an der Vielfalt der zu berücksichtigenden Komponenten. Im Rahmen einer Tagung erstellten die anwesenden Personen eine Liste mit dem Ziel, die verschiedenen Facetten von Incentives zu erfassen und zu sortieren²². Diese Liste dient folgend als Grundlage für die Erläuterung der verschiedenen Aspekte von Incentives.

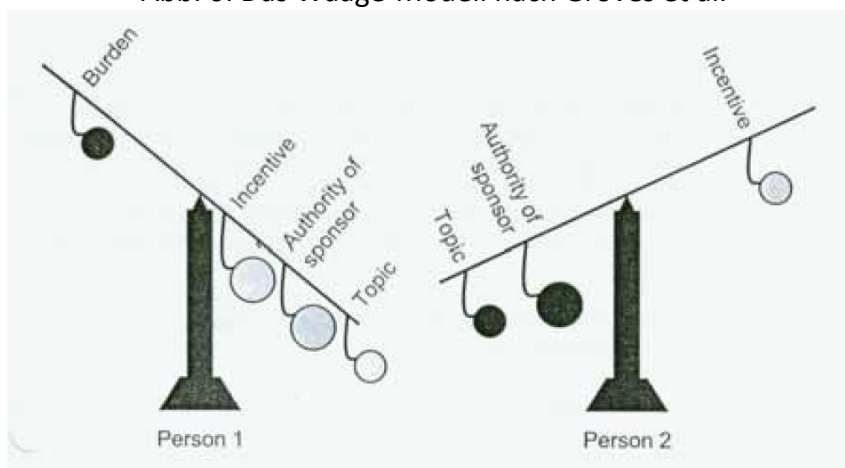
Zuerst kann zwischen monetären und nicht-monetären Incentives unterschieden werden. Bei einem monetären Incentive handelt es sich um Geld oder geldäquivalente Gegenstände (z.B. Bonuspunkte). In Studien werden sehr unterschiedliche monetäre Beträge ausbezahlt, welche aber üblicherweise zwischen 1 Euro und 20 Euro liegen (Stadtmüller (2009); Fick & Diehl (2013), Pforr (2015)). Nicht-monetären Incentives (z.B. Kugelschreiber, Schlüsselanhänger) wird, im Gegensatz zu monetären Incentives, die prinzipielle Universalität als Tauschmittel abgesprochen. Sie sind damit eher nach ihrer Nützlichkeit als nach ihrem materiellen Wert zu beurteilen. Weiter gilt es zu beachten, dass ein Incentive zu verschiedenen Zeitpunkten ausgegeben werden kann. Ein Incentive, welches vor Beginn der Befragung ausgegeben wird, wird als Pre-Paid oder unkonditional bezeichnet. Es ist insofern unkonditional, als dass es nicht an eine bestimmte Leistung des Befragten gekoppelt ist (z.B. die Teilnahme, oder der erfolgreiche Abschluss der Befragung). Wird das Incentive erst im Nachhinein ausgegeben, so spricht man von einem Post-Paid oder konditionalen Incentive. Des Weiteren gibt es die Möglichkeit ein Incentive, neben der (verbindlichen) Zusage, auch über eine Verlosung in Aussicht zu stellen. Damit hat der Befragte eine bestimmte, für ihn aber meist unbekannte Wahrscheinlichkeit,

²² vgl.: www.copafs.org/reports/providing_incentives_to_survey_respondents.aspx#defining

ein monetäres oder nicht-monetäres Incentive zu gewinnen. Die Art und Höhe eines Incentives sollte dabei im Hinblick auf die Art der Empfänger gewählt werden. Hier kann zwischen Individuen und Kollektiven unterschieden werden. Als Kollektiv kann zum Beispiel ein Haushalt oder eine Institution betrachtet werden.

Dem Rational Choice Ansatz folgend ist die subjektive Wahrnehmung und Gewichtung des Incentives für dessen Wirkung relevant: So mag beispielhaft ein Incentive in Höhe von 5 Euro bei Hochverdienern anders wahrgenommen werden als bei Studierenden. Groves et al. (2004) berücksichtigen dies in dem sog. Waage-Modell zur Erklärung der Teilnahme an Befragungen:²³

Abb. 6: Das Waage-Modell nach Groves et al.



Quelle: Groves et al. (2004), S. 177.

Bei Person 1 kann das Incentive überzeugen an der Befragung teilzunehmen und den damit verbundenen Aufwand akzeptiert. Dies bedeutet, dass bei Person 1 das Incentive den Nutzen so weit steigerte, dass es den antizipierten Kosten überwiegt. Bei Person 2 hingegen konnte das Incentive nicht genügend „Gewicht“ aufbringen. Der Aufwand für die Teilnahme an einer Befragung erscheint trotz Incentive zu hoch, was zu einer Teilnahmeverweigerung führt. Der Nutzen konnte in diesem Fall nicht überwiegen. Unter Nutzen kann, neben dem Erhalt eines

²³ Das Waage-Modell wird im Rahmen der Leverage-Saliency-Theory von Groves et al. (2000) eingeführt.

Incentives, z.B. auch die Partizipation an Entscheidungen (über Meinungsbefragungen), die Unterstützung von bestimmten Forschungszweigen, die Mitwirkung bei der Erhebung neuer Forschungsergebnisse oder Selbstverwirklichung verstanden werden. Auf der Kostenseite liegen Abwägungen z.B. über die vermeintliche Länge der Befragung, die Entfernung zum Ort der Befragung, ob es sich um eine einmalige oder zum Auftakt zu einer langjährigen Befragung handelt, der Modus der Datenerhebung (z.B. mündlich, schriftlich), die verlangten Informationen (z.B. sensible Daten, als unangenehm eingestufte Fragebereiche, Durchführung von medizinischen Tests) und die damit antizipierten Konsequenzen (z.B. bei Angabe von illegalen Aktivitäten). Auch Sponsor-Effekte können die Abwägung zur Teilnahme negativ beeinflussen. Bei einer Teilnahmepflicht können Incentives die negative Wahrnehmung des Zwangsgefühls mindern und die Teilnahme ggf. dennoch als positiven Akt erscheinen lassen.

Darüber hinaus werden auch häufiger ethische Bedenken gegen die Vergabe von Incentives hervorgebracht (vgl. Grant (2012); Singer & Couper (2008)). Diese liegen z.B. darin begründet, dass bei einer sehr hohen Incentivierung die potentiellen Probanden dem Incentive nicht mehr widerstehen können und an Studien teilnehmen, welche für sie hohe, bzw. nicht abschätzbare Risiken bergen können (z.B. Medikamentenstudien).²⁴

Unter Berücksichtigung all der angedeuteten Facetten von Incentives erscheint es nicht verwunderlich, dass die Forschung über Incentives sehr vielseitig ist und nur in wenigen Bereichen von einem tendenziell einheitlichen Ergebnis gesprochen werden kann. Incentives wird üblicherweise eine Wirkung auf vier Zielgrößen zugesprochen (vgl. Stadtmüller (2009)): a) der Rücklaufquote, b) der Rücklaufgeschwindigkeit, c) der Zusammensetzung der Stichprobe und

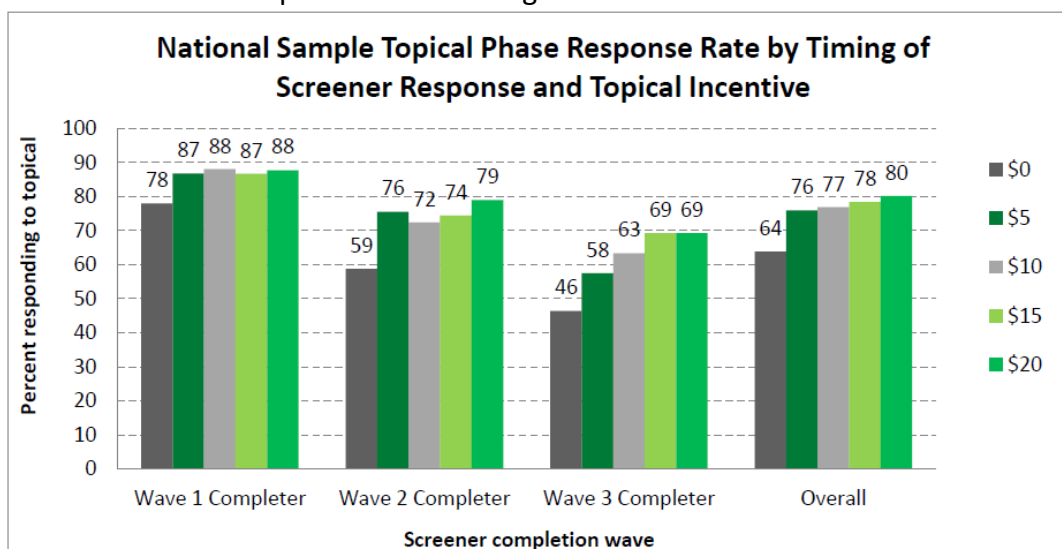
²⁴ Singer & Couper (2008) führten dafür eine Vignettenstudie durch, konnten aber den vermuteten Effekt nicht nachweisen.

d) der Datenqualität. Nachfolgend werden Studien dargestellt, welche sich dadurch kennzeichnen, dass sie stellvertretend für eine Vielzahl ähnlicher Untersuchungen stehen oder aber entgegengesetzte Ergebnisse bieten.

3.1 Die Rücklaufquote

Es liegt eine hohe Anzahl an Studien vor, die sich mit der Wirkung von Incentives auf die Rücklaufquote beschäftigen. Hierbei konnte festgestellt werden, dass ein unkonditionales Incentive in verschiedenen Settings tendenziell einen positiven Effekt auf die Rücklaufquote aufweist (vgl. die Meta Analysen von Church (1993) und Singer (1998)), wobei mit unterschiedlicher Höhe des Incentives auch verschiedene Effektstärken beobachtet werden konnten. Die Wirkungen scheinen dabei abhängig von der Zielgruppe zu sein. In der Studie von McPhee & Hasted (2012) wird daher der Zusammenhang von Höhe und Teilnahme ausführlich an einer Bevölkerungsstichprobe untersucht:

Abb. 7: Response-Raten untergliedert nach Höhe des Incentive



SOURCE: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey (NHES): 2011 Field Test.

Quelle: McPhee & Hasted (2012), S. 10.

Es ist der Grafik zu entnehmen, dass ein Schwellenwert für die Rücklaufquote vorliegt, ab welchem eine Erhöhung des Incentives nur noch (wenn überhaupt) einen geringen Zuwachs bewirkt (5\$). Dieser Schwellenwert kann allerdings nicht global interpretiert werden, da in spezifischen Zielgruppen abweichende Schwellenwerte bzw. Schwellenwertstrukturen vorliegen können. In der Studie von Dykema et al. (2011) konnte z.B. gezeigt werden, dass erst bei einem sehr hohen monetären Incentive auch sonst schwer zu erreichende Bevölkerungsgruppen (z.B. Ärzte) befragt werden konnten (die Teilnahmequote lag bei ca. 4.5% wenn kein Incentive versprochen wurde, bei ca. 14% bei einem Incentive in Höhe von 50\$ und bei ca. 25% bei einem Incentive in Höhe von 100\$). In der Studie von James & Bolstein (1992) konnte hingegen aufgezeigt werden, dass zu hohe un konditionale Incentives (20\$ zu 40\$) eine Abnahme der Rücklaufquote bewirken kann. Die Autoren finden jedoch keine schlüssige Erklärung für diesen Befund.

Im Vergleich zu Pre-Paid Incentives weisen Post-Paid Incentives²⁵ tendenziell eine geringere Wirkung auf (vgl. Yu & Cooper (1983), Singer et. al (1999)). Bei Lotterien kann fast kein Effekt auf die Rücklaufquote festgestellt werden (vgl. Warriner et al. (1996)). Ausnahmen hierfür liefern die Befunde der Studien von Hubbard & Little (1988), sowie Bosnjak & Tuten (2003). Dort konnte gezeigt werden, dass auch eine Lotterie einen positiven Effekt auf die Rücklaufquote aufweisen kann: „Prize draws significantly increase the willingness to participate and, eventually, also the number of sample units starting the survey. Prize draws also increase actual participation (completion rates) and tend to reduce the number of incompleteness compared to no incentives“ (Bosnjak & Tuten (2003), S. 215).

²⁵ In der Meta-Analyse von Church (1993) konnte überhaupt kein Effekt des Post-Paid Incentives (im Vergleich zur Kontrollgruppe) nachgewiesen werden.

Es ist anzumerken, dass in der Studie von Robertson & Bellenger (1978) die Möglichkeit zur Spende eines monetären Incentives an Dritte zu einem enormen Anstieg der Rücklaufquote geführt hat (41.3% vs. 23.3% bei der wohltätigen Spende von 1\$). Dieses Ergebnis konnte jedoch nicht mehr reproduziert werden. Auch nicht-monetäre Incentives weisen einen positiven Effekt auf die Rücklaufquote auf. So konnte im Rahmen einer Online-Erhebung unter Teilnehmern eines Online-Rollenspieles (Williams et al. (2011)) gezeigt werden, dass ein passender Anreiz (in diesem Fall ein digitales Item zur Verbesserung eines Rollenspielcharakters: Greatstaff of the Sunserpent“) eine sehr hohe Teilnahmebereitschaft bewirkt hat.²⁶ Es ist zu betonen, dass ein solcher Zielgruppeneffekt bei Ryu et al. (2006) in einem anderen Setting nicht repliziert werden konnte. Dort wurde den Befragten, auf Basis einer Randomisierung, entweder ein Incentive in Höhe von 5\$ oder eine Eintrittskarte in den Metropark (Wert ca. 12\$) für die Teilnahme versprochen. Die Ergebnisse entsprachen hierbei nicht den Erwartungen der Forscher, da die Personen, welche in ihrer Freizeit den Metropark besuchen (und damit an der Eintrittskarte interessiert sein könnten), nicht häufiger in der Befragung vertreten waren, als die Personen, denen 5\$ versprochen wurde.

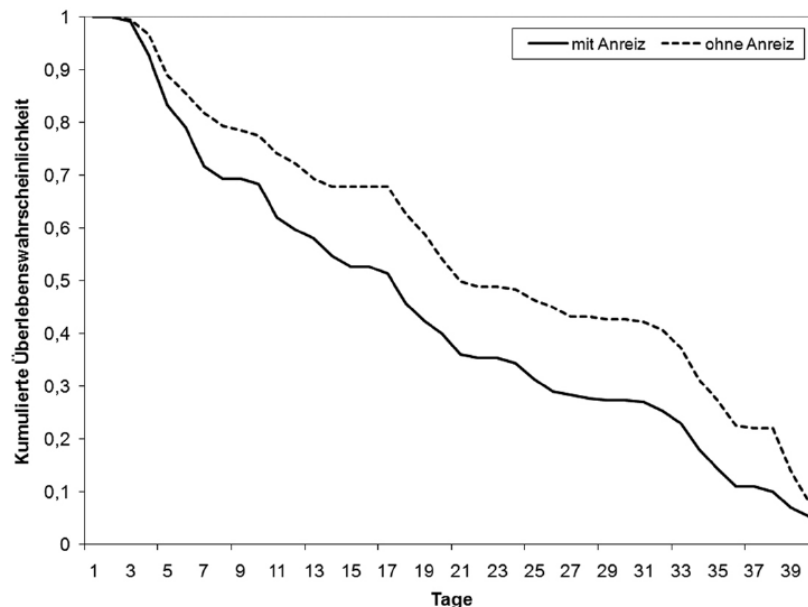
Insgesamt kann aus den Ergebnissen der Studien abgeleitet werden, dass der Einsatz von Incentives die Rücklaufquote tendenziell erhöht (Pforr (2015); Dillman et al. (2009); Stadtmüller & Porst (2005)).

²⁶ Es muss zu dieser Studie bemerkt werden, dass es keine Kontrollgruppe gab und daher der Effekt nicht kontrolliert wurde.

3.2 Die Rücklaufgeschwindigkeit

Neben der Rücklaufquote wird auch die Wirkung von Incentives auf die Rücklaufgeschwindigkeit untersucht. Die Forschungsergebnisse weisen hierbei tendenziell darauf hin, dass ein unkonditionales Incentive eine Erhöhung der Rücklaufgeschwindigkeit bewirken kann. So zeigt sich in der postalischen Befragung von Stadtmüller (2009) folgendes Ergebnis: Nach 20 Tagen standen bei der Gruppe mit monetärem Anreiz noch knapp 40% der Fragebögen aus, wobei in der Kontrollgruppe zum gleichen Zeitpunkt noch 54% ausstanden (vgl. Abb. 8):

Abb. 8: Rücklaufgeschwindigkeit bei einer Umfrage mit unkonditionalem 1€-Incentive

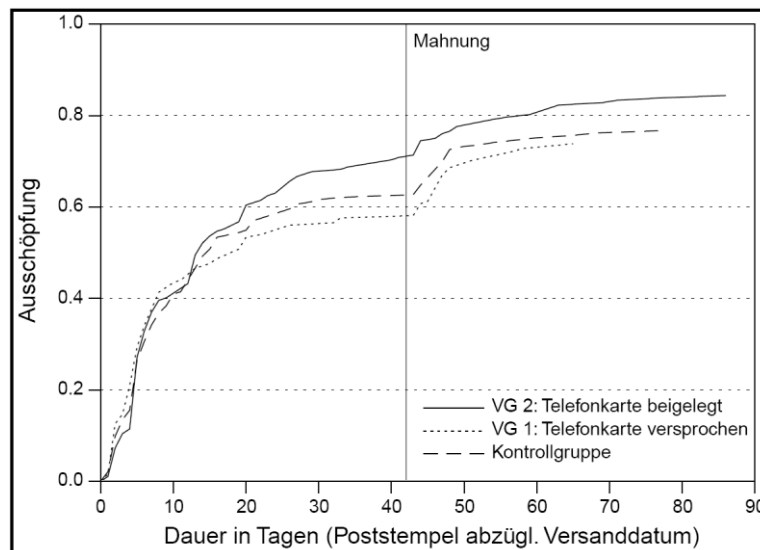


Quelle: Stadtmüller (2009), S. 178.

Diekmann & Jann (2001) kommen zu einem ähnlichen Ergebnis. Hierbei wurden Telefonkarten im Wert von 10 Schweizer Franken vor der Umfrage an die Teilnehmer ausgegeben bzw. den potentiellen Teilnehmern bei Teilnahme versprochen. In Abb. 9 kann die Entwicklung der Rücklaufquote nachvollzogen werden.²⁷

²⁷ Hierbei ist auch der bereits beschriebene Befund der Teilnahmesteigerung ersichtlich. Das Post-Paid Incentive führt zu einer geringeren Ausschöpfungsquote als das Pre-Paid Incentive. Es soll ebenfalls darauf hingewie-

Abb. 9: Rücklaufgeschwindigkeit bei Vergabe oder Ankündigung einer 10 Sfr Telefonkarte



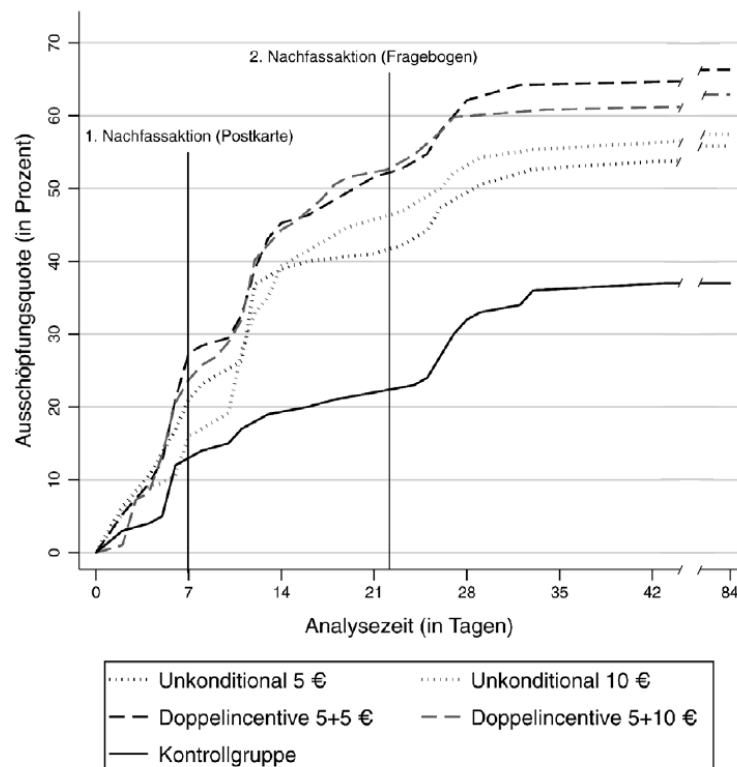
Quelle: Diekmann & Jann (2001), S. 24.

Es kann der Grafik entnommen werden, dass die Befragten mit beigelegter Telefonkarte die Fragebögen früher zurückschickten, als die Befragten ohne bzw. mit versprochenem Incentive. So wird die Ausschöpfungsquote von 60% bei un konditionaler Vergabe einer Telefonkarte nach ca. 20 Tagen erreicht, während in der Gruppe der Befragten mit der versprochenen Telefonkarte ca. 45 Tage gewartet werden musste. Es ist hervorzuheben, dass in dieser Darstellung Unterschiede in den Versuchsgruppen erst bei einer Ausschöpfungsquote von ca. 50% deutlich werden und somit erst im fortgeschrittenen Verlauf ein Effekt auf die Rücklaufgeschwindigkeit erkennbar ist.

Fick & Diehl (2013) nutzen in ihrer Studie eine doppelte Incentivierungsstrategie. Hierbei wurde bei zwei der fünf Versuchsgruppen zu Beginn einer postalischen Befragung ein un konditionales Incentive übersendet und bei erfolgreichem Abschluss der Befragung ein weiteres konditionales Incentive ausgezahlt.

sen werden, dass die Versuchsgruppe 1 (VG 1) und die Kontrollgruppe nicht signifikant im Rahmen der Rücklaufgeschwindigkeit unterscheiden ($p = 0.359$) und das versprochene Incentive damit keinen Einfluss auf die Rücklaufgeschwindigkeit aufweist.

Abb. 10: Die Rücklaufgeschwindigkeit bei einfacher und doppelter Vergabe von Incentives



Quelle: *Deutsche/r bleiben?* (n=483)

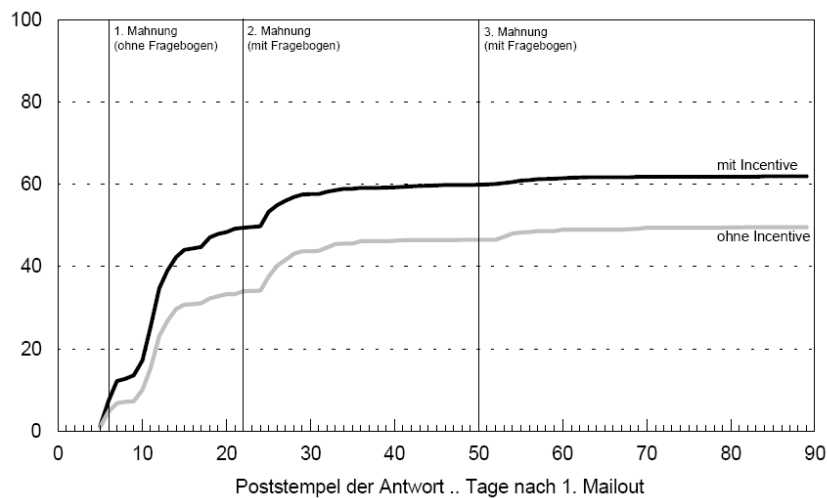
Quelle: Fick & Diehl (2013), S. 79.

Trotz der variierenden Verläufe konnten die Autoren keine signifikanten Unterschiede zwischen den Versuchsgruppen feststellen.²⁸ Es ist dennoch hervorzuheben, dass bei Verwendung des Doppelincentive die Rücklaufgeschwindigkeit, im Vergleich zur Kontrollgruppe, in der Studie erhöht war und damit einen positiven Effekt aufweist.

Arzheimer & Klein (1998) haben die Rücklaufgeschwindigkeit bei Panelbefragungen analysiert, wobei un konditionale Incentives (Telefonkarte im Wert von 6 DM) eingesetzt wurden. In der ersten Erhebungswelle (Abb. 11) konnte eine signifikant erhöhte Rücklaufgeschwindigkeit festgestellt werden.

²⁸ Die Autoren verweisen als Erklärung auf die geringe Fallzahl.

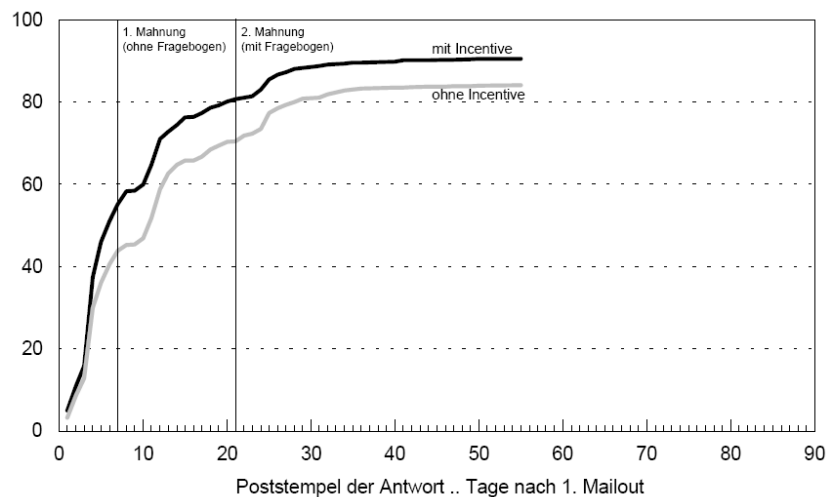
Abb. 11: Die Rücklaufgeschwindigkeit bei Beilegung eines monetären Incentives – 1. Welle



Quelle: Arzheimer & Klein (1998), S. 15.

Zu Beginn der zweiten Welle wurde erneut eine Telefonkarte im Wert von 6 DM dem Fragebogen beigelegt. Auch hier zeigten sich signifikante Unterschiede: Die Befragten mit einem Incentive haben signifikant schneller geantwortet als die Befragten ohne Incentive.

Abb. 12: Die Rücklaufgeschwindigkeit bei Beilegung eines monetären Incentives – 2. Welle



Quelle: Arzheimer & Klein (1998), S. 15.

Zusammengefasst lauten die Empfehlungen daher überwiegend, dass ein Pre-Paid Incentive einen positiven Effekt aufweist und folglich eingesetzt werden sollte. Dies fördert nicht nur einen Anstieg der Rücklaufgeschwindigkeit, und damit einen schnelleren Zugang zu Daten, sondern bewirkt zudem auch eine Kostenersparnis, da z.B. Personalkosten (Interviewer) oder

Sachkosten (z.B. Raummiete) gesenkt werden können (vgl. Fick & Diehl (2013), Berlin et al. (1992)).

3.3 Die Zusammensetzung der Stichprobe

Da nach den Befunden aus Kapitel 3.1.1 angenommen werden kann, dass bei Vergabe eines Incentives die Teilnahmebereitschaft positiv beeinflusst wird, stellt sich die Frage, ob und inwiefern sich dadurch die Zusammensetzung der Stichprobe verändert. Singer et al. (1999) stellten im Rahmen ihrer Meta Analyse fest: „(...) there is an indication that paying an incentive may be useful in obtaining higher numbers of respondents in demographic categories that might otherwise tend to be underrepresented in sample surveys (e.g., low income or nonwhite race)“ (Singer et al. (1999), S. 225). In der Literaturzusammenfassung von Simmons & Wilmot (2004) wird sogar ein starker Effekt skizziert: „Incentives have repeatedly been found to increase co-operation rates among certain groups: low-income and low-education groups, larger households and households with dependent children, minority ethnic groups and younger respondents“ (Simmons & Wilmot (2004), S. 6). Jedoch sind diese Ergebnisse nicht so eindeutig fundiert wie sie von Simmons & Wilmot (2004) dargestellt werden. Dies zeigt sich bereits in den teilweise widersprüchlichen Daten der Meta Analyse von Singer et al. (1999). Diese Widersprüchlichkeit lässt sich auch anhand von aktuellen Studien skizzieren: In der Studie von Stadtmüller (2009) ist in der Versuchsgruppe mit einem un konditionalen Incentive der Anteil der Hauptschüler signifikant geringer, während in der Studie von Fick & Diehl (2013) bei Vergabe eines Doppelincentives der Anteil der Befragten mit Abitur gesunken

ist.²⁹ Bei Arzheimer & Klein (1998), sowie Warriner (1996) kann hingegen überhaupt kein empirischer Zusammenhang zwischen dem sozioökonomischen Status und einem Incentive festgestellt werden.³⁰ Zusammengefasst kann also keine eindeutige Aussage bezüglich der Wirkung von Incentives auf die Stichprobenzusammensetzung getroffen werden.

3.4 Die Qualität der Umfrage

Eine Vielzahl an Studien hat sich ebenfalls mit der Frage beschäftigt, inwiefern die Vergabe eines Incentives einen Einfluss auf die Qualität der Befragung aufweist. Prinzipiell wird angenommen, dass die Qualität der Daten - unter Verwendung von Incentives - (wenn überhaupt) eine leichte Verbesserung aufweist (vgl. Pforr (2015), Boulianne (2008)). Der Großteil der Studien nutzte Item-Nonresponse und/oder die Anzahl an Worten bei offenen Fragen als Indikatoren (vgl. Stadtmüller (2009); Singer et al. (2000); Willimack et al. (1995); James & Bolstein (1990)). Seit den letzten 15 Jahren werden nun, auch aufgrund technischer Möglichkeiten, vermehrt alternative Indikatoren zur Messung der Datenqualität genutzt (z.B. Bereitschaftserklärung für Urinproben (Krenzke et al. (2005)) oder höhere Informationsvielfalt bei Kontaktinformationen für Folgebefragungen (Shettle & Mooney (1999))). Die Ergebnisse verschiedener Indikatoren können inhaltlich jedoch nur schwer miteinander verknüpft werden, gerade da – wie im zweiten Kapitel beschrieben – keine Definition der angestrebten Qualitätsdimension vorgenommen wird. Medway (2012) hat bisher die umfangreichste Arbeit zur Wirkung

²⁹ Bei der Interpretation ist zu beachten, dass eine Migrantenstichprobe zugrunde liegt. Es ist dabei zusätzlich unklar, ob tatsächlich weniger Befragte mit Abitur in der Stichprobe enthalten sind. Die Analyse der genannten Bildungserfolge im Mikrozensus (Schimpl-Neimanns (2006)), weisen deutliche Inkonsistenzen auf, welche auf Übertreibungen des eigenen Bildungsabschlusses zurückgeführt werden können. Eine Abnahme an Abiturienten könnte demnach alternativ auch als ein Anzeichen für ein ehrlicheres Beantworten aufgrund der hohen Doppelincentivierung gewertet werden.

³⁰ Darüber hinaus bleibt unklar, inwiefern eine Veränderung der Stichprobenzusammensetzung als positiv oder negativ wahrgenommen wird. Unter einer positiven Stichprobenveränderung kann beispielhaft das Erreichen schwer zugänglicher Personengruppen gezählt werden, z.B. ein erhöhter Anteil an Ärzten, aufgrund eines hohen Incentives. Negativ wird hingegen die Teilnahme z.B. finanziell schwacher Personen angesehen, da diese Incentives als besonders begehrt empfinden können und in der Folge in der Stichprobe überrepräsentiert sind.

von Incentives auf die Qualität von Daten verfasst und untersuchte insgesamt 12 Indikatoren der „Datenqualität“³¹, mitsamt Strukturmerkmalen im Datensatz (z.B. Stabilität von Parametern in Strukturgleichungsmodellen). Insgesamt konnten bei Vergabe eines unkonditionalen Incentives (5\$) in einer telefonischen Befragung (n = 900) zwischen den Versuchsgruppen so gut wie keine signifikanten Unterschiede im Bearbeitungsverhalten festgestellt werden. Ausnahmen bildeten die Befunde zu Item-Nonresponse (Reduktion von 3% auf 2%) und der Dauer des Interviews (Reduktion von 17 Sekunden pro Frage auf 16 Sekunden pro Frage). Unklar ist jedoch, inwiefern diese Ergebnisse verallgemeinert werden können, da Bonke & Fallesen (2010) in ihrer Studie aufzeigten, dass die Wirkung von Incentives auf die Qualität (gemessen über Item-Nonresponse und die Anzahl an freiwillig gegebenen Informationen) durch den Erhebungsmodus beeinflusst werden kann. Sie berichteten, dass bei Telefoninterviews eine signifikant geringere Qualität festgestellt werden konnte, als bei einer parallel durchgeführten Online-Befragung.³² In der Studie von Medway (2012) bleibt darüber hinaus unklar, ob und inwiefern die Höhe des Incentive auf die „Datenqualität“ wirkt. Während James & Bolstein (1990) lineare Zusammenhänge zwischen der Höhe des unkonditionalen Incentives (0\$, 0.25\$, 0.5\$, 1\$ und 2\$) und der Anzahl an geschriebenen Worten aufdecken konnten, wurden bei Davern et al. (2003) keine Zusammenhänge bei Vergabe eines Incentives von 10\$ bzw. 20\$ auf die „Datenqualität“ festgestellt.³³ Gleichzeitig stellten die Autoren aber auch fest, dass die

³¹ Zu den Indikatoren zählten: Item Nonresponse, geschriebene Worte in offenen Fragen, Nicht-Differenzierung in Rating Skalen, Akquieszenz, Reihenfolge-Effekte, Aufmerksamkeitsmessungen, Rundungen von Häufigkeitsangaben, Erinnerungsstrategien für Häufigkeitsangaben, Filterverhalten, Dauer des Interviews, Validierungsfragen und eine Einschätzung der Telefoninterviewer über die Bemühungen der Befragten. Leider wurde ein Großteil der Indikatoren nicht theoretisch begründet und keine Definition der Antwortqualität vorgenommen.

³² Das Incentive war in dieser Studie die Teilnahme an einer Lotterie.

³³ Von den Autoren wurden das Editieren und Ergänzen von Informationen (im Vergleich zur Vorbefragung) und Item Nonresponse als Indikatoren für die „Datenqualität“ gewählt.

Befragten den Auftraggeber der Studie mit steigendem Incentive besser bewerteten.³⁴ Hervorzuheben ist an dieser Stelle der Befunde von Barge & Gehlbach (2012): Sie fanden einen negativen Effekt auf die „Datenqualität“ bei einem zu hohen Incentive (15\$, bzw. 20\$).³⁵

Insgesamt bleibt festzuhalten, dass die Frage zur Wirkung von Incentives auf die „Datenqualität“, noch nicht geklärt ist. Dies wird u.a. dadurch deutlich, dass der Begriff der „Datenqualität“ nicht definiert wird und die Auswahl der Indikatoren „Datenqualität“ zwar zumeist begründet, aber nicht theoretisch eingeordnet werden. So ist beispielhaft die Nutzung von Item-Nonresponse als Indikator problematisch, da bei einem diffusen Begriff der „Datenqualität“ unklar ist, welche Qualitätsdimension angesprochen wird. So kann z.B., je nach Argumentation Item-Nonresponse die „Datenqualität“ senken oder erhöhen: Auf der einen Seite kann die „Datenqualität“ sinken, da aufgrund der geringeren Anzahl an gültigen Fällen die Standardfehler steigen. Auf der anderen Seite kann die „Datenqualität“ auch steigen, da die Befragten aufgrund von Item-Nonresponse weniger Falschantworten geben können.³⁶

Neben einer fehlenden Definition kann auch ein deutliches Defizit in der theoriegeleiteten Prüfung von Wirkbeziehungen zwischen Incentives und der Antwortqualität festgestellt werden. Mithilfe von geprüften Theorien über die Wirkweise von Incentives könnten Vorhersagen abgeleitet und die bisher nicht erklärte Heterogenität bezüglich der Befunde verschiedener Studien erklärt werden.

³⁴ Es bleibt allerdings unklar, ob dies für oder gegen die „Datenqualität“ spricht.

³⁵ Leider führen die Autoren keine Erklärung für dieses Ergebnis an.

³⁶ Wie bereits beschrieben weisen die Antwortmöglichkeiten „weiß nicht“ oder „Keine Angabe“ auch eine inhaltliche Ebene auf und können daher bei Item-Nonresponse aufgrund unzureichender Antwortkategorien eigentlich nicht gewählt werden. Eine Antwort, welche nicht zu den vorgegebenen Antwortkategorien passt ist etwas anderes als Unwissen oder Antwortverweigerung.

4. Theorie

Aufgrund der in Kapitel 3.1.4 aufgezeigten Defizite in der aktuellen Forschung soll in dieser Studie ein umfassender Beitrag zur Erklärung der Wirkung von Incentives auf die Antwortqualität geleistet werden. Hierfür werden zwei theoretische Ansätze herangezogen, welche aufgrund ihrer thematischen Ausrichtung als besonders geeignet erscheinen: die 1) Cognitive Evaluation Theory und 2) die Reziprozitätshypothese. Im Rahmen der Cognitive Evaluation Theory wird unterstellt, dass unter bestimmten Voraussetzungen eine Belohnung die intrinsische Motivation korrumpieren kann und somit die sorgfältige Bearbeitung eines Fragebogens beeinträchtigt. Im Rahmen der Reziprozitätshypothese wird hingegen angenommen, dass sich die Befragten aufgrund des Erhalts eines Incentives in einem asymmetrischen Schuldverhältnis sehen, welches durch ein sorgfältiges Bearbeiten eines Fragebogens aufgelöst werden kann. Diese zwei Ansätze sollen im Folgenden genauer erläutert werden.

4.1 Cognitive Evaluation Theory

Die Anfänge der Cognitive Evaluation Theory (auch CET) werden auf Deci (1971) zurückgeführt, welcher in seinem Artikel *Externally Mediated Rewards And Intrinsic Motivation* (1971) auf mögliche negative Folgen von Belohnungen verweist.³⁷ Die Cognitive Evaluation Theory wird heute als Subtheorie der Self-Determination Theory angesehen (vgl. Deci & Ryan (1985)), welche sich mit der Erklärung zur Entstehung und Aufrechterhaltung verschiedener Arten von Motivation beschäftigt. Der Begriff der Motivation wird daher für die Cognitive Evaluation Theory aus der Self-Determination Theory von Deci & Ryan (1985) übernommen, bei welcher zwischen drei Arten von Motivation unterschieden wird (vgl. Deci & Ryan (1985;1993;2000)),

³⁷ Die Grundidee, dass Belohnungen einen Einfluss auf die Wahrnehmung oder das Verhalten haben können wurde von anderen Autoren bereits vor Deci (1971) formuliert (vgl. Festinger (1957), DeCharms (1968)).

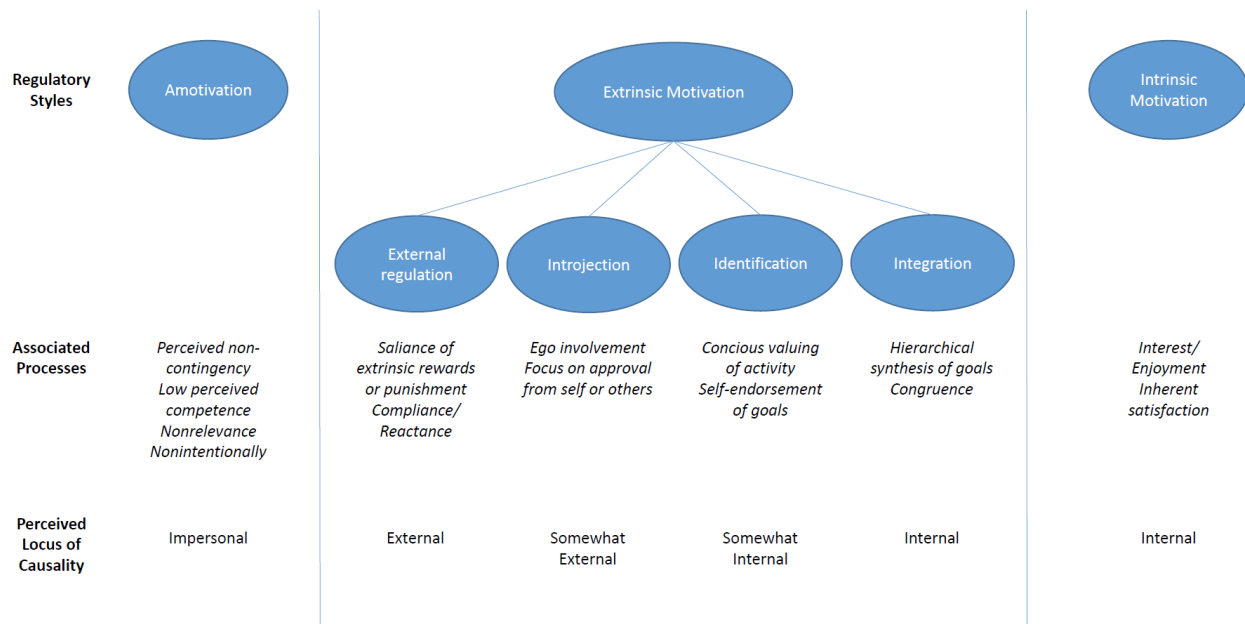
Ryan & Deci (2000)): die Amotivation, die extrinsische Motivation und die intrinsische Motivation.³⁸ Zur Amotivation „gehören z.B. Verhaltensweisen, die kein erkennbares Ziel verfolgen (z.B. dösen, herumlungern), oder die einem unkontrollierten Handlungsimpuls folgen (z.B. Wutanfall). Auch die amotivierten Verhaltensweisen sind energetisiert und psychologisch erklärbar. Aber wir bezeichnen sie nicht als motiviert, weil sie nicht durch intentionale Prozesse gesteuert werden“ (Deci & Ryan (1993), S. 224). Die Amotivation kann nach Deci & Ryan (2000) aus drei Gründen erfolgen: „Amotivation results from not valuing an activity (Ryan, 1995), not feeling competent to do it (Deci, 1975), or not believing it will yield a desired outcome (Seligman, 1975)“ (Deci & Ryan (2000a), S. 61).³⁹

Die extrinsische Motivation wird von Deci & Ryan (1985) in vier Komponenten unterteilt: 1) externale Regulation, 2) introjezierte Regulation, 3) identifizierte Regulation und 4) integrierte Regulation. Die vier Komponenten der extrinsischen Motivation werden dabei durch den jeweiligen Grad an Regulation und Internalisierung unterschieden:

³⁸ Es wird von den Autoren nicht zwischen einer initiiierenden und durchführungsbezogenen Handlungsmotivation unterschieden. Dies bedeutet, dass z.B. die intrinsische Motivation eine Handlung initiieren und beständig aufrecht erhalten kann.

³⁹ Die Amotivation kann, wie auch die folgenden vorgestellten Arten der Motivation auch als Erwartung x Wert - Funktionen verstanden werden. Die Unterscheidung der einzelnen Kategorien kann dann durch verschiedene Erwartungswahrscheinlichkeiten und Wertigkeiten vorgenommen werden.

Abb. 13: Verschiedenen Arten der Motivationen nach Ryan & Deci



Quelle: eigene Darstellung, nach Ryan & Deci (2000a), S. 61.

„External reguliertes Verhalten ist zwar intentional, aber von äußeren Anregungs- und Steuerungsfaktoren abhängig. Es entspricht weder dem Prinzipien der Autonomie noch der Freiwilligkeit“ (Deci & Ryan (1993), S. 227). Dabei gilt zu beachten: „Individuals typically experience externally regulated behavior as controlled or alienated, and their actions have an external perceived locus of causality“⁴⁰ (Deci & Ryan (2000b), S. 70). Die extern regulierte Motivation ist, wie in Abb. 13 dargestellt, neben der Amotivation verortet, da die Amotivation aufgrund externer Einflüsse in eine extrinsische Motivation überführt werden kann. Weniger extrinsisch, aber immer noch sehr stark external wirken die introjezierten Regulationen: „(...) sie beziehen sich auf Ereignisse, die für die Selbstachtung relevant sind. Man tut etwas, „weil es sich gehört“ oder „weil man sonst ein schlechtes Gewissen hätte“ (Deci & Ryan (1993), S. 227). Hier ist im Vergleich zur externalen Regulation ein höherer Grad an Internalisierung gegeben:

⁴⁰ Der Locus of Causality wird von den Autoren vom Locus of Control unterschieden: The term locus of control refers to whether people believe that outcomes are controllable, in other words whether outcomes are believed to be contingent on one's behavior. (...) Locus of causality, on the other hand, refers to the perceived source of initiation and regulation of behavior“. (Deci & Ryan (1985), S. 166)

„Introjected regulation is, of course, more stable than external regulation because it does not require the presence of external contingencies“ (Deci & Ryan (2000), S. 136). Noch stärker ist der Grad der Internalisierung bei der identifizierten Regulation: „Man tut etwas nicht einfach deshalb, weil man das Gefühl hat, es tun zu sollen, sondern weil man es für wichtig hält. Diese persönliche Relevanz resultiert daraus, daß man sich mit den zugrunde liegenden Werten und Zielen identifiziert und sie in das individuelle Selbstkonzept integriert hat“ (Deci & Ryan (1993), S. 228). Hierin kann eine Vorstufe zur Entwicklung einer intrinsischen Motivation gesehen werden, da Tätigkeiten aus einer eigenen Überzeugung heraus begangen werden. Noch deutlicher ist die Nähe zur intrinsischen Motivation bei der integrierten Regulation: „Sie ist das Ergebnis der Integration von Zielen, Normen und Handlungsstrategien, mit denen sich das Individuum identifiziert und die es in das kohärente Selbstkonzept integriert hat“ (Deci & Ryan (1993), S. 228). Obwohl diese Darstellung schon als intrinsische Motivation definiert werden könnte, wird hier von Deci & Ryan (1993, 2000) noch unterschieden: „Der Unterschied ist, daß intrinsisch motivierte Verhaltensweisen autotelischer Natur sind, während integriertes (extrinsisches) Verhalten eine instrumentelle Funktion besitzt, aber freiwillig ausgeführt wird, weil das individuelle Selbst das Handlungsergebnis subjektiv hoch bewertet“⁴¹ (Deci & Ryan (1993), S. 228).

Die intrinsische Motivation wird wie folgt definiert: „When people are intrinsically motivated, they experience interest and enjoyment, they feel competent and self-determining, they perceive the locus of causality for their behavior to be internal, and in some instances they experience flow“ (Deci & Ryan (1985), S. 34). Die umfangreiche Definition der intrinsischen Motivation wird von Deci & Ryan (1985) in einer kürzeren Arbeitsdefinition zusammengefasst: „(...)

⁴¹ Der Begriff autotelisch bedeutet, „dass mit der a. Aktivität kein sekundärer Zweck erreicht wird, sondern allein der Ablauf dieser Aktivität selbst das Ziel ist.“ (Bergius (2014), S. 242)

we infer intrinsic motivation for an activity when a person does the activity in the absence of a reward contingency or control“ (Deci & Ryan (1985), S. 34). Die intrinsische Motivation entspringt damit originär den Wünschen und Interessen einer Person.⁴²

Die vorgestellten Konzeptionen zu den drei Motivationsarten sind gemäß Ryan & Deci (2000a) nicht als kategorial abgrenzend zu verstehen, sondern als kontinuierlicher Verlauf: „Thought of as a continuum, the concept of internalization describes how one’s motivation for behavior can range from amotivation or unwillingness, to passive compliance, to active personal commitment. With increasing internalization (and its associated sense of personal commitment) come greater persistence, more positive self-perceptions, and better quality of engagement“ (Ryan & Deci (2000a), S. 60f.). Sie betonen dabei, dass nicht jede Regulation der extrinsischen Motivation auch durchlaufen werden muss, um eine intrinsische Motivation aufzubauen.

Die oben beschriebenen Definitionen der Motivationen wurden, wie auch die Cognitive Evaluation Theory, von Deci angepasst und erweitert. So wird der Verlust an intrinsischer Motivation aufgrund externer Belohnungen von Deci (1971) noch wie folgt erklärt: „(...) it is suggested that money is frequently used as a means of “buying“ services which would probably not otherwise be rendered. Perhaps, then, the presence of money as an external reward suggests to the subjects that they “should probably not render this activity without pay,” that is, they should not be so intrinsically motivated to do the activity“ (Deci (1971), S. 107). Vierzehn Jahre später wenden Deci und Ryan (1985) einen differenzierteren Ansatz an, welcher auch in dieser Studie zugrunde gelegt wird:⁴³ „CET proposes that underlying intrinsic motivation are

⁴² Deci & Ryan (1985) sehen eine Unabhängigkeit von der intrinsischen Motivation zu den sogenannten Drives (Trieben): „The evidence is indisputable that intrinsic motivation exists and that it involves non-tissue-based, drive-independent needs.“ (Deci & Ryan (1985), S. 39)

⁴³ Die Cognitive Evaluation Theory wird von Deci bereits 1975 ausführlich dargestellt, jedoch gab es auch hier inhaltliche Umgestaltungen, so dass auf die gemeinsame Arbeit von Deci & Ryan (1985) verwiesen wird.

the innate psychological need for competence and self-determination. According to the theory, the effects on intrinsic motivation of external events such as the offering of rewards, the delivery of evaluations, the setting of deadlines, and other motivational inputs are a function of how these events influence a person's perception of competence and self-determination. Events that decrease perceived self-determination (i.e., that lead to a more external perceived locus of causality) will undermine intrinsic motivation, whereas those that increase perceived self-determination (i.e., that lead to a more internal perceived locus of causality) will enhance intrinsic motivation" (Deci et al. (2001a), S. 3).⁴⁴ Diese zwei Mechanismen, "need for competence" und "self-determination" weisen nach Deci et al. (1999) – je nach Belohnungsgrundlage – unterschiedliche Wirkungen auf die intrinsische Motivation auf.

Abb. 14: Die Wirkung von Belohnungen auf die intrinsische Motivation

<i>Task non-contingent</i>	<i>Task-contingent: Engagement-contingent</i>	<i>Task-contingent: Completion-contingent</i>	<i>Task-contingent: Performance-contingent</i>
(Belohnung unabhängig von Aufgabe und Leistung)	(Belohnung abhängig von dem Beginn der Aufgabe)	(Belohnung abhängig von der Beendigung der Aufgabe)	(Belohnung abhängig von der qualitativen bzw. quantitativen Leistung)
Keine Wirkung auf die intrinsische Motivation	Schwache Wirkung auf die intrinsische Motivation	Mittlere Wirkung auf die intrinsische Motivation	Hohe Wirkung auf die intrinsische Motivation

Quelle: eigene Darstellung, nach Deci et al. (1999), S. 628.

Den Autoren zufolge weist beispielhaft eine Belohnung, welche Performance-contingent vergeben wird, einen starken negativen Effekt auf die intrinsische Motivation auf. Dies erklärt

⁴⁴ Es ist hierbei deutlich hervorzuheben, dass die intrinsische Motivation durch die Vergabe von Belohnungen auch gesteigert werden kann. Deci (1975), sowie Deci und Ryan (1985) führen diesen zwar Aspekt aus, weisen aber dennoch zumeist auf einen negativen Effekt von Belohnungen hin.

sich daraus, dass die Durchführung der Aufgabe weniger als selbstbestimmt wahrgenommen und damit eher der Belohnung zugesprochen wird. Dem zufolge sinkt die intrinsische Motivation, da sich der wahrgenommene Locus of Causality verschiebt bzw. verschieben kann. Eine Belohnung, welche hingegen unabhängig von der Aufgabe und Leistung (Task non-contingent) vergeben wird, sollte hingegen keinen Effekt aufweisen: „The explanation for why task-contingent rewards are more detrimental to intrinsic motivation than task-noncontingent rewards is based in the degree of control conveyed by the two types of rewards. If someone must complete a task to get a reward, the task is more likely to be seen as instrumental to the reward. The task is something one must do to get the reward, so the reward is more controlling than a task noncontingent reward, which one gets independent of whether one finishes the task“ (Deci & Ryan (1985), S. 77).

Zusammengefasst bedeutet dies, dass die Wirkung eines Incentives von der Wahrnehmung und Bewertung dessen abhängt. Wird ein Incentive als verhaltenskontrollierend wahrgenommen, so ist gemäß Deci & Ryan (1985) davon auszugehen, dass die intrinsische Motivation sinkt und daraufhin Regulatoren der extrinsischen Motivation wirken. Dies würde zu einem Verlust an Ich-Bezug führen, was sich wiederum negativ auf die Performance auswirkt. Wird ein Incentive hingegen als bestätigend und unterstützend wahrgenommen, dann kann die intrinsische Motivation steigen. Wird das Incentive weder als kontrollierend noch bestätigend wahrgenommen, dann sollte die intrinsische Motivation unberührt bleiben.

4.1.1 Forschungsstand zur Cognitive Evaluation Theory

Zu Beginn soll zuerst auf die widersprüchlichen und inkonsistenten Befunde der bisher durchgeführten Studien aufmerksam gemacht werden. Die Heterogenität der Befunde spiegelt sich dabei auch in den durchgeführten Meta Analysen wider (vgl. Cameron & Pierce (2002); Deci

& Ryan (2001); Cameron et al. (2001); Deci et al. (1999); Tang & Hall (1995); Cameron & Pierce (1994); Wiersma (1992); Rummel & Feinberg (1988)). Die teilweise sehr deutlich auftretenden Unterschiede in den Ergebnissen und Schlussfolgerungen der Meta Analysen wurden jedoch zumeist als Resultat von Fehlklassifikationen gewertet.⁴⁵ Dies führte zu sieben verschiedenen Meta Analysen in 10 Jahren, wobei stets darauf verwiesen wurde, dass dabei auch neuere Studien berücksichtigt wurden. Dies kann jedoch nicht darüber hinwegtäuschen, dass die Aussagen der Meta Analysen dennoch teilweise auf sehr wenigen Studien basieren. So liegen z.B. in der letzten Meta Analyse von Cameron & Pierce (2002) nur zwei Veröffentlichungen vor, welche die Wirkung von unkonditionalen Belohnungen (task non-contingent) thematisieren. Bei genauerer Betrachtung der beiden Studien fallen zudem Einschränkungen auf, welche die Aussagekraft schmälern: In der Studie von Pallak et al. (1982) wird der Motivationsverlauf von Kindergartenkindern untersucht. Hierfür wurde in der Experimentalgruppe ein Video gezeigt, bei dem ein Kind für das Ausmalen von Bildern von einem Forscher beschenkt wird. In der Kontrollgruppe wurde das Video vorher gestoppt, so dass die Szene über die Beschenkung nicht gezeigt wurde. Nach dem Video wurde gemessen wie lange die Kinder freiwillig an (von den Forschern verteilten Bildern) malten. Hier muss jedoch kritisch aufgezeigt werden, dass eine Verallgemeinerung der Ergebnisse aufgrund der Eigenschaften der Probanden (= Kindergartenkinder) und der geringen Fallzahlen ($n_{\text{Experimentalgruppe}} = 15$ und $n_{\text{Kontrollgruppe}} = 12$) mit starken Unsicherheiten verknüpft ist. In den beiden Experimenten von Pretty & Seligman (1984) wurden Studierende an einem College gebeten ein SOMA-Puzzle⁴⁶ zu lösen. Es wurde nach der einmaligen Übergabe einer Belohnung in der Experimentalgruppe gemessen, wie lange

⁴⁵ Die Debatten werden hierbei teilweise sehr emotional geführt. Dies zeigt sich beispielhaft an Kritik von Lepper et al. (1996) an der Meta Analyse von Cameron et al. (1994): „(...) the sorts of procedures employed by Cameron and Pierce could be used to turn silk purses into sows' ears“ (Lepper et al. (1996), S. 26).

⁴⁶ Das SOMA-Puzzle besteht aus sieben verschieden geformten Bauklötzen, welche zu vorgegebenen Formen zusammengesetzt werden sollen. Die Bauklötze sind hierbei so gestaltet, dass keines identisch ist.

sich die Probanden in einer folgenden Wartezeit mit dem Puzzle freiwillig weiter beschäftigten oder alternative Tätigkeiten (z.B. Zeitschriften lesen) aufnahmen. Cameron & Pierce (2001) kritisieren hierbei die häufig genutzte Messung der intrinsischen Motivation über die freiwillig aufgewendete Zeit und damit die zentralen Interpretationen der Ergebnisse: „Consider the fact that individuals who were promised a reward for solving puzzles or drawing pictures did more of this activity in the reward phase than the control group, which was not promised a reward. When placed in the free-time setting, the rewarded participants could now distribute their time to the experimental task or the others activities. Because they had spent more time doing the experimental task in the reward phase, the alternative activities in the choice phase would tend to become more attractive. That is, the experimental participants would choose the alternate activities rather than spend more time on puzzle solving or drawing“ (Cameron & Pierce (2001), S. 27). Damit wäre eine kürzere Beschäftigung mit einem SOMA-Puzzle nicht der Belohnung zuzuschreiben, sondern einem Sättigungseffekt. Die Probanden haben sich ausführlich mit der Aufgabe beschäftigt, so dass sie in der freien Arbeitsphase weniger Zeit darauf verwenden. Deci & Ryan (1985) scheint dieses Problem bewusst zu sein, auch wenn sie in ihren Veröffentlichungen nicht direkt darauf eingehen. Vielmehr schlagen sie eine diffuse Erweiterung des Messkataloges vor: „(...) we sometimes look at the quality of performance or of outcomes as indicators of intrinsic motivation“ (Deci & Ryan (1985), S. 35).

4.1.2 Hypothesen zur Cognitive Evaluation Theory

Da die oben aufgezeigten empirischen Befunde zur Stützung der Cognitive Evaluation Theory unzureichend erscheinen und der theoretische Aufbau damit nicht als empirisch abgesichert gelten kann, soll in dieser Studie ein zentraler Aspekt der Cognitive Evaluation Theory untersucht werden: der Mechanismus des Locus of Causality. Hierfür werden Hypothesen bezüglich

(unerwarteter) Task non-contingent Belohnungen formuliert, welche – gemäß der Cognitive Evaluation Theory – keine Effekte auf die intrinsische Motivation aufweisen dürfen, da zum einen das Erhalten der Belohnung keine Zielvorgabe darstellt, und zum anderen das Gefühl der Autonomie bei den Befragten unberührt bleiben sollte. Der oft erwartete Korrumpierungseffekt einer Belohnung auf die intrinsische Motivation sollte daher ausbleiben.

H1a: Der Erhalt eines unkontingenten Incentive hat keinen Einfluss auf die Höhe der intrinsischen Motivation.

Da Deci & Ryan (1985) argumentieren, dass eine hohe intrinsische Motivation mit einer intensiven Auseinandersetzung einer Aufgabe zusammenhängt. Daraus wird die Schlussfolgerung gezogen, dass – unabhängig von einer Belohnung – mit steigender intrinsischer Motivation auch die Antwortqualität steigt, da die gestellten Aufgaben vertiefter bearbeitet werden.

H1b: Je höher die intrinsische Motivation, desto höher die Antwortqualität.

4.2 Reziprozitätshypothese

Die Idee von Reziprozität bzw. reziprokem Verhalten besagt, dass ein positiv wahrgenommenes Geschenk zu einer positiven Erwidern und damit Gegengabe des Empfängers führen kann. Nach Stegbauer können vier Arten von Reziprozität näher bestimmt werden: a) die direkte Reziprozität, b) die generalisierte Reziprozität, c) die Reziprozität der Positionen und d) die Reziprozität der Perspektiven⁴⁷ (vgl. Stegbauer (2011), S. 29). Bei der direkten Reziprozität

⁴⁷ Auf die Reziprozität der Positionen und der Perspektiven wird in dieser Arbeit nicht genauer eingegangen. Prinzipiell fasst Stegbauer diese wie folgt zusammengefasst: „Die dritte und vierte Bedeutung hängt eng miteinander zusammen. Reziprozität von Positionen meint, dass beispielsweise in Rollensystemen eine bestimmte Rolle immer gleichzeitig auch einen Gegenpart besitzt. Die Rollen passen ineinander, ja eine bestimmte Rolle

wird nach Überreichung des Geschenks die Erwidierung in Form einer gleichwertigen Gabe erwartet. Unter generalisierter Reziprozität versteht Stegbauer: „Generalisierte Reziprozität ist eine Leistung, die erbracht wird, ohne auf einen direkten Ausgleich hoffen zu können. Dieser Begriff wird häufig in Verbindung mit Gruppenzugehörigkeit gebraucht. Gruppenzugehörigkeit kann in diesem Zusammenhang sowohl eine konkrete Gruppe (Familie, Sippe, Verein, Gleichaltrigengruppe etc.), als auch eine soziologische Gruppe (etwa die Gruppe der Arbeiter, die der Fernfahrer oder der Ärzte) meinen, die selbst untereinander gar nicht in Kontakt stehen. Die ersten beiden Bedeutungen stellen den Austausch selbst in das Zentrum“ (Stegbauer (2011), S. 29). Hierbei muss beachtet werden: „Die Erwartung der Reziprozität ist unbestimmt. Gewöhnlich hängen Zeit und Wert der Rückgabe nicht nur davon ab, was der Geber gegeben hat, sondern auch davon, was er braucht und wann dies der Fall ist, ebenso davon, was der Empfänger aufbringen kann und zu welchem Zeitpunkt dies möglich ist. Der Empfang von Gütern beinhaltet die unbestimmte Verpflichtung, die Gabe zurückzuerstatten, wenn der Geber sie benötigt und/oder dies dem Empfänger möglich ist. Die Abgeltung kann daher sehr bald oder auch niemals erfolgen“ (Sahlins (2005), S. 82).⁴⁸

mag ohne die andere gar nicht denkbar sein, man denke etwa an den Arzt, der ohne Patient kaum seiner ärztlichen Position gerecht werden kann. Die Gegenseitigkeit des Verhaltens entspricht nicht einer klaren Definition einer bestimmten Gabe und einem entsprechenden Äquivalent, sondern der Austausch ist durch die spezifische Position der Beteiligten reguliert. Nicht immer werden die Partner als gleichberechtigt angesehen, oft finden sich in solchen Rollensystemen Hierarchien, die eine wertorientierte Äquivalenz der Tauschgüter oder -leistungen ausschließen. Diese Bedeutung von Reziprozität sagt also etwas darüber aus, was legitime Tauschgüter für die Beteiligten sind, und wie diese, abhängig von den Positionen bewertet werden“ (Stegbauer (2011), S. 29ff.).

⁴⁸ Stegbauer untergliedert die generalisierte Reziprozität noch in zwei Komponenten: „eine Generalisierung über einen längeren Zeitraum und über eine bestimmte Gruppe, der man sich zugehörig fühlt“ (Stegbauer (2011), S. 67).

In der Überblicksliteratur wird ferner zwischen zwei theoretischen Strömungen unterschieden: den sogenannten normativistisch und individualistisch orientierten Theorieansätzen (vgl. Adloff & Mau (2005), Ekeh (1974)).⁴⁹

Innerhalb der als normativistisch bezeichneten Theorieansätze wird auf die gesellschaftliche Wirkung und Bedeutung der Reziprozität verwiesen. Gouldner (1960) formuliert hierbei: „A norm of reciprocity is, I suspect, no less universal and important than the incest taboo, although, similarly, its concrete formulations may vary with time and place“ (Gouldner (1960), S. 171). Damit wird sie als allgegenwärtig formuliert und folgt einem einfachen Mechanismus: „Specifically, I suggest that a norm of reciprocity, in its universal form makes two interrelated, minimal demands: (1) people should help those who have helped them, and (2) people should not injure those who have helped them“ (Gouldner (1960), S. 171).

Bei den individualistisch orientierten Theorieansätzen steht das Individuum mit seinen Entscheidungen im Vordergrund. So überrascht es nicht, dass viele Werke auf der Idee des Rational Choice basieren, bei welcher die Reziprozitätsnorm als Teil des Entscheidungsprozesses modelliert werden kann (vgl. Coleman (1991), Falk & Fischbacher (2000)). Falk & Fischbacher (2000) stellen ein umfangreiches Konzept zur Reziprozität zur Verfügung⁵⁰. Der Begriff der Reziprozität wird bei den Autoren jedoch nicht direkt definiert. Perugini et al. (2003) stellen fest: „Despite the amount of work on reciprocal behavior, less effort has been devoted to providing a clear theoretical definition of what reciprocity is. In most of these studies reciprocity has

⁴⁹ Diese Trennung wird von anderen Disziplinen (z.B. Wirtschaftswissenschaften) nicht vorgenommen und dient in den Sozialwissenschaften primär um (vermeintlich neue) integrative Konzepte von individualistischer und normativer Reziprozität vorzustellen. So kann eine Norm (mittlerweile) im Rational Choice Ansatz integriert und dargestellt werden.

⁵⁰ Im Rahmen der Spieltheorie wird üblicherweise von einer altruistischen Reziprozität gesprochen: „Altruistisch ist diese Form der Reziprozität (...), wenn andere Personen von der Revanche profitieren.“ (Diekmann (2008), S. 14f.) Der Begriff der Revanche bezieht sich auf die Wiederholungen innerhalb spieltheoretischer Experimente, d.h. die Probanden wiederholen das Szenario und können dabei auf positive und negative Vorerfahrungen in die Entscheidungen einfließen lassen

been defined as a strategy applicable to repeated interaction, mainly described as repeated social dilemma (or two-person mixed motive) games“ (Perugini et al. (2003), S. 252). Den Autoren zufolge kann zwischen einer positiven und negativen Reziprozität unterschieden werden. Darüber hinaus definieren sie drei Komponenten, welche sie als Voraussetzungen für ein positives reziprokes Verhalten ansehen:

„First, positive reciprocators are expected to be particularly prone to react to positive interpersonal behaviours whereas negative reciprocators are expected to be particularly reactive to negative ones. Therefore, whereas the former should be especially sensitive to kind behavior, the latter should be especially sensitive to unkind behaviour.

Second, positive reciprocators are expected to be particularly willing to perform positive behaviours, or to deliver positive sanctions, following the other’s positive action (e.g. be kind with someone if the other is kind to you) whereas negative reciprocators should be especially willing to perform negative behaviours, or to negatively sanction, when receiving negative behaviours from the other (e.g. retaliate against someone who has behaved negatively towards you). (...)

The third feature has to do with how and when a behaviour is perceived to be fair. Fairness is an elusive concept that can be achieved in different ways, for instance by splitting endowments equally (equality), by balancing out inputs and outputs in the transaction (equity), or by reciprocating in kind (reciprocity)“ (Perugini et al. (2003), S. 255).⁵¹

Werden die Darstellungen der Reziprozitätshypothese zusammengefasst, so ergeben sich für die Wirkung von Incentives folgende Vorhersagen: Wird ein unkonditionales Incentive positiv bewertet und liegt zudem eine stark verinnerlichte Reziprozitätsnorm vor, dann kann dies zu einem positiven reziproken Verhalten führen. Ein solches positives Verhalten könnte im Rahmen einer Befragung zu einer erhöhten Antwortqualität führen. Wird die Gabe hingegen als

⁵¹ Die Autoren nutzen, genau wie auch Falk & Fischbacher (2000) die theoretischen Grundlagen von Gouldner (1960) als Ausgangsposition.

negativ bewertet, so kann dies zu einem negativen reziproken Verhalten führen, was sich beispielhaft in einer eher geringen Antwortqualität niederschlagen kann.

4.2.1 Forschungsstand zur Reziprozitätshypothese

In den Sozialwissenschaften ist die Idee der Reziprozität über verschiedene theoretische Ansätze erklärt worden (z.B. Coleman (1991); Blau (1964); Mauss (1923)), jedoch gibt es aktuell nur wenige empirische sozialwissenschaftliche Studien, welche reziprokes Verhalten als empirischen Forschungsgegenstand thematisieren (Wolbring & Hellmann (2010); Wolbring & Groß (2009); Voswinkel (2005))⁵² und es ist damit nicht verwunderlich, dass nur wenige Operationalisierungen für reziprokes Verhalten vorliegen. In den Wirtschaftswissenschaften wird versucht, reziprokes Verhalten mithilfe spieltheoretischer Ansätze nachzuweisen (vgl. Dohmen et al. (2009); Falk (2003)). Hierfür wurden vorzugsweise trust games⁵³ oder public goods games genutzt. Die Ergebnisse der Studien können zentral zusammengefasst werden: „Being positively reciprocal predicts higher work effort, lower unemployment and also higher subjective well-being“ (Dohmen et al. (2009), S. 609). Diese Spiele sind größtenteils Verlaufsspiele, also mit Wiederholung, dennoch weist Falk (2003) darauf hin, „daß Reziprozität auch in nicht-wiederholten Spielen auftritt und deshalb etwas anderes ist, als „kooperative“ Gleichge-

⁵² In der Sozialpsychologie ist die Studie von Regan (1971) sehr prominent. Hierbei konnte aufgezeigt werden, dass reziprokes Verhalten auch experimentell herbeigeführt werden kann. Hierfür wurde an die Probanden der Experimentalgruppe kostenlose Erfrischungen ausgegeben und wenig später Lose zum Verkauf angeboten. Die Probanden der Kontrollgruppe bekamen keine Getränke angeboten, konnten später allerdings auch Lose für eine Tombola erwerben. In der Experimentalgruppe wurde fast die doppelte Anzahl an Losen im Vergleich zur Kontrollgruppe verkauft. Dieser Effekt wird auf die Wirkung der Reziprozitätshypothese zurückgeführt.

⁵³ Hierunter zählt z.B. das Spiel „gift exchange“: „Das Experiment simuliert einen Arbeitsmarkt, wobei die Probanden die Rolle von Arbeitnehmern und Arbeitgebern einnehmen. Jede von mehreren Spielrunden mit wechselnden Partnern besteht aus zwei Phasen. In der ersten Phase können die Arbeitnehmer in einer Auktion einen Kontrakt mit einer bestimmten Lohnhöhe ersteigern. In der zweiten Phase bestimmen die Arbeitnehmer ihren Arbeitseinsatz. Je höher der Einsatz, desto geringer die Auszahlung an die Arbeitnehmer und desto höher der Gewinn der Arbeitgeber. Hoher Arbeitseinsatz entspricht positiver, altruistischer Reziprozität, falls ein großzügiges Lohnangebot vorlag.“ (Diekmann (2008), S. 12)

wichtsstrategien in wiederholten Spielen“ (Falk (2003), S. 142). Die Erkenntnisse der spieltheoretischen Experimente werden auch in den Sozialwissenschaften wahrgenommen und diskutiert. Diekmann & Voss (2003) sehen in den spieltheoretischen Ansätzen ein großes Potential für die Sozialwissenschaften und empfehlen daher eine verstärkte Anwendung (vgl. Diekmann (2008); Diekmann & Voss (2003)).⁵⁴ Die in den Spielen zugrunde liegende Laborsituation muss an dieser Stelle jedoch auch problematisiert werden: „While experimental evidence has shown the behavioural relevance of reciprocity in laboratory settings, there is less evidence for the role of reciprocity outside the laboratory“ (Dohmen et al. (2009), S. 592). Damit wird deutlich, dass die externe Validität aufgrund der künstlichen Situation eingeschränkt sein kann. Darüber hinaus werden für die spieltheoretischen Experimente oft Studierende herangezogen, so dass auch daraus Limitationen der Verallgemeinerbarkeit der Ergebnisse ergeben können.

4.2.2 Hypothesen zur Reziprozität

Aus den oben vorgestellten Ansätzen wird abgeleitet, dass ein unkonditionales Incentive einen positiven Einfluss auf die Reziprozität aufweist. Es wird hierbei vermutet, dass der Erhalt des Incentives die verinnerlichte Reziprozitätsnorm aktiviert und somit zu Reziprozität führen kann. Hierfür muss allerdings unterstellt werden, dass ein Incentive auch positiv bewertet wird. Da eine Bewertung in den späteren Analysen nur von Empfängern eines Incentives vorgenommen werden kann, werden die Hypothesen über einen Proxi formuliert: die Bewertung der Person, welche das Incentive übergab. Diese Person wird allen Personen bekannt sein und

⁵⁴ Auch in dieser Studie werden spieltheoretische Erkenntnisse und Messinstrumente genutzt. Die Möglichkeit zur Prüfung der Wirkung von Incentives auf die Antwortqualität in einer authentischen Erhebungssituation überweg hierbei der Konzeption eines spieltheoretischen Experiments.

es wird angenommen, dass eine positive Bewertung des Incentives zu einer positiven Bewertung des Belohnungsgebers führt. Der daraus entstehende (positive) Reziprozitätsdruck kann von den Befragten über eine generalisierte Reziprozitätsleistung aufgelöst werden, d.h. über eine ungleiche, aber angemessene Gegengabe. Eine solche Gegengabe kann im Rahmen der Befragung eine erhöhte Sorgfalt bei der Bearbeitung des Fragebogens darstellen und sich folglich in einer erhöhten Antwortqualität widerspiegeln. Aus diesen Annahmen ergeben sich folgende Hypothesen:

H2a: Je höher das Incentive, desto stärker wirkt die verinnerlichte Reziprozitätsnorm positiv auf die Antwortqualität.⁵⁵

H2b: Je höher das Incentive, desto positiver wird der Belohnungsgeber wahrgenommen.

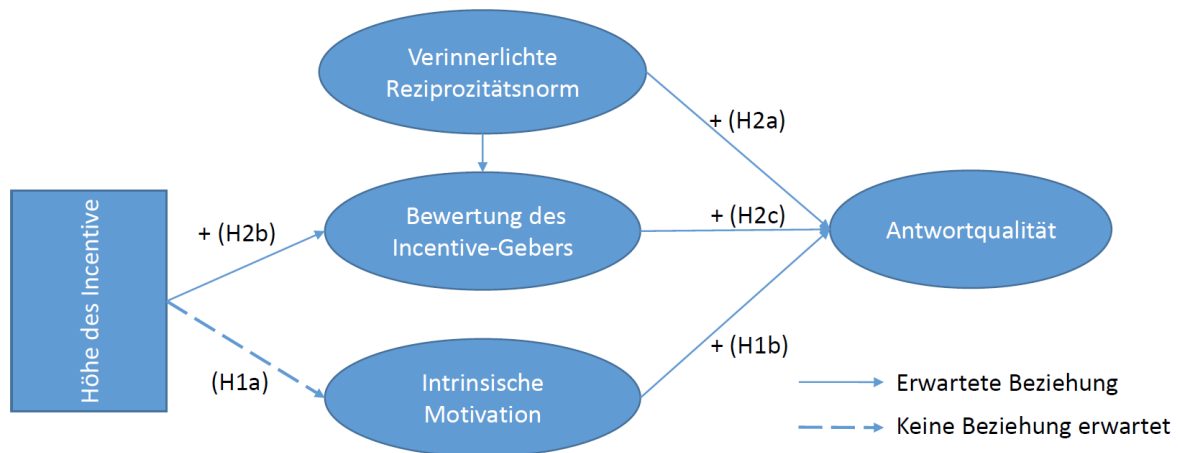
H2c: Je positiver der Belohnungsgeber wahrgenommen wird, desto höher die Antwortqualität.

Die formulierten Hypothesen zur Wirkung der intrinsischen Motivation und der Reziprozität lassen sich auch graphisch in einem Strukturmodell zusammenfassen⁵⁶:

⁵⁵ Dies ist eine Interaktionshypothese und postuliert eine Variation der Effektstärken bei unterschiedlichem Incentivierungsgrad. Dies wird in den späteren Hypothesenprüfungen berücksichtigt, auch wenn es in den folgenden Grafiken additiv dargestellt wird.

⁵⁶ Es wird zusätzlich ein gerichteter Zusammenhang zwischen der verinnerlichteten Reziprozitätsnorm und der Bewertung des Incentive-Gebers erwartet. Dies begründet sich darin, dass sich bei aktivierter Reziprozitätsnorm Reziprozität auch im Sinne einer Beziehungseröffnung manifestieren kann (vgl. Stegbauer (2011), S. 93ff).

Abb. 15: Die Darstellung der Zusammenhänge der intrinsischen Motivation und der Reziprozität auf die Antwortqualität



Quelle: eigene Darstellung.

4.3 Zusammenführung der Cognitive Evaluation Theory und der Reziprozitätshypothese

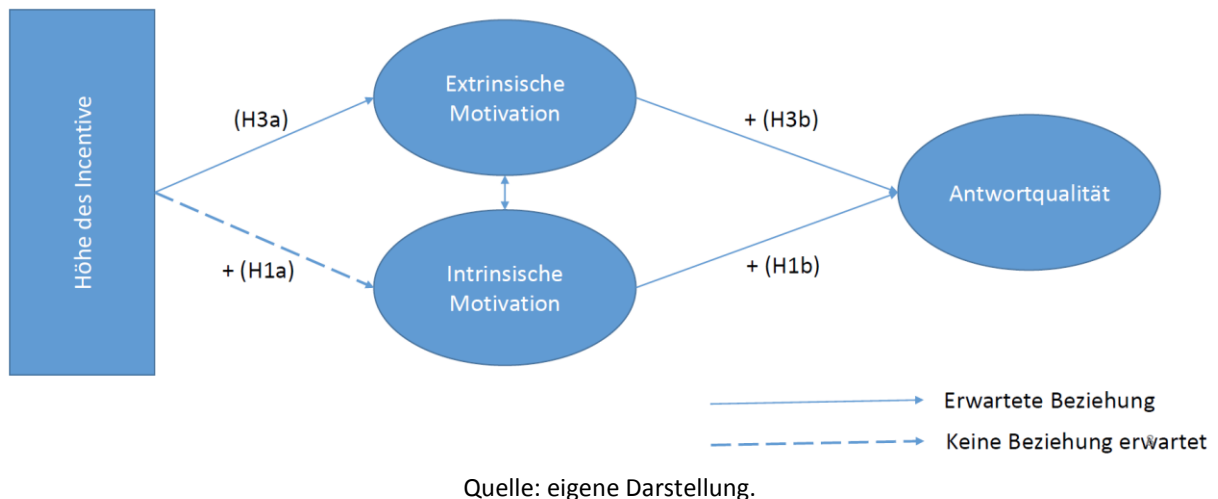
Gemäß der Cognitive Evaluation Theory ist eine Handlung aufgrund des Befolgens einer Norm nicht intrinsisch, da extern regulierte Mechanismen vorliegen. Die motivierende Wirkung der Reziprozitätsnorm wird damit der extrinsischen Motivation zugewiesen. Deci & Ryan (1985) unterscheiden innerhalb der extrinsischen Motivation zwischen vier Regulationen (externe Regulation, introjezierte Regulation, identifizierte Regulation und integrierte Regulation), welche sich in dem Grad der Internalisierung und Handlungszuschreibung (Locus of Causality) unterscheiden (vgl. Abb. 14). Die identifizierte Regulation umfasst Deci & Ryan (1985) zufolge normkonformes Handeln, wenn die Norm als wichtig und persönlich bedeutsam erachtet wird. Demzufolge kann ein Incentive dann zu Reziprozität führen, wenn sich die Empfänger bewusst für ein normkonformes Verhalten entscheiden und somit eine positive Wirkung auf die Antwortqualität entfalten. Auch für diese Annahmen werden Hypothesen formuliert:

H3a: Je höher das unktionale Incentive, desto höher die extrinsische Motivation.

H3b: Je höher die extrinsische Motivation, desto höher die Antwortqualität.

Die Hypothesen werden auch hier graphisch in einem Strukturmodell dargestellt.

Abb. 16: Die Darstellung der Zusammenhänge der intrinsischen und der extrinsischen Motivation auf die Antwortqualität



5. Aufbau der Studie

Im vorherigen Kapitel wurden Kausalhypothesen formuliert, welche im Rahmen dieser Arbeit geprüft werden sollen. Damit die Ergebnisse später auch als intern und extern valide gelten können, muss das Design der Studie auf die Fragestellung zugeschnitten sein. In diesem Kapitel wird daher der Aufbau dieser Studie ausführlich dargestellt und erläutert.

Es wurde in dieser Studie ein experimentelles Design genutzt, da dieses zur Prüfung von Kausalhypothesen sehr geeignet erscheint (vgl. Kühnel & Dingelstedt (2014); Shadish et al. (2002)). Shadish et al. (2002) unterscheiden hierbei zwischen drei bzw. zwei Arten von Experimenten:

Abb. 17: Die drei Arten eines Experiments nach Shadish et al.

„Rando- mized Expe- riment:	An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.
Quasi-Expe- riment:	An experiment in which units are not assigned to conditions randomly.
Natural Ex- periment:	Not really an experiment because the cause cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition.“

Quelle: eigene Darstellung, nach Shadish et al. (2002), S. 12.

Das randomisierte Experiment weist nach Shadish et al. (2002) die höchste interne Validität auf, da die Veränderung der abhängigen Variable auf die Variation der unabhängigen Variable zurückgeführt werden kann.

Auch in diese Studie wurde daher ein randomisiertes Experiment durchgeführt. Hierfür muss zum einen das Treatment (Incentive) in den Versuchsgruppen gezielt manipuliert werden und zum anderen die Teilnehmer mithilfe eines Randomisierungsverfahrens einer der drei Versuchsgruppen zugeordnet werden. Die genaue Umsetzung wird nun folgend beschrieben:

a) Beschreibung der Versuchsgruppen und des Treatments

Im Rahmen dieser Forschungsarbeit werden drei Versuchsgruppen unterschieden: In der ersten Versuchsgruppe wurde kein Incentive ausgehändigt, diese entspricht damit der Kontrollgruppe des Experiments. In der zweiten Versuchsgruppe wurde ein unkonditionales Incentive von 5 Euro an die Probanden ausgegeben. Dieser Betrag wurde gewählt, da ab dieser Höhe in

Studien zur Teilnahmebereitschaft zumeist ein positiver Effekt von Incentives festgestellt werden konnte.⁵⁷ Auf dieser Grundlage wurde vermutet, dass ein Incentive in Höhe von 5 Euro ausreicht, um eine positive Wirkung auf die Antwortqualität zu entfalten. In der dritten Versuchsgruppe beträgt die Höhe des Incentives 20 Euro. Das Incentive wurde auf diesen Betrag festgelegt, da die Ergebnisse der Studie von Barge & Gehlbach (2012) auf mögliche negative Effekte bezüglich der Antwortqualität bei einem hochwertigen Incentive (15\$, bzw. 20\$) aufmerksam machen. Ein solcher Befund widerspricht den im vierten Kapitel formulierten theoretischen Annahmen zur positiven Wirkung von Incentives, da diese demnach nicht monoton mit Höhe des Incentives steigt. In dieser Studie sollte daher ebenfalls ein höheres monetäres Incentives eingesetzt werden, um ggf. negativ auftretende Effekte auf die Antwortqualität erfassen zu können.

Die Entscheidung für die Vergabe von unkonditionalen Incentives erklärt sich aus der geplanten Prüfung der Reziprozitätshypothese. Im Rahmen der Reziprozitätshypothese wird angenommen, dass eine unkonditionale Gabe beim Empfänger Reziprozitätsdruck auslöst und dies zu einer (generalisierten) Gegengabe führen kann. Ein konditionales Incentive widerspricht dieser Konzeption, da dann die Gegengabe vor der eigentlichen Gabe stattfinden müsste. Zur Prüfung der Reziprozitätshypothese muss daher das Incentive unkonditional ausgegeben werden. Die Vergabe unkonditionaler Incentives sollte, gemäß der Cognitive Evaluation Theory, keinen Einfluss auf die intrinsische Motivation aufweisen. Aufgrund der geringen Studienlage zu task non-contingent Belohnungen wird es auch als wichtig betrachtet diese postulierte Unabhängigkeit zu prüfen.

⁵⁷ Es konnten zur Bestimmung des Betrags leider keine Studien zur Antwortqualität herangezogen werden, da einheitlichen Befunde zur Wirkung unter Kontrolle der Höhe des Incentives fehlen. Aus diesen Gründen wird auf die Teilnahmebereitschaft verwiesen. Darüber hinaus wird ein Incentive in Höhe von 5 Euro oft in Studien eingesetzt und erscheint als Testhöhe für die Versuchsgruppe besonders relevant.

Desweiteren wurde festgelegt, dass die Probanden erst bei Teilnahme an der Studie von dem un konditionalen Incentive erfahren sollten (vorausgesetzt sie sind in einer entsprechenden Versuchsgruppe). Dieses Vorgehen wurde gewählt, damit sich die Befragten unter gleichen Vorinformationen zur Teilnahme einfinden und unterschiedliche Effekte bezüglich der Antwortqualität allein auf das jeweilige Treatment der zugeordneten Versuchsgruppe zurückzuführen sind. Eine Vorinformation über eine Incentivierung könnte einen zusätzlichen Stimulus darstellen und damit die interne Validität einschränken.

b) Beschreibung der Versuchspersonenanwerbung

Die Zielgruppe dieses Experiment waren alle zum Zeitpunkt der Studie immatrikulierten Studierenden der Georg-August-Universität Göttingen. In die Stichprobe einbezogen werden konnten allerdings nur die Studierenden, welche zum Sommersemester 2015 über einen gültigen Studierendenaccount des StudIT⁵⁸ verfügten und damit über eine personalisierte eMail-Adresse (bestehend aus dem Vor- und Nachnamen) erreichbar waren. Diese Einschränkung lag darin begründet, da keine umfassenden eMail-Adresslisten aller Studierenden für die Kontaktaufnahme zur Verfügung standen und deshalb auf onomastische Verfahren zurückgegriffen werden musste, um eMail-Adressen zu generieren⁵⁹. Aufgrund dieser hohen potentiellen Probandenzahl, wurde die Fallzahl für die Versuchsgruppen auf 100 festgelegt, so dass insgesamt 300 Studierende befragt werden sollten, bei einer Erhebungsdauervon vier Wochen.

⁵⁸ Für Informationen zum StudIT siehe: studit.uni-goettingen.de/

⁵⁹ An dieser Stelle möchte ich mich bei Stephan Schlosser für seine umfangreiche Unterstützung bei der eMail-Generierung bedanken. Aus den häufigsten deutschen und englischen Vor- und Nachnamen konnte ein Datensatz mit eMail-Adressen generiert werden, wobei sich der Anteil der nutzbaren Adressen auf ca. 20.000 belief. Somit konnten theoretisch fast zwei Drittel aller Studierenden der Universität (= 28320) erreicht werden.

Es wurde während der konzeptionellen Phase des Designs erwogen die Wohnbevölkerung der Stadt Göttingen anstatt der Studierenden als potentielle Probanden heranzuziehen. Dieser Gedanke musste jedoch aufgrund begrenzter zeitlicher und finanzieller Ressourcen aufgegeben werden. Es soll an dieser Stelle aber hervorgehoben werden, dass der Fokus auf die Studierenden neben der Erreichbarkeit auch noch über ein weiteres Argument gestützt werden kann: ein großer Teil der universitären Forschung stützt sich auf studentische Probanden und es erscheint daher umso wichtiger für diese Gruppe die Konsequenzen einer Incentivierung auf die Antwortqualität aufzudecken.

Mit Beginn des Sommersemesters 2015 wurden innerhalb der ersten zwei Vorlesungswochen die Teilnahmeeinladungen⁶⁰ an alle generierten eMail-Adressen versandt. Der Einladungstext war für alle Studierenden gleich und umfasste, neben dem Erhebungszeitraum⁶¹, dem Erhebungsort⁶² und den Teilnahmezeiten⁶³, auch einen Hinweis über eine durchschnittliche Bearbeitungsdauer von ca. 27 Minuten. Dies erschien notwendig, damit die Probanden eine längere Befragungszeit einplanen und somit Abbrüche aufgrund terminlicher Verpflichtungen vermieden werden sollten.⁶⁴

⁶⁰ Der Einladungstext ist im Anhang wiedergegeben, S. 202.

⁶¹ Ursprünglich wurde eine Erhebungsphase von 4 Wochen eingeplant.

⁶² Der Erhebungsort war der CIP-Pool der sozialwissenschaftlichen Fakultät im Oeconomicum. Dieser Raum wurde freundlicherweise für die Erhebung zur Verfügung gestellt.

⁶³ Eine Teilnahme war innerhalb der Erhebungsphase immer montags bis freitags, von 15:00 Uhr bis 20:00 Uhr möglich. Die Teilnahmezeiten ergaben sich aus der Verfügbarkeit des CIP-Pools, da dieser bis 15:00 Uhr von den Studierenden noch regulär genutzt wurde.

⁶⁴ Aufgrund des im Anschreiben formulierten Teilnahmeaufwands (längere Befragungsdauer und zurückzulegende Wegstrecke zum Befragungsort) wird davon ausgegangen, dass die teilnehmenden Studierenden eine hohe intrinsische Motivation aufweisen. Dies ist bei der Interpretation der Ergebnisse zu beachten.

c) Umsetzung des Versuchs

Bei Erscheinen der Probanden am Erhebungsort wurde ein kurzes Vorbereitungsgespräch in einem Besprechungszimmer geführt. In diesem wurden die Studierenden gebeten ihr Mobiltelefon auszuschalten und während der Erhebung nicht mit anderen Personen zu sprechen. In diesem Kontext wurde auch erfragt, inwiefern bei den Probanden Zeitdruck aufgrund von Folgeterminen besteht.⁶⁵ Dann wurden die Probanden, unter Nutzung einer Liste mit Zufallszahlen, zu einer der drei Versuchsgruppen zugeordnet und somit ggf. ein Incentive überreicht. Nach dem Vorgespräch nahmen die Probanden im angegliederten Erhebungslabor an einem Computer Assisted Self-Interview (CASI) teil und füllten somit an einem PC eigenständig eine standardisierte Befragung aus.⁶⁶ Für die Befragung wurde ein Fragebogen entwickelt und konstruiert⁶⁷, auf dessen Grundlage die in Kapitel 2.4 dargestellten vier verschiedenen Facetten der Antwortqualität⁶⁸ gemessen werden sollten. Die Studie wurde offiziell als Befragung zum Thema „Persönlichkeit & Identität“ angekündigt, da die Nennung des wahren Forschungsthemas zu Verzerrungen und Problemen der externen Validität führen kann.⁶⁹ Die Operationalisierungen zur Messung der einzelnen Facetten der Antwortqualität können dem sechsten Kapitel entnommen werden. Es soll an dieser Stelle erwähnt werden, dass der Fragebogen bewusst so konzipiert wurde, dass ein hoher kognitiver Aufwand zur Beantwortung aller Fragen von Nöten ist. Dies wurde angestrebt, damit eine erhöhte Trennschärfe zur Messung zu den Facetten der Antwortqualität erreicht werden kann.

⁶⁵ Wurde Zeit, bzw. Termindruck berichtet, so wurden die Befragten gebeten zu einem späteren Zeitpunkt an der Befragung teilzunehmen.

⁶⁶ In dem CIP-Pool sind 15 PCs, so dass auch mehrere Probanden zeitgleich in diesem Raum an der Studie teilnehmen konnten. Die Anzahl der anwesenden Studierenden wurde mithilfe einer Rückfrage im Fragebogen kontrolliert.

⁶⁷ Zur Programmierung wurde Unipark® von Questback genutzt.

⁶⁸ Die Facetten für die Antwortqualität lauten: durchdacht, (situational) wahrheitsgemäß, vollständig und anweisungsbefolgend.

⁶⁹ Beispielhaft könnten Selbst-Selektionseffekte auftreten oder die Probanden weisen aufgrund der Kenntnis des wahren Themas eine besonders hohe Antwortqualität auf.

Der Fragebogen ist vor Anwendung in der Studie mithilfe mehrerer Pretest-Verfahren auf inhaltliche Fehler und technische Probleme geprüft worden. Es sind zum einen Expert-Reviews durchgeführt worden, bei welchen über ausgewählte Frageformulierungen sowie den Gesamtablauf des Fragebogens diskutiert wurde. Darüber hinaus kamen für ausgewählte Fragen auch kognitive Pretestverfahren mit Studierenden⁷⁰ zum Einsatz. Zuletzt wurden, ebenfalls mit Studierenden, Durchführungstests für den gesamten Fragebogen vorgenommen, um eventuell auftretende technische oder konzeptionelle Probleme aufzudecken (z.B. Darstellungsfehler aufgrund unterschiedlicher Auflösungen oder Probleme bei der Filterführung).⁷¹ Aufgrund dieses Vorgehens konnten viele Fehler (z.B. Tippfehler bei Frageformulierungen) korrigiert und der reibungslose Ablauf während der Erhebung gesichert werden.

Am Ende der Befragung wurden die Befragten in Einzelgesprächen kurz über den Inhalt der Studie aufgeklärt und gebeten Verschwiegenheit über die Inhalte der Befragung sowie dem ggf. erhaltenen Incentive zu wahren. Dies wurde als besonders wichtig angesehen und im Gespräch betont, da die Verbreitung der Information über eine mögliche Belohnung das gesamte Design korrumpieren und damit die Ergebnisse verfälschen kann.

d) Rekrutierung der Teilnehmer

Die Erhebungsphase begann mit dem ersten Versenden der Einladungen am 13.04.2015. Nachdem Einladungstexte an alle generierten eMail-Adressen versandt waren, wurde schnell deutlich, dass die ursprüngliche Planung bezüglich der Teilnehmerquoten zu optimistisch war und nicht beibehalten werden konnte: in den ersten drei Wochen erschienen nur 21 Studierende, um an der Studie teilzunehmen. Aufgrund dessen wurde die Fallzahl von 100 auf 60

⁷⁰ Die hierfür ausgewählten Studierenden gehörten auch der Universität Göttingen an und konnten demnach in der Haupterhebung nicht mehr befragt werden.

⁷¹ Die Teilnehmer an den Pretests wurde stets gebeten keine Informationen bezüglich der getesteten Fragebogenbereiche an dritte Personen weiterzugeben.

Probanden pro Versuchsgruppe reduziert und Maßnahmen zur Steigerung der Teilnahmebereitschaft überlegt. Eine Maßnahme bestand darin, dass ab der vierten Erhebungswoche Rekrutierungsscouts zum Einsatz kamen, welche auf zentralen Plätzen der Universität Flyer über die Studie verteilten.⁷² Die Scouts wurden hierfür genau instruiert, um bei Rückfragen von Studierenden nur auch der eMail zu entnehmende Informationen wiederzugeben.⁷³ Darüber hinaus wurden auch Flyer in Seminarräumen und Vorlesungssälen ausgelegt. Der Anstieg der Teilnahme in der vierten Woche kann folglich auf diese erweiterten Rekrutierungsstrategien zurückgeführt werden.

Abb. 18: Die Darstellung der Rücklaufquote unter Berücksichtigung der Einführung alternativer Rekrutierungsstrategien.⁷⁴

Erhebungswoche	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Teilnehmer	9	6	6	20	2	26	20	31	13	17	9	5	12	10
	Anschreiben													
					Scouts verteilen Flyer									
						Erinnerung								
										Werbung in Veranstaltungen				

Quelle: eigene Darstellung.

Ab der sechsten Woche wurden Erinnerungsschreiben per eMail an die Studierenden versandt, was ebenfalls zu einer Steigerung der Teilnahmebereitschaft führte. Aufgrund erneut sinkender Teilnahmezahlen wurde in der zehnten und elften Woche zusätzlich in Großveranstaltungen verschiedener Fakultäten der Universität für die Studie geworben (z.B. in der „Vor-

⁷² Der Flyertext ist im Anhang wiedergegeben, S. 204.

⁷³ Für den Fall das die Scouts nach dem persönlichen Grund für das Verteilen der Flyer gefragt werden, sollten diese darauf Antworten: „Es ist ein Studentenjob“. Damit sollte vermieden werden, dass die Studierenden den Scouts zuliebe an der Studie teilnehmen. Um die Effekte der neuen Rekrutierungsmaßnahme abschätzen zu können, wurde der Fragebogen ganz am Ende der Befragung um eine letzte Seite erweitert: Es wurde zusätzlich erhoben aufgrund welcher Rekrutierungsstrategie die Studierenden an dieser Studie teilnahmen.

⁷⁴ In den Erhebungszeitraum fielen auch einige gesetzliche Feiertage, so dass sich auch daraus Schwankungen der Teilnahmebereitschaft erklären lassen: am Donnerstag der fünften Woche war „Christi-Himmelfahrt“ und damit der darauf folgende Freitag ein oft genutzter „Brückentag“. In der siebten Woche lag der Pfingstmontag.

lesung in die Volkswirtschaftslehre“ der Fakultät für Argarwissenschaften oder in der „Systematische Überblicksvorlesung zum Thema “Religionsstifter““ der Theologischen Fakultät). Am 14.07.2015 (und damit zum Ende der Vorlesungszeit) konnte die Minimalstichprobe von 180 gültig befragten Probanden erreicht werden.⁷⁵

e) Beschreibung der realisierten Stichprobe

Im Folgenden werden die Teilnehmer dieser Studie genauer beschrieben: Das Durchschnittsalter der Probanden beträgt 24.2 Jahre. Der Anteil an weiblichen Befragten ist mit 57.2% leicht höher als der Anteil der männlichen Befragten (38.9%).⁷⁶

Da es sich bei den Probanden um Studierende handelt, werden nun folgend Verteilungsinformationen zum Studium dargestellt:

Tab. 4: Aktuell angestrebter Studienabschluss

	Häufigkeit	Gültige Prozent
Bachelor	118	67.8
Master	44	25.3
Magister	3	1.7
Staatsexamen	4	2.3
kirchliches Examen	1	0.6
Promotion	4	2.3
Gesamtsumme	174	100

Quelle: eigene Daten.

Die Probanden geben zumeist den Bachelor als angestrebten Abschluss an. Die durchschnittliche Anzahl an Hochschulsesemestern liegt dabei bei den Bachelor-Studierenden bei 4.64 und sie befinden sich somit tendenziell im fortgeschrittenen Studium. Die Master-Studierenden geben einen durchschnittlichen Wert von 10 Hochschulsesemestern an und befinden sich damit

⁷⁵ Gesamt wurden 186 Personen befragt, wobei sechs Befragte aufgrund unsystematischer Ausfälle (Konnektivitätsprobleme mit dem Internet, gesundheitliche Beschwerden) nicht in den Analysen berücksichtigt werden. Innerhalb der gültigen Fälle gibt es nur eine Person, welche den Fragebogen aus inhaltlichen Gründen abgebrochen hat. Diese Person ist der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro zugeordnet.

⁷⁶ 3.9% (7 Befragte) verweigerten die Angabe des Geschlechts.

– gemessen an der Regelstudienzeit – eher am Ende des Studiums.⁷⁷ Die Studierenden verteilen sich dabei wie folgt auf die Fakultäten:

Tab. 5: Die Aufteilung der befragten Studierenden nach Fakultäten

	Antworten		Prozent der Fälle
	Häufigkeit	Prozent	
Fakultät für Agrarwissenschaften	7	3.3%	3.9%
Fakultät für Biologie und Psychologie	7	3.3%	3.9%
Fakultät für Chemie	2	0.9%	1.1%
Fakultät für Forstwissenschaften und Waldökologie	3	1.4%	1.7%
Fakultät für Geowissenschaften und Geographie	8	3.7%	4.5%
Fakultät für Mathematik und Informatik	2	0.9%	1.1%
Fakultät für Physik	2	0.9%	1.1%
Juristische Fakultät	5	2.3%	2.8%
Philosophische Fakultät	42	19.6%	23.5%
Sozialwissenschaftliche Fakultät	102	47.7%	57.0%
Theologische Fakultät	4	1.9%	2.2%
Universitätsmedizin	1	0.5%	0.6%
Wirtschaftswissenschaftliche Fakultät	29	13.6%	16.2%
Gesamtsumme	214	100.0%	119.6%

Quelle: eigene Daten. Es waren Mehrfachantworten möglich.

Es ist ersichtlich, dass Studierende der Sozialwissenschaftlichen Fakultät überproportional häufig vertreten sind. Dies kann zum einen auf die hohe Anzahl an ausgelegten Flyern in den Räumlichkeiten der sozialwissenschaftlichen Fakultät zurückgeführt werden und zum anderen auf die Bekanntheit des Versuchsleiters aufgrund von Lehrveranstaltungen. Bei Rückfrage gaben 33.3% aller Befragten an, die für die Studie verantwortliche Person zu kennen.⁷⁸

Da in dieser Studie die Wirkung monetärer Incentives auf die Antwortqualität geprüft werden soll, wurde auch der finanzielle Hintergrund der Probanden erfragt: Das durchschnittliche im Monat zur Verfügung stehende Geld (inklusive Miete usw.) liegt bei 726.5€. Um die Höhe des durchschnittlichen Betrags einschätzen zu können, soll dieser mit einem Referenzwert vergli-

⁷⁷ Die durchschnittlichen Hochschulsemeister der anderen angestrebten Abschlüsse lauten: Magister = 5.67; Staatsexamen = 8.00; kirchliches Examen = 8.00 und Promotion = 16.25.

⁷⁸ Die Studierenden sind in diesem Merkmal unsystematisch über die drei Erhebungsgruppen verteilt, folglich liegen keine signifikanten Unterschiede zwischen den Gruppen vor.

chen werden. Die Universität Göttingen nennt hierfür geschätzte monatliche Lebenshaltungskosten von 767.05€⁷⁹ und weist damit einen höheren Wert aus. Dennoch geben die Befragten auf einer 7er-Skala an, tendenziell mit dem monatlich zur Verfügung stehenden Geld zufrieden zu sein (Mittelwert = 4.84). Die genutzten Quellen zur Finanzierung des Studiums sind hierbei:

Tab. 6: Finanzierungsquellen während des Studiums.

	Antworten		Prozent der Fälle
	Häufigkeit	Prozent	
Eltern	135	34.6%	75.0%
andere Verwandte	25	6.4%	13.9%
eigene Erwerbstätigkeit	94	24.1%	52.2%
BaföG	56	14.4%	31.1%
eigene Ersparnisse	58	14.9%	32.2%
Studienkredit	9	2.3%	5.0%
Stipendium	10	2.6%	5.6%
Erbschaft	3	0.8%	1.7%
Sonstiges, und zwar: ⁸⁰	17	4.2%	9.4%
Gesamtsumme	390	100.0%	216.7%

Quelle: eigene Daten. Es waren Mehrfachantworten möglich.

Mit 34.6% werden die Eltern als häufigste Finanzierungsquelle genannt. Die zweithäufigste Finanzierung ist mit 24.1% die eigene Erwerbstätigkeit.

Die bisher dargestellten Informationen über die Befragten können zu einem Profil zusammengefasst werden: Die Teilnehmer sind im Verlauf des Studiums tendenziell fortgeschritten und weisen zumeist einen sozial- oder geisteswissenschaftlichen Hintergrund auf. Das ihnen monatlich zur Verfügung stehende Geld wird als angemessen empfunden, wobei vor allem die Eltern als Finanzierungsquelle genannt werden.

⁷⁹ Die geschätzten Lebenshaltungskosten für Juni 2015 können folgender Homepage entnommen werden: <http://www.studieren-in-goettingen.de/wohnen-essen-finanzen/lebenshaltungskosten-goettingen.html> [letzter Zugriff: 30.09.2015]

⁸⁰ Unter der Kategorie *Sonstiges* wurde der Erhalt von Kindergeld, Waisenrente, Wohngeld oder die Unterstützung von Personen ohne Verwandtschaftsbeziehung (z.B. Freunde oder Lebenspartner) genannt.

6. Operationalisierung

Damit eine Prüfung der im vierten Kapitel beschriebenen Hypothesen erfolgen kann, werden zuerst die zu messenden Konstrukte operationalisiert. Der Fragebogen wurde bei der Konzeption in drei Bereiche untergliedert, welche den Befragten zu Beginn auch kurz vorgestellt wurden:

- Persönlichkeit und Identität
- Gesundheit und Demographie
- Feedback zur Umfrage

Die drei Abschnitte umfassen jeweils eine ungefähr gleiche Anzahl an Fragebogenseiten und dienen zum einen der Strukturierung und zum anderen als Begründung für die verschiedenen eingesetzten Skalen⁸¹ und Fragewiederholungen.

In diesem Kapitel werden die Operationalisierungen für die unabhängigen (Kapitel 6.1) und abhängigen Variablen (Kapitel 6.2) vorgestellt.

6.1 Die Messung der unabhängigen Variablen

6.1.1 Intrinsische Motivation

Zur Messung der intrinsischen Motivation wird auf vier verschiedene Operationalisierungen zurückgegriffen (Selbsteinstufungen auf vorgegebenen Skalen, die Bearbeitungszeiten, die Wiederbefragungsbereitschaft und die Anzahl an Worten in einer freiwilligen Zusatzfrage), welche von Deci & Ryan entweder selbst genutzt oder zumindest vorgeschlagen wurden.

⁸¹ Die Antwortskalen der ordinalen Items weisen üblicherweise folgende Kodierung auf: 1 = *Trifft überhaupt nicht zu* und 7 = *Trifft voll zu*. Abweichungen werden explizit erwähnt.

a) *Selbsteinstufungen auf Basis der Situational Motivation Scale*⁸²

Unter Anwendung der Situational Motivation Scale (SIMS) soll die intrinsische Motivation der Befragten auf Basis von Selbsteinstufungen gemessen werden. Hierfür liegen vier Items vor, welche direkt zu Beginn der Befragung gestellt wurden.⁸³ Die Formulierung des ersten Items wurde für diese Befragung leicht angepasst, damit es auf den Befragungskontext angewendet werden kann.

Motivation: Warum nehmen Sie gerade an dieser Umfrage teil?	
Intrinsische Motivation	Weil ich denke, dass die Umfrage interessant ist.
	Weil ich es für angenehm halte.
	Weil es mir Spaß macht.
	Weil ich mich gut fühle, sobald ich daran teilnehme.

b) *Bearbeitungszeit*

Neben der Selbsteinstufung soll auch die aufgewendete Zeit während der Befragung herangezogen werden, da nach Deci & Ryan (1985) die Dauer der freiwilligen Auseinandersetzung als Indikator für die intrinsische Motivation gesehen werden kann.⁸⁴ Sie

⁸² Die SIMS wurde von Deci bisher nicht als Messinstrument eingesetzt, erscheint aber aufgrund der Konzeption und Konstruktvalidität (vgl. Guay et al (2000)) als angemessen zur Messung der intrinsischen Motivation bei Bearbeitung von Umfragen. Sie umfasst 16 Items, wobei jeweils vier verschiedenen Motivationstypen zugeordnet werden können: Amotivation, externale Regulation, identifizierte Regulation und intrinsische Motivation. In diesem Fragebogen wurde die gesamte Situational Motivation Scale erhoben, aber nur die relevanten Motivationsdimensionen werden für die Analysen herangezogen. Die SIMS liegt in englischer Version vor (Guay et al. (2000)) und wurden in der Übersetzung von Vogt (2004) übernommen und für die Befragungssituation angepasst.

⁸³ Diese Items stehen nicht direkt hintereinander, da innerhalb der Situational Motivation Scale die Items zur Messungen der vier Teildimensionen alternierend dargestellt werden.

⁸⁴ Üblicherweise wird die Bearbeitungszeit als direkter Maßstab für die Antwortqualität genutzt. Dies ist jedoch nicht unproblematisch: Im Rahmen der Theorie des kognitiven Antwortprozesses (vgl. Tourangeau & Rasinski (1988)) kann argumentiert werden, dass bei einem Durchlaufen des vollständigen Prozesses mehr Zeit für das Beantworten von Fragebogenfragen benötigt wird, als bei einem unvollständigen oder abgekürzten Vorgehen. Demgemäß kann eine niedrige/ hohe Bearbeitungszeit für eine niedrige/ hohe Antwortqualität sprechen. Jedoch kann, mit Verweis auf die Editierungsmöglichkeiten der Befragten eine hohe Bearbeitungszeit mit einer niedrigen Antwortqualität verbunden sein. Als stabile Annahme kann in dieser Hinsicht nur erhalten bleiben, dass der betriebene kognitive Aufwand mit der Zeit zusammenhängen kann, jedoch unabhängig von der Richtigkeit der genannten Antworten.

gehen folglich davon aus, dass mit steigender Bearbeitungszeit auch mehr Interesse an der Durchführung einer Tätigkeit gezeigt wird und somit eine erhöhte intrinsische Motivation vorliegt.⁸⁵

Im Rahmen der vorliegenden Studie wurden Bearbeitungszeiten für jede dargestellte Fragebogenseite gemessen. Die Zeitmessung wurde hierbei im Bereich der Millisekunden erfasst und gespeichert. Dabei wurde die Ladezeit (aufgrund der Internetverbindung) kontrolliert, so dass die reine Bearbeitungszeit gemessen werden konnte. Da die Bearbeitungszeit jedoch auch die Lesezeit inkludiert, kann nicht von einer Antwortreaktionszeit gesprochen werden. „Typischerweise wird die Antwortreaktionszeit als das Zeitintervall zwischen der Präsentation eines Fragestimulus und der Initialisierung der Reaktion einer Person auf diesen Stimulus definiert“ (Mayerl & Urban (2008), S. 8).⁸⁶ Die Erhebung der Bearbeitungszeit erscheint als Messung dennoch angemessen, da bei Messung der intrinsischen Motivation die Länge der freiwilligen Auseinandersetzung mit einer Tätigkeit im Vordergrund steht und nicht die Reaktionszeit.

⁸⁵ Die Teilnahme an dieser Studie wurde daran geknüpft, dass die Probanden nicht unter Zeitdruck stehen. Die Bearbeitungszeit erscheint damit auch in dieser Studie als zulässiger Indikator für die intrinsische Motivation. Es ist darauf hinzuweisen, dass eine Person die Befragung unter Angabe von terminlichen Gründen abbrach. In diesem Fall lagen allerdings während der Befragung starke Netzwerkprobleme vor, so dass sich die Bearbeitungszeit künstlich durch lange Ladezeiten zwischen den Fragebogenseiten verlängert hat. Unklar ist deswegen, ob die befragte Person tatsächlich keine Zeit mehr hatte oder dies lediglich als Vorwand nutzte um die unangenehme Befragungssituation zu beenden. Aufgrund der Konnektivitätsprobleme wurden an diesem Tag keine weiteren Befragungen mehr durchgeführt und die Angaben der Person als ungültig betrachtet.

⁸⁶ Dies löst allerdings nicht das Problem, dass die Präsentation der Frage bei den Befragten ein unterschiedliches (kognitives) Ver- und Bearbeiten bewirken kann.

c) Wiederbefragungsbereitschaft

Die Wiederbefragungsbereitschaft kann ebenfalls als Indikator für die intrinsische Motivation angesehen werden. Hierbei wird vermutet, dass sich Probanden eher dazu bereit erklären an einer Folgebefragung teilzunehmen, wenn sie intrinsisch motiviert sind.

Da die Wiederbefragungsbereitschaft die intrinsische Motivation erfassen soll, muss den Probanden aufgezeigt werden, dass bei einer zukünftigen Teilnahme keine Incentives zu erwarten sind. Dies entspricht dem Konzept der freiwilligen Beschäftigungsphase nach Deci und Ryan (1985), nur dass diese auf einen späteren, unbekanntem Zeitpunkt verschoben wird. Es wurde ebenfalls bewusst auf eine genaue Benennung des Wiederbefragungszeitpunkts verzichtet, damit die Probanden nicht aus terminlichen Gründen die Zusage verweigern.

Die Wiederbefragungsbereitschaft wird im letzten Drittel des dritten Blocks⁸⁷ abgefragt und über die folgende Frage operationalisiert:

Für die weitere Forschung ist es sehr wichtig, dass der Entwicklungsverlauf von Persönlichkeit- und Identitätsbildung erfasst wird.

Sie würden die Arbeit daher sehr unterstützen, wenn ich Sie in ein paar Wochen (unentgeltlich) erneut befragen könnte. Die Folgeumfrage wird online erfolgen, das heißt Sie müssen nicht mehr ins Oec kommen und können bequem von zuhause aus die Fragen beantworten.

Sie würden mit einer weiteren Teilnahme wirklich sehr helfen und die Qualität der Arbeit verbessern!

- Ja, ich nehme an einer Folgeumfrage teil.
- Nein, ich nehme nicht an einer Folgeumfrage teil.

Weiter



⁸⁷ Die Wiederbefragungsbereitschaft wird erst am Ende der Befragung erhoben, da den Befragten die potentiellen Konsequenzen einer erneuten Teilnahme verdeutlicht werden. Damit sollten die Befragten mit einer geringeren intrinsischen Motivation eine Wiederbefragung eher verneinen.

Die Befragung läuft nach dieser Frage regulär weiter, bis das Ende des dritten Blockes erreicht ist. Die Befragten, welche vorher eine positive Antwort bezüglich der Wiederbefragungsbereitschaft gegeben haben, werden erst dann aufgefordert eine eMail-Adresse zu hinterlassen. Hierbei wird den Befragten zugleich eine Möglichkeit zum Widerruf der ursprünglichen Zusage gegeben:

Sie haben ein paar Fragen vorher angegeben, dass Sie gern an der Folgebefragung teilnehmen möchten: Vielen Dank!

Damit Sie erneut befragt werden können wird eine gültige eMail-Adresse von Ihnen benötigt.

- Meine eMail-Adresse lautet:
- Ich möchte doch nicht mehr an einer Folgeumfrage teilnehmen.

Weiter



d) Anzahl an Worten in einer freiwilligen Zusatzfrage

Im Rahmen einer Befragung wird von den Teilnehmern erwartet, dass diese den Fragebogen vollständig ausfüllen. Dies ist mit einem mehr oder weniger hohen kognitiven und zeitlichen Aufwand verbunden und kann – dem theoretischen Ansatz von Deci & Ryan (1985) zufolge – bei einer hohen intrinsischen Motivation dazu führen, dass Fragen intensiver bearbeitet werden. So soll in dieser Befragung die Intensität der Bearbeitung als Messung der intrinsischen Motivation genutzt werden. Hierfür wurde kurz vor Ende des dritten Fragebogenabschnitts eine offene Frage gestellt, welche für die vermeintliche Fragestellung „Persönlichkeit und Identität“ keine direkten inhaltlichen Schwerpunkte aufweist und damit als freiwillige Zusatzinformation interpretiert werden kann:

Wenn Sie allgemeine Kritik oder Anregungen haben, dann können Sie diese hier eintragen.

Es wird hierbei unterstellt, dass mit steigender Anzahl an Worten, bzw. Zeichen auch eine höhere intrinsische Motivation einhergeht.⁸⁸

6.1.2 Reziprozität

Gemäß den theoretischen Vorgaben von Perugini et al. (2003) wird zur Prüfung der Reziprozitätshypothese die verinnerlichte Reziprozitätsnorm und die subjektive Bewertung des Versuchsleiters operationalisiert.

a) Die verinnerlichte Reziprozitätsnorm

Zur Messung der verinnerlichten Reziprozitätsnorm wird die Skala zur positiven und negativen Reziprozität von Perugini et al. (2003) genutzt.⁸⁹ Das Konzept der positiven Reziprozität erscheint hierbei zentral, da geprüft werden soll, inwiefern die Befragten ein unterstützendes reziprokes Verhalten aufweisen. Die Items zur Messung der negativen Reziprozität werden als Kontrollvariablen ergänzend erhoben. Die Formulierungen für die Messungen der verinnerlichten Reziprozitätsnorm lauten wie folgt:

⁸⁸ Üblicherweise wird die Anzahl an Worten als Indikator für die „Datenqualität“ genutzt. In dieser Konzeption wird davon abgewichen, da im Gegensatz zu den inhaltlichen Fragen die Rückfrage nach allgemeiner Kritik und Verbesserungsvorschlägen von den Befragten als optional wahrgenommen werden kann. Dies würde dann der Konzeption einer freiwilligen Beschäftigung entsprechen und entspricht damit einem Indikator für die intrinsische Motivation.

⁸⁹ Die Skalen werden in dieser Studie in der Übersetzung des SOEP (2005) verwendet.

Verinnerlichte Reziprozitätsnorm	
Positiv	<p>Wenn mir jemand einen Gefallen tut, bin ich bereit, diesen zu erwidern.</p> <p>Ich strenge mich besonders an, um jemanden zu helfen, der mir früher schon mal geholfen hat.</p> <p>Ich bin bereit, Kosten auf mich zu nehmen, um jemanden zu helfen, der mir früher einmal geholfen hat.</p>
Negativ	<p>Wenn mir schweres Unrecht zuteil wird, werde ich mich um jeden Preis bei der nächsten Gelegenheit rächen.</p> <p>Wenn mich jemand in eine schwierige Lage bringt, werde ich das Gleiche mit ihm machen.</p> <p>Wenn mich jemand beleidigt, werde ich mich im gegenüber auch beleidigend verhalten.</p>

b) Bewertung des Versuchsleiters

Da die Bewertung des Incentives als „fair“ oder „unfair“ von einer vorherigen Vergabe abhängt, können die Befragten ohne Incentive keine Bewertung vornehmen. Dies führt zu dem Problem, dass die Hypothesen nicht global getestet werden können, da keine gruppenübergreifenden Messungen vorliegen. Daher soll die Bewertung des Incentives, wie in Kapitel 4.2.2 dargestellt, über eine Proxi-Variable gemessen werden: die Bewertung des Versuchsleiters. Dies ist möglich, da der Versuchsleiter zum einen mit allen Probanden die Einführungsgespräche durchführte und zum anderen auch die Incentives vergab. Es wird hierbei angenommen, dass eine positive Bewertung des Incentives zu einer positiveren Bewertung des Versuchsleiters führt. Die Bewertung der Person erfolgt über zwei Items:

Wie bewerten Sie die Person, welche für die Umfrage verantwortlich ist? ⁹⁰
Unfreundlich – Freundlich
Nicht Hilfsbereit – Hilfsbereit

⁹⁰ Vor Beginn der Befragung wurden alle Probanden darüber informiert, dass die Person, welche das Einführungsgespräch leitet auch zugleich die verantwortliche Person für diese Umfrage ist. Die Bewertungen der Person werden auf zwei 5stufigen bipolaren Skalen vorgenommen.

6.1.3 Identifizierte Regulation (extrinsische Motivation)

Die identifizierte Regulation ist, gemäß dem Ansatz von Deci & Ryan (1985) eine Teilkomponente der extrinsischen Motivation. Demzufolge bedeutet ein Anstieg an identifizierter Regulation gleichzeitig einen Anstieg der extrinsischen Motivation. Zur Messung der identifizierten Regulation werden die dazugehörigen Items der Situational Intrinsic Motivation Scale (Guay et al (2000), in der Übersetzung von Vogt (2004)) genutzt. Diese Dimension der SIMS wird über vier Items gemessen, deren Formulierungen für die Befragungssituation angepasst wurden:⁹¹

Identifizierte Regulation	Weil ich es zu meinem eigenen Wohl tue. Weil ich die Teilnahme für gut für mich halte. Aus persönlichem Entschluss. Weil ich glaube, dass eine Teilnahme wichtig für mich ist.
------------------------------	---

6.2 Die Messung der abhängigen Variablen: die Indikatoren für Antwortqualität

Zur Messung der Antwortqualität wurden verschiedene Operationalisierungen für die einzelnen definitorischen Komponenten *durchdacht, (situational) wahrheitsgemäß, vollständig und anweisungsbefolgend* ausgewählt. Diese Operationalisierungen werden nun folgend beschrieben.

6.2.1 Indikatoren für ein durchdachtes Bearbeiten eines Fragebogens

a) Akquieszenz und der Status Quo-Effekt als Ergebnis von Satisficing

1) Akquieszenz als Indikator für weak Satisficing

Akquieszenz wird in dieser Befragung über eine umfangreiche Item-Batterie zum Thema Drogen gemessen, wobei diese den Befragten so dargestellt wird, dass sie nach unten

⁹¹ Diese Items stehen nicht in direkter Abfolge, da innerhalb der Situational Motivation Scale die Items zur Messungen der vier Teildimensionen alternierend dargestellt werden.

scrollen müssen, um alle Fragen beantworten zu können.⁹² Diese Darstellungsform soll für die Befragten einen hohen Beantwortungsaufwand verdeutlichen und damit die Trennschärfe der Akquieszenz-Messung erhöhen. Dies unterstützend wurden die Drogen-Items stets so formuliert, dass eine Zustimmung zu den Aussagen durchgängig auf dem rechten Pol der Antwortskala verortet werden kann. Die verwendeten Items lauten wie folgt:⁹³

Akquieszenz
Der Konsum der Droge "Marihuana" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Heroin" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Kokain" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "LSD" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Ecstasy" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Speed" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Crystal" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "LA-42" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Haschisch" führt zu starken gesundheitlichen Problemen.
Der Konsum der Droge "Marihuana" führt zur sozialen Isolation.
Der Konsum der Droge "Heroin" führt zur sozialen Isolation.
Der Konsum der Droge "Kokain" führt zur sozialen Isolation.
Der Konsum der Droge "LSD" führt zur sozialen Isolation.
Der Konsum der Droge "Ecstasy" führt zur sozialen Isolation.
Der Konsum der Droge "Speed" führt zur sozialen Isolation.
Der Konsum der Droge "Crystal" führt zur sozialen Isolation.
Der Konsum der Droge "LA-42" führt zur sozialen Isolation.
Der Konsum der Droge "Haschisch" führt zur sozialen Isolation.

2) Status Quo-Effekt als Indikator für strong Satisficing

Neben der Akquieszenz als Indikator für ein weak Satisficing wird der Status Quo-Effekt als Messung für ein strong Satisficing herangezogen. Nach Krosnick (1991) führt der Status Quo-Effekt dazu, dass die Befragten bei einer Auswahl an konkurrierenden Szenarien jenes präferieren, welches entweder am ehesten ihrer aktuellen Situation entspricht oder die geringste Variation zur aktuellen Situation aufweist. Dies wird dadurch

⁹² Die Befragten müssen hierbei auch ganz nach unten scrollen, da sie nur so den Link für die nächste Fragebogenseite erreichen können.

⁹³ Die hervorgehobenen Drogen existieren nicht, da sie für die Operationalisierung von Pseudo-Opinions erdacht wurden.

begründet, dass ein Durchdenken und Bewerten alternativer Szenarien für die Befragten einen hohen kognitiven Aufwand bedeutet und daher bei Satisficing eher vermieden wird.⁹⁴ Im Fragebogen wird der Status Quo-Effekt über das folgende Item gemessen:

10. Angenommen Sie könnten den Aufbau Ihres Studiengangs beeinflussen. Würden Sie etwas ändern?

- Ja, ich würde folgendes ändern:
- Nein, ich würde alles so belassen.
-

Der Druck zur Wahl des Status Quo (= „Nein, ich würde alles so belassen“) wird aufgrund der gewählten Frageformulierung hierbei erhöht, da eine Abweichung vom Status Quo zusätzlich noch schriftlich begründet werden soll.

b) Offenlegung von Erinnerungsstrategien für Häufigkeitsangaben

Im Rahmen der Befragung wird die Frage gestellt „Wie oft haben Sie in den vergangenen vier Wochen Sport getrieben?“. Zur Beantwortung dieser Frage können verschiedene Erinnerungsstrategien genutzt werden. Hierfür werden auf der Folgeseite eine Auswahl an möglichen Erinnerungsstrategien vorgegeben, aus welchen sich die Befragten die zutreffende, bzw. zutreffenden auswählen können. Die folgenden Antwortstrategien werden den Befragten vorgegeben:

Ich habe an jedes einzelne Mal gedacht, wo ich Sport getrieben habe.

Ich habe daran gedacht wie oft ich normalerweise Sport treibe.

Ich habe an die verschiedenen Sportarten gedacht, die ich getrieben habe.

Ich habe auf Basis meines generellen Eindrucks geschätzt.⁹⁵

⁹⁴ Es soll noch einmal hervorgehoben werden, dass die Wahl des Status Quo nicht per se ein Anzeichen für Satisficing ist, da auch nach reiflicher und reichlicher Überlegung ein alltagsnahes Szenario gewählt werden kann.

⁹⁵ Es sind Mehrfachantworten möglich. Die Fragen wurden hierbei von Medway (2012) übernommen und eigenständig übersetzt. Es werden zur späteren Analyse zwei Kodierungen unterschieden: a) die strenge Definition einer hohen Antwortqualität für die Befragten, welche ausschließlich angeben *an jedes einzelne Mal gedacht zu haben* und b) die weichere Definition einer hohen Antwortqualität für die Befragten, welche auch angeben *an jedes einzelne Mal gedacht zu haben*.

Die Wahl der ersten Erinnerungsstrategie weist, gemäß Conrad et al. (1998), auf einen hohen kognitiven Aufwand hin, da jede einzelne Situation mit sportlicher Aktivität bedacht wird. Die alternativen Erinnerungsstrategien erfolgen primär auf Basis von Schätzungen, bzw. Verallgemeinerungen und implizieren damit einen geringeren kognitiven Aufwand.

c) Konsistenz

Die Messungen von Konsistenzen von Antworten sollen ebenfalls als Indikatoren für eine hohe Antwortqualität genutzt werden. „Dies liegt nahe, da bei einer hohen Qualität des Antwortverhaltens ein einmal gebildetes Urteil wiederholt werden sollte“ (Häder & Kühne (2010), S. 115). Es ist jedoch anzumerken, dass sich die Meinung aufgrund des Bearbeitens des Fragebogens ändern kann. Es liegt dann ein „Reifungseffekt“ vor, welcher von einer Inkonsistenz nicht zu trennen ist. Darüber hinaus kann der Versuch einer Konsistenz, wenn sie von den Befragten angestrebt wird (siehe Kapitel 2.2.3) aufgrund fehlerhafter Rückerinnerungen verhindert werden. Die Herbeiführung von Konsistenz ist daher auch an die Aufmerksamkeit und Erinnerungsleistung gekoppelt. Van Meurs & Saris (1990) sind hierbei der Ansicht, dass ungefähr 20 Minuten ausreichen damit die Erinnerung an Vorfragen stark verblasst und ein konsistentes Beantworten ohne stabile Einstellungswerte erschwert wird.⁹⁶ Aus diesen Argumentationen folgend soll Konsistenz über zwei Ebenen gemessen werden: zum einen über eine kurze zeitliche Abfolge, so dass Inkonsistenzen aufgrund von Reifungs- oder Erinnerungseffekte tendenziell ausgeschlossen werden können und zum anderen über eine lange zeitliche Abfolge, wobei mehr als 25 Minuten Bearbeitungszeit zwischen den zwei Messungen vorliegen sollen. Die Messungen für die beiden Reichweiten werden nun vorgestellt:

⁹⁶ Es ist anzumerken, dass Saris & Gallhofer (2007; 2014) im Rahmen der Aktualisierung des Buches *Design, Evaluation, and Analysis of Questionnaires for Survey Research* die Zeit von 20 Minuten auf 25 Minuten erhöht haben (vgl. S. 220, bzw. S. 209). Eine Begründung wird von den Autoren nicht genannt.

1) *Kurze zeitliche Abfolge (auf derselben Fragebogenseite)*

Hierfür wurden vier Fragen erstellt, welche unabhängig von Reifungseffekten oder Erinnerungsverlusten Konsistenz messen sollen. Zwei zusammengehörige Fragen bilden jeweils ein Paar und wurden gemeinsam auf einer Fragebogenseite dargestellt. Die erste Frage wurde hierbei positiv formuliert, während die zweite Frage zum gleichen Sachverhalt gegenläufig formuliert wurde. Die Beantwortung der beiden Fragen erfordert jeweils eine numerische Angabe, welche sich für spätere Analysen zusammenrechnen lassen. Die daraus resultierenden Abweichungen stellen die Konsistenzmessung dar. Die in der Befragung genutzten Fragen lauten:

Konsistenz (kurze zeitliche Abfolge)	
<i>kurze zeitliche Abfolge</i>	Wie viele Stunden schlafen Sie durchschnittlich?
	Wie viele Stunden sind Sie durchschnittlich wach?
	Wie viel Prozent Ihres aktuellen Studiums haben Sie bisher abgeschlossen?
	Wie viel Prozent Ihres aktuellen Studiums fehlt Ihnen noch, damit es abgeschlossen ist?

2) *Lange zeitliche Abfolge (auf verschiedenen Fragebogenseiten)*


Die lange Abfolge wurde gewählt, da geprüft werden soll, ob ein gebildetes Urteil wirklich stabil ist. Hierfür werden Bewertungen von Politikern auf einer Sympathieskala genutzt. Zur Messung der Konsistenz werden hierfür zu Beginn und am Ende der Befragung dieselben Fragen gestellt, es liegt folglich eine Messwiederholung vor. Um die Fragewiederholungen zu plausibilisieren, wurde kurz vor Ende der Befragung ein Speicherfehler vorgetäuscht. Hierfür erschien folgende Nachricht:

Datenbanküberprüfung!

Unipark durchsucht gerade die Speicherdatenbank und prüft ob alle Daten korrekt gesichert wurden. Dieser Vorgang dauert einige Sekunden.

Es kann passieren, dass aufgrund der Konnektivität einige Dateneingaben nicht korrekt erfasst wurden. Die Software wird dann ggf. die nicht korrekt gespeicherten Daten erneut abfragen.

Bitte haben Sie etwas Geduld!

 **Comment**

Diese wurde nach 5 Sekunden automatisch um zwei weitere Zeilen ergänzt:

Datenbanküberprüfung!


Unipark durchsucht gerade die Speicherdatenbank und prüft ob alle Daten korrekt gesichert wurden. Dieser Vorgang dauert einige Sekunden.

Es kann passieren, dass aufgrund der Konnektivität einige Dateneingaben nicht korrekt erfasst wurden. Die Software wird dann ggf. die nicht korrekt gespeicherten Daten erneut abfragen.

Bitte haben Sie etwas Geduld!

Es wurde(n) insgesamt 2 Seite(n) gefunden, die nicht korrekt gespeichert wurde(n).

Mit dem Klick auf "Weiter" wird die Information neu abgerufen! Bitte füllen Sie die fehlenden Informationen erneut aus!

 **Comment**

Die Items zur Messung der Konsistenz lauten nun wie folgt:

Konsistenz (lange zeitliche Abfolge)
Wie unsympathisch bzw. sympathisch stufen Sie die folgenden Politikerinnen und Politiker der Bundesregierung ein?
Angela Merkel Sigmar Gabriel Frank-Walter Steinmeier Andrea Nahles Ursula von der Leyen Wolfgang Schäuble Alexander Dobrindt Heiko Maas Raphael Zastel Peter Altmaier

d) Anzahl der Worte bei offenen Fragen

Es wird angenommen, dass der kognitive und zeitliche Aufwand zur Bearbeitung einer offenen Frage höher ist als bei einer geschlossenen Frage und die Anzahl der geschriebenen Worte somit ein Indikator für die Antwortqualität darstellt. Aus dieser Vermutung wird üblicherweise der Schluss gezogen, dass eine befragte Person, welche sich einem erhöhten Aufwand aussetzt und viele Worte zur Beantwortung nutzt, auch im restlichen Fragebogen ein hohes Maß an Antwortqualität aufweist. Für die Messung bedeutet dies: ein „Mehr“ an Wort ist ein „Mehr“ an Antwortqualität.⁹⁷ Die Messung ist jedoch nicht unproblematisch, da zum einen nicht kontrolliert wird, ob der Aufwand einer offenen Frage tatsächlich bei den Befragten als „höher“ wahrgenommen wird und zum anderen nicht Alternativerklärungen für eine erhöhte Wortnutzung denkbar sind.⁹⁸

Zur Zählung der Worte, später auch der Zeichen, werden insgesamt zwei Fragen verwendet:

- 1. Bitte überlegen Sie sich mindestens zwei Möglichkeiten, die Sie sehen um die Antwortqualität von Befragten bei Umfragen zu verbessern.*
- 2. Es ist nun auch sehr wichtig zu erfahren, warum Sie ursprünglich an dieser Umfrage teilgenommen haben. Bitte erläutern Sie daher die Gründe für Ihre Teilnahme.*

⁹⁷ Diese folgenden offenen Fragen werden nicht als Indikator für die intrinsische Motivation herangezogen, da sie keinen optionalen Charakter aufweisen, sondern inhaltlich mit den Vorfragen verknüpft sind. Sie erscheinen damit als ein wichtiger, integrierter Bestandteil der gesamten Befragung und weisen damit eine Abhängigkeitsstruktur zur intrinsischen Motivation auf.

⁹⁸ Beispielsweise kann eine geringe Schreib- und Sprachkompetenz einen Einfluss auf eine geringe Anzahl von geschriebenen Worten haben und damit nicht zwangsweise mit der Antwortqualität zusammenhängen. Bei Blick auf den kognitiven Antwortprozess (vgl. Tourangeau & Rasinski (1988)) kann es dennoch plausibel erscheinen, dass aufgrund einer geringen Schreib- und Sprachkompetenz das Comprehension beeinflusst wird und dadurch eine Verschlechterung der Antwortqualität gefördert wird.

6.2.2 Indikatoren für ein (situational) wahrheitsgemäßes Bearbeiten des Fragebogens

a) Validierungen

Zur Prüfung der Verlässlichkeit von Antworten wurden sog. Validierungsstudien durchgeführt, welche die Antworten zu Faktfragen an bestehenden „objektiven“ Wissensbeständen überprüfen (vgl. Skarbek-Kozietulska et al. (2012), Häder & Kühne (2010)). Diese Informationen beruhen auf Erinnerungen und es wird angenommen: je präsenter die Information, desto eher ist mit einer korrekten Erinnerungsleistung zu rechnen. Die Validierungsstudien nutzen Erinnerungsfragen mit unterschiedlicher Reichweite. Eine kurze Reichweite kann durch sogenannte Pseudo-Opinions geprüft werden. Hierfür werden beispielhaft Einstellungsabfragen gegenüber fiktiven Gesetzesvorlagen oder fiktiven Politikern in den Fragebogen eingebaut. Häder & Kühne (2010) weisen darauf hin, dass eine Bewertungsangabe zu einem fiktiven Sachverhalt ein guter Indikator für die Antwortqualität darstellt, denn „eine positive Antwort auf eine solche Frage deutet auf eine geringe Antwortqualität hin“ (Häder & Kühne (2010), S. 113).⁹⁹

Im Rahmen der Prüfung der Wirkung von Incentives auf die Qualität von Umfragen wurden bisher nur in sehr geringem Umfang Validierungsverfahren eingesetzt (vgl. Medway & Tourangeau (2015); Medway (2012); Tourangeau et al. (2010); Krenzke et al. (2005); McDaniel & Rao (1980); Godwin (1979)).¹⁰⁰

⁹⁹ Bei Pseudo-Opinions muss darauf geachtet werden, dass keine Ähnlichkeiten zu existierenden Antwortmöglichkeiten bestehen, da sonst von den Befragten ein Fehlschluss vorgenommen werden kann.

¹⁰⁰ Zusammengefasst kann über die Ergebnisse der Studien berichtet werden: die jüngeren Studien finden keinen Zusammenhang zwischen einer genaueren Berichterstattung von Fakten in Abhängigkeit eines Incentives. Jedoch ist entweder die Messung sehr ungenau und verfälscht oder die notwendige Erinnerungsleistung ist wirklich nur sehr schwer zu erbringen. Darüber hinaus kann bei einer Gruppierung zu Pre-Paid- und Post-Paid-Studien auch ersehen werden, dass widersprüchliche Ergebnisse vorliegen und unklar bleibt woran dies liegt. Weitergehend wurde oft nur eine Incentivestufe (5\$) gewählt, so dass Schwellenwerteffekte nicht kontrolliert werden können.

Innerhalb dieser Befragung ist in jedem der drei Abschnitte des Fragebogens mindestens ein Sachverhalt zu bewerten, welcher – auch bei Nutzung des Internets zu Recherchezwecken – inhaltlich nicht sinnvoll bewertet werden kann. Es soll folglich das Antwortverhalten mit Bezug auf Pseudo-Opinions gemessen werden. Die Fragen lauten:

- 1) *Wie unsympathisch bzw. sympathisch stufen Sie die folgenden Politikerinnen und Politiker der Bundesregierung ein? [Raphael Zastel]¹⁰¹*
- 2) *Der Konsum der Droge "LA-42" führt zu starken gesundheitlichen Problemen.*
- 3) *Der Konsum der Droge "LA-42" führt zur sozialen Isolation.*
- 4) *Haben Sie schon etwas von der Droge „LA-42“ gehört?*
- 5) Die Bewertung von nicht existenten Deckblättern:

Es gibt die Überlegung die Umfrage auch als Papierbogen in anderen Universitäten zu verteilen. Dafür wurden zwei Deckblätter entwickelt. Bitte sagen Sie uns welches Deckblatt Ihnen mehr zusagt:



Das linke Deckblatt Das rechte Deckblatt

Weiter



¹⁰¹ Die Frage wird zu Beginn und am Ende der Befragung gestellt. Der Name des fiktiven Politikers Raphael Zastel wurde so erdacht, dass die Verwechslungsgefahr mit einem existierenden Bundespolitiker oder einer anderen Person des öffentlichen Lebens sehr gering ist.

b) Soziale Erwünschtheit

Im Rahmen der Forschung zur sozialen Erwünschtheit kann auf eine Vielfalt an Definitionen und Konzeptualisierungen geblickt werden (vgl. Hartmann (1991), Lischewski (2015)). Hartmann fasst die zentralen Elemente der sozialen Erwünschtheit wie folgt zusammen:

„Wenn man soziale Erwünschtheit primär als Validitätsproblem versteht, ist der zentrale Begriff der des *SD Bias*. Der gemeinsame Nenner verschiedener Konzeptionen läßt sich am besten in folgender Weise skizzieren. Es geht um einen *nichtzufälligen Fehler* (Bias) bzw. eine *Tendenz zur Abgabe systematisch fehlerhafter Antworten* (Response Set) bei der Erhebung von Daten durch Befragung. Der Fehler besteht darin, daß die gegenüber dem Interviewer geäußerte Meinung von der "wahren" Meinung abweicht oder daß der Befragte über ein Verhalten falsch berichtet. Wir haben es mithin mit einer Verletzung der Befragtenrolle zu tun. Weiterhin geht es nicht um beliebige Diskrepanzen, sondern nur um solche, bei denen die *diskrepante Selbstdarstellung, gemessen an den tatsächlichen Verhältnissen, zu positiv ist*“ (Hartmann (1991), S. 235).

Die positive Selbstdarstellung, bzw. Selbstwahrnehmung soll in dieser Studie ebenfalls berücksichtigt werden. Zur Prüfung von sozialer Erwünschtheit kommt der *Balanced Inventory of Social Desirability Responses* (BIDR) zum Einsatz, welcher die Dimensionen der Fremdtäuschung (Impression Management; kurz: IM) und Selbsttäuschung (Self-Deception Enhancement; kurz: SDE) umfasst:¹⁰²

¹⁰² Die BIDR-Skala wurde gewählt, da diese trotz der Messprobleme noch als vergleichsweise reliabel angesehen werden kann (vgl. Lischewski (2015)).

Kurzsкала des BIDR	
<i>SDE</i>	Mein erster Eindruck von Menschen stellt sich gewöhnlich als richtig heraus Ich bin mir oft unsicher in meinem Urteil Ich weiß immer genau, wieso ich etwas mag
<i>IM</i>	Ich habe schon mal zuviel Wechselgeld zurückbekommen und nichts gesagt Ich bin immer ehrlich zu anderen Ich habe gelegentlich mal jemanden ausgenutzt

Als primär relevante Dimension für ein wahrheitsgemäßes Beantworten des Fragebogens erscheint die Fremdtäuschung, da diese bewusst gesteuert und folglich als Messung für ein sozial erwünschtes Beantworten interpretiert werden kann.¹⁰³ Je höher dieser Wert, desto eher ist von einer geringen Antwortqualität auszugehen.

6.2.3 Indikatoren für ein vollständiges Bearbeiten eines Fragebogens

a) Item-Nonresponse

In der Definition von Antwortqualität wird der Begriff der „Vollständigkeit“ genutzt. Wie in Kapitel 2.3.2 ausgeführt ist es jedoch schwer die Gründe für eine Nichtbeantwortung einer Fragebogenfrage herauszufinden und damit zu unterscheiden, ob Meinungslosigkeit, ein kognitives Unvermögen, eine unvollständige Antwortskala oder eine Aufwandsminimierungsstrategie ursächlich ist. Aus diesem Grund wurde ein neues Vorgehen zur Operationalisierung gewählt: Die Befragten wurden zu Beginn der Befragung darüber aufgeklärt, dass bei einer Nichtbeantwortbarkeit einer Frage ein Kommentar hinterlassen werden soll. Dementsprechend gab es im Fragebogen, bis auf eine Ausnahme¹⁰⁴, auch keine „weiß nicht“ und „keine Angabe“

¹⁰³ Es muss noch darauf hingewiesen werden, dass die Fremdtäuschung und die Selbsttäuschung von Paulhus (1984) als Trait definiert werden. Daher dürften keine Abhängigkeitsstrukturen im Rahmen von Incentives vorliegen. Es wird jedoch ein Zusammenhang zwischen der Antwortqualität und der Fremddarstellungstendenz (Impression Management) unterstellt.

¹⁰⁴ Es handelt sich hierbei um die Frage nach dem Geschlecht. Die Begründung liegt in der Forschungsethik, da damit der Empfehlung des deutschen Ethikrates gefolgt wird den Befragten direkt die Möglichkeit zu „anderes“ oder „keine Angabe“ ohne Diskriminierung gegeben wird (vgl. Deutscher Ethikrat (2012)). Da die Frage nach dem Geschlecht erst am Ende des zweiten Fragebogenblocks kommt wurde angenommen, dass aufgrund des

Kategorien.¹⁰⁵ Hiermit ergibt sich die Möglichkeit zur Unterscheidung der Ursachen für Item-

Nonresponse:

Meinungslosigkeit	Hierfür kann eine kurze Notiz hinterlassen werden
kognitives Unvermögen	Hierfür kann die unverständliche oder mehrdimensionale Komponente der Frage in dem Kommentarfeld eingegeben werden
fehlende Antwortkategorie	Hierfür kann die nicht vorhandene, aber gewünschte Antwortkategorie in dem Kommentarfeld eingegeben werden.
Aufwandsminimierung	Hierfür wird die Frage übersprungen, ohne einen Kommentar zu hinterlassen. ¹⁰⁶

Die relevante Komponente zur Messung der Antwortqualität ist die Aufwandsminimierung und diese kann wie folgt berechnet werden:

$$\begin{aligned} & \text{Anzahl der Items die übersprungen wurden} \\ & - \underline{\text{Anzahl der Items die übersprungen, aber durch Kommentare begründet wurden}} \\ & = \text{Anzahl der Items die aus Gründen des Aufwands, ohne Kommentar übersprungen wurden} \end{aligned}$$

b) Filterfragen

Im Rahmen der Vollständigkeit soll auch das Verhalten der Befragten bei Filterfragen untersucht werden, da einige Autoren ein Ausnutzen von Filtern zur Aufwandsminimierung unterstellen (vgl. Faulbaum et al. (2009), Kreuter et al. (2011)). In der Studie von Kreuter et al. (2011) wurde untersucht, welches Filterformat am ehesten bei den Befragten ein Überspringen der Filter provoziert: Grouped Version oder Interleafed Version.

bis dahin geübten Antwortverhaltens die einmalige Sicht der „keine Angabe“ Kategorie die Nutzung der Kommentarfunktion nicht vergessen lässt oder verdrängt.

¹⁰⁵ Ermöglicht wurde dieses Vorgehen durch die Anwendung der Pretest-Funktion von Unipark. Diese Kommentare werden seitenweise gespeichert, so dass diese später den jeweiligen Items noch zugeordnet werden müssen.

¹⁰⁶ Hierbei muss natürlich beachtet werden, dass eine der anderen drei Kategorien ursächlich für den Item-Nonresponse sein können, allerdings ist die Verweigerung des Aufwands zur Kommentarerstellung ein plausibles Indiz für eine Aufwandsminimierung.

Abb. 19: Darstellung verschiedener Filterformate

Grouped version	Interleafed version
<i>In the past 3 months, have you purchased a coat?</i>	<i>In the past 3 months, have you purchased a coat?</i>
<i>In the past 3 months, have you purchased a shirt?</i>	Please briefly describe your most recent coat purchased.
<i>In the past 3 months, have you purchased pants?</i>	For whom was it purchased?
<i>In the past 3 months, have you purchased a suit?</i>	In what month did you purchase it?
<i>In the past 3 months, have you purchased a dress?</i>	How much did it cost?
 	<i>In the past 3 months have you purchased a shirt?</i>
FOR EACH YES:	Please briefly describe your most recent shirt purchased.
Please briefly describe your most recent [item] purchased.	For whom was it purchased?
For whom was it purchased?	In what month did you purchase it?
In what month did you purchase it?	How much did it cost?
How much did it cost?	AND SO ON FOR Pants, Suits and Dresses

Quelle: Kreuter et al. (2011), S. 90.

Es konnte festgestellt werden, dass bei der Interleafed Version die Befragten signifikant häufiger ein Verneinungsverhalten aufwiesen und damit öfter Filterfragen übersprangen. Dies wird darauf zurückgeführt, dass mittels eines solchen Vorgehens der Aufwand für die Beantwortung der Filterfragen abschätzbar wird und die Befragten als Vermeidungsstrategie zu Verneinungen neigen. In dieser Studie soll ebenfalls die Filterführung gemäß der Interleafed Version genutzt werden, da den Befragten bewusst eine Verneinungsstrategie ermöglicht werden soll und damit eine Verschlechterung der Antwortqualität gemessen werden kann.¹⁰⁷

Es ist hierbei natürlich wichtig zu unterscheiden, ob eine Filterfrage aus Gründen der Unkenntnis einer Droge oder aus Aufwandsminimierung übersprungen wird. Um diese Unterscheidung

¹⁰⁷ In der Studie von Medway (2012) wird dieses Verfahren bereits mit Blick auf die Wirkung von Incentives angewendet. Als Ergebnis konnte kein signifikanter Unterschied im Verneinen von Filterfragen zwischen der Experimentalgruppe (\$5) und der Kontrollgruppe festgestellt werden. Medway kritisiert allerdings selbst, dass die drei Folgefragen (pro Filter) vielleicht nicht aufwändig genug gewesen seien und das die Filterfragen am Anfang des Fragebogens standen (fünf Fragen waren davor erst gestellt worden) und die Befragten damit noch motiviert sind und folglich keine Ausweichstrategien nutzen.

ermöglichen zu können, wurden in der Befragung zuerst die Bewertungen der Drogen und dann folgend die Kenntnis abgefragt. Somit können über das Antwortverhalten und hinterlassene Kommentare Inkonsistenzen aufgedeckt werden. Es muss jedoch angemerkt werden, dass die Befragten erahnen können, dass die (nicht-existente) Droge LA-42 ebenfalls als Filterfrage abgefragt wird. Dies könnte dazu führen, dass die Befragten die vorherigen Filterfragen nicht überspringen, da sie mindestens einen aufwandsmindernden Übersprung antizipieren können. Die Filterfragen mit den jeweiligen Filterwegen können der folgenden Darstellung entnommen werden. Unter jedem Hauptfilter liegt eine Fragebogenseite, welche eine offene und zwei geschlossene Fragen umfasst. Die Filterführung ist folgend dargestellt:



6.2.4 Indikatoren für ein anweisungsfolgendes Bearbeiten des Fragebogens

Den Befragten werden im Rahmen der Befragung mehrere Anweisungen zur Bearbeitung des Fragebogens gegeben. Diese Anweisungen bergen teilweise einen hohen Aufwand in sich, so dass ein Nichtbefolgen für die Befragten eine Zeit- und Aufwandsersparnis bedeuten kann. Das Einhalten der Anweisungen kann unter Verwendung von Para-Daten¹⁰⁸ überprüft und ausgewertet werden.

- a) *Die Nutzung des Windows-Taschenrechners zur Berechnung eines Glückszahlkoeffizienten.*

Es wird oft vermutet, dass bestimmte Persönlichkeitstypen zu bestimmten Zahlen tendieren (Lieblingszahlen). Führen Sie daher die kurze Rechenoperation aus, um Ihren Glückszahlkoeffizienten zu berechnen. Bitte nutzen Sie zur Berechnung den Windows Taschenrechner (ist bereits im Hintergrund geöffnet) und geben Sie das genaue Ergebnis mit allen Nachkommastellen an.

Denken Sie an eine beliebige Zahl von 1 bis 9.
Teilen Sie die von Ihnen ausgewählte Zahl durch 12,5.
Multiplizieren Sie das Ergebnis mit 35.
Addieren Sie auf das neue Ergebnis 2.

Wie lautet das finale Ergebnis (mit allen Nachkommastellen)?

Weiter



- b) *Die Nutzung von Google Maps® zur Bestimmung der Entfernung des Wohnortes zum Erhebungslabor.*

Bitte geben Sie die genaue Entfernung (in KM) von Ihrer Unterkunft bis zum OEC (Platz der Göttinger Sieben 3) an. Mit Unterkunft ist der Ort gemeint, an welchem Sie seit Beginn des Jahres am häufigsten übernachtet haben.

Öffnen Sie hierfür bitte einen neuen TAB in diesem Browser und starten Sie den Routenplaner von Google Maps (maps.google.de). Sie werden verschiedene Routenmöglichkeiten angeboten bekommen, zur Beantwortung dieser Frage wählen Sie bitte die kürzeste Strecke, welche für Fußgänger angegeben wird.

Weiter



¹⁰⁸ Es wird hierfür erhoben, ob die Probanden das Browser-Fenster mit dem Fragebogen verlassen und sich damit anderen Homepages oder Computerprogrammen widmen.

7. Erste Analysen zur Bewertung der Messinstrumente der erklärenden Variablen

In diesem Kapitel werden erste Analysen auf Basis der Operationalisierungen der intrinsischen Motivation (Kapitel 7.1), der Reziprozitätshypothese (Kapitel 7.2) und der extrinsische Motivation (Kapitel 7.3) durchgeführt.¹⁰⁹ Es wird in diesem Kontext eine Auswahl an geeigneten Indikatoren für die später folgenden multivariaten Analysen vorgenommen. Darüber hinaus werden auch erste Analysen der Zusammenhangsstrukturen zwischen den Indikatoren und den drei Versuchsgruppen vorgenommen.¹¹⁰

7.1 Indikatoren der intrinsischen Motivation

7.1.1 Der Zusammenhang zwischen den Messungen für intrinsische Motivation

Es wird zuerst untersucht, inwiefern die vier Operationalisierungen der intrinsischen Motivation (Selbsteinstufungen auf Basis der SIMS, Bearbeitungszeiten, Wiederbefragungsbereitschaft und Anzahl an Worten in offener Frage) zusammenhängen und damit als äquivalent angesehen werden können. Hierfür wird eine explorative Faktorenanalyse durchgeführt:

¹⁰⁹ Der Mittelwertindex ist so kodiert, dass hohe Werte für eine hohe intrinsische Motivation stehen.

¹¹⁰ Die maximal zulässige Irrtumswahrscheinlichkeit wird für diese Studie aufgrund der geringen Fallzahl auf 10% festgelegt. Unabhängig davon werden auch die Vorzeichen und die damit verbundenen Effektgrößen vorsichtig interpretiert.

Tab. 7: Faktorladungen der explorativen Faktorenanalyse für die ausgewählten Indikatoren der intrinsischen Motivation

	Mustermatrix		
	Faktor I	Faktor II	Faktor III
Bearbeitungsdauer - Block I	0.913	-0.006	-0.130
Bearbeitungsdauer - Block II	0.934	0.003	-0.072
Bearbeitungsdauer - Block III	0.700	-0.016	0.286
Weil ich denke, dass die Umfrage interessant ist.	0.117	0.639	-0.013
Weil ich es für angenehm halte.	-0.018	0.804	-0.047
Weil es mir Spaß macht.	-0.102	0.813	-0.048
Weil ich mich gut fühle, sobald ich daran teilnehme.	-0.024	0.596	0.095
Wiederbefragungsbereitschaft - final	0.049	0.103	0.569
Häufigkeit von Worten in offener Frage	-0.068	-0.113	0.844

Quelle: eigene Daten. Die Fallzahl beträgt n = 176. Die Faktoren basieren auf einem Eigenwertkriterium von >= 1. Es wurde die Oblimin-Rotation genutzt, da die einzelnen Faktoren untereinander korrelieren dürfen.

Die Berechnungen der explorativen Faktorenanalyse ergeben drei Faktoren. Der erste Faktor umfasst die Items der Bearbeitungszeit in den drei Abschnitten des Fragebogens, der zweite Faktor die Messungen zur Selbsteinstufung auf Basis der Situational Intrinsic Motivation Scale und der dritte Faktor speist sich aus der Wiederbefragungsbereitschaft und der Worthäufigkeit.¹¹¹ Die vorliegende Faktorenlösung ist überraschend, da die Operationalisierungen der intrinsischen Motivation nicht einen gemeinsamen Faktor bilden und die einzelnen extrahierten Faktoren nur sehr schwach untereinander korrelieren.

¹¹¹ Der dritte Block endet, bevor die offene Frage gestellt wird. Damit wird die Beziehung zwischen der Bearbeitungszeit und der offenen Frage nicht konfundiert, da das Schreiben vieler Worte mit einem erhöhten zeitlichen Aufwand einhergeht.

Tab. 8: Die Korrelationen zwischen den drei extrahierten Faktoren

Faktor I	1		
Faktor II	0.039	1	
Faktor III	0.149	0.066	1
n = 180	Faktor I	Faktor II	Faktor III

Quelle: eigene Daten.

Dies weist darauf hin, dass die Operationalisierungen entweder unterschiedliche Dimensionen der intrinsischen Motivation widerspiegeln, oder zu invaliden Messungen führten.

Der dritte Faktor umfasst zwei Operationalisierungen der intrinsischen Motivation und könnte damit am ehesten der intrinsischen Motivation zu entsprechen. Unter Berücksichtigung der sehr geringen internen Konsistenz (Cronbachs Alpha = 0.015) löst sich dieser Zusammenhang jedoch auf und deutet damit auf einen sehr instabilen Faktor hin. Dieser Befund wird auch durch den Vergleich der empirischen Korrelationen zwischen der Wiederbefragungsbereitschaft und der Anzahl der Worte ($r = 0.088$) mit den modellimplizierten Korrelationen der Faktorenanalyse gestützt ($r = 0.466$) gestützt, da diese von letzteren deutlich überschätzt werden. Zusammengefasst bedeutet dies, dass die Messungen nicht äquivalent genutzt werden können und daher für eine Nutzung in den multivariaten Analysen gegeneinander abgewogen werden müssen. Hierfür werden zuerst die einzelnen Messungen genauer betrachtet und anschließend eine Auswahl vorgenommen.

7.1.2 Die Selbsteinstufung zur intrinsischen Motivation auf Basis der SIMS

Die Messung der intrinsischen Motivation auf Grundlage der Situational Motivation Scale basiert auf vier items. Es werden nun zuerst die Korrelationen zwischen den einzelnen Items der Selbsteinstufung berechnet, da die interne Konsistenz der Messungen geprüft werden soll.

Tab. 9: Korrelationen zwischen den SIMS-Items der intrinsischen Motivation

1. Weil ich denke, dass die Umfrage interessant ist.	1			
5. Weil ich es für angenehm halte.	0.347 (p = 0.000) n = 175	1		
9. Weil es mir Spaß macht.	0.386 (p = 0.000) n = 175	0.524 (p = 0.000) n = 174	1	
13. Weil ich mich gut fühle, sobald ich daran teilnehme.	0.158 (p = 0.037) n = 175	0.354 (p = 0.000) n = 174	0.342 (p = 0.000) n = 174	1
	1. Weil ich denke, dass die Umfrage interessant ist.	5. Weil ich es für angenehm halte.	9. Weil es mir Spaß macht.	13. Weil ich mich gut fühle, sobald ich daran teilnehme.

Quelle: eigene Daten.

Die Beziehungen zwischen den Items sind alle signifikant, wobei die Zusammenhangsstärken variieren. Eine recht geringe Korrelation mit $r = 0.158$ kann zwischen dem ersten und dem letzten Item festgestellt werden, wobei eine sehr starke Korrelation von 0.524 zwischen dem zweiten und dritten Item vorliegt. Die eher geringeren Korrelationen können jedoch in den Strukturgleichungsmodellen für die Hypothesenprüfungen zu Problemen führen, da sie dann keinen gemeinsamen Faktor bilden. Daher wurde ergänzend eine Reliabilitätsanalyse für die vier Items durchgeführt. Die Berechnungen ergaben, dass ein Auslassen des ersten und letzten Items zu einer tendenziellen Verbesserung der Reliabilität führt (Cronbachs Alpha steigt von 0.666 auf 0.682; die Trennschärfe der beiden Items beträgt < 0.4).

Aufgrund dieser Befunde wurde ein Mittelwertindex berechnet, welcher nur die beiden Items berücksichtigt. Die Mittelwerte für die drei Versuchsgruppen sind in der folgenden Tabelle zusammengefasst:

Tab. 10: Mittelwerte für die stark zusammenhängenden Items der intrinsischen Motivation

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	4.15	58	1.47
Incentive 5€	4.32	58	1.25
Incentive 20€	4.36	60	1.21
Gesamtsumme	4.28	176	1.31

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.431$; $p = 0.651$.

Es ist für die Versuchsgruppen mit einem Incentive ein leichter Anstieg der Mittelwerte festzustellen, jedoch sind die Unterschiede nicht signifikant ($p > 0.100$). Dieser Befund entspricht damit den Vorhersagen der Cognitive Evaluation Theory. Bei Betrachtung der Mittelwerte kann darüber hinaus festgestellt werden, dass die intrinsische Motivation mit einem durchschnittlichen Wert von 4.28 nur etwas über dem Mittelwert der Skala liegt. Dies ist insofern überraschend, da den Befragten eine hohe intrinsische Motivation unterstellt wird. Diese Annahme ergibt sich daraus, dass sich die Probanden zum einen bewusst für eine zeitaufwendige Studie entscheiden und zum anderen für eine Teilnahme im Erhebungslabor erscheinen müssen. Wie sind diese niedrigen Mittelwerte nun zu interpretieren? Zum Vergleich werden die Ergebnisse der Studie von Vogt (2004) herangezogen, da diese zum einen auf derselben Übersetzung der SIMS erfolgten und zum anderen auch auf einer studentischen Stichproben basieren. Die berichteten Mittelwerte zur intrinsischen Motivation liegen dort durchschnittlich bei 3.6 und damit unter den festgestellten Mittelwerten dieser Studie.¹¹² Dies kann als Anzeichen dafür gesehen werden, dass die Befragten dieser Studie tatsächlich eine erhöhte intrinsische

¹¹² Vogt (2004) untersucht in ihren Studien die Wirkung von Belohnungen (performance contingent) auf das social loafing in Kooperationssituationen. Die Versuchsgruppe mit Incentive weist in der ersten Studie einen durchschnittlichen Mittelwert von ca. 3.3 auf, während die Versuchsgruppe ohne Incentive einen Mittelwert von ca. 3.8. In der zweiten Studie wird die Belohnung erst nach einer Übungsphase ausgegeben, der Mittelwert der intrinsischen Motivation liegt dann bei ca. 3.7. Es muss beachtet werden, dass Vogt (2004) alle vier Items zur Berechnung des Mittelwertindex herangezogen hat. Werden die in dieser Studie vorher ausgeschlossenen Items der SIMS im Mittelwertindex berücksichtigt, so liegen die Mittelwerte durchschnittlich bei 4.6.

Motivation aufweisen. Die insgesamt relativ niedrigen Mittelwerte der Messungen der SIMS legen damit zwei Erklärungen nahe: a) Die Teilnahme an Studie ist prinzipiell mit keiner sehr hohen intrinsischen Motivation verknüpft, oder b) die zugrunde liegenden Messungen sind tendenziell ungenau.¹¹³

7.1.3 Die Bearbeitungszeit als Indikator der intrinsischen Motivation

Die Bearbeitungszeiten wurden während der Befragung in Millisekunden gemessen, wobei die Probanden – um Verzerrungen zu vermeiden – erst am Ende der Befragung über die erfolgte Messung aufgeklärt wurden. Die einzelnen Bearbeitungszeiten für den ersten, zweiten und dritten Abschnitt des Fragebogens wurden in einem Summenindex zusammengefasst¹¹⁴, wobei zur einfacheren Interpretation der Ergebnisse die Einheit in Minuten umgerechnet wurde.¹¹⁵ Werden die Mittelwerte, unter Berücksichtigung der Versuchsgruppen berechnet, so ergeben sich folgende Werte:

Tab. 11: Mittelwerte für die Bearbeitungszeit für den gesamten Fragebogen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	44.02	60	12.22
Incentive 5€	45.92	60	15.54
Incentive 20€	47.57	59	13.64
Gesamtsumme	45.83	179	13.87

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.976$; $p = 0.379$.

¹¹³ Bei Betrachtung der zugrunde liegenden Items („Weil ich es für angenehm halte.“ und „Weil es mir Spaß macht“) erscheint die Frageformulierung für die eher geringen Mittelwerte verantwortlich.

¹¹⁴ Die Zeitmessungen für Fragen, welche nicht alle Befragten erhalten haben sind in dem Index nicht berücksichtigt.

¹¹⁵ Es ist zu beachten, dass die Nachkommastellen nicht dem Sekundenprinzip (60 Sekunden = 1 Minute) folgen, sondern als Anteile der Basis 1 auftreten. Ein Wert von 0.5 entspricht damit 30 Sekunden.

Es ist festzustellen, dass die Bearbeitungszeit bei Vergabe eines Incentives ansteigt. Die Mittelwertdifferenzen sind nicht signifikant, so dass auch bei diesem Indikator von einer statistischen Unabhängigkeit des Incentives gesprochen werden kann. Medway & Tourangeau (2015) vermuten, dass sich die Bearbeitungszeit im Verlauf des Fragebogens ändern kann, d.h. sich die Probanden z.B. zu Beginn der Befragung mehr Zeit lassen als am Ende der Befragung (sog. Speeding) und deshalb die Nutzung der Messung für die gesamte Bearbeitungszeit nicht angemessen ist. Diese Vermutung soll auch in dieser Studie geprüft werden. Da der Fragebogen für die Befragten wahrnehmbar in drei Abschnitte untergliedert wurde, werden für diese Bereiche nun die durchschnittlichen Bearbeitungszeiten, aufgegliedert nach den drei Versuchsgruppen, analysiert:

Tab. 12: Mittelwerte für die Bearbeitungsdauer, aufgegliedert nach den drei Fragebogenblöcken

Block I	<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
	Kein Incentive	17.25	60	5.51
	Incentive 5€	17.79	60	7.87
	Incentive 20€	18.39	60	6.81
	Gesamtsumme	17.81	180	6.78
Block II	<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
	Kein Incentive	14.26	60	4.07
	Incentive 5€	15.10	60	5.65
	Incentive 20€	15.57	60	4.17
	Gesamtsumme	14.97	180	4.69
Block III	<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
	Kein Incentive	7.72	60	2.16
	Incentive 5€	8.20	60	2.66
	Incentive 20€	8.40	59	2.92
	Gesamtsumme	8.11	179	2.60

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab für Block I: $F = 0.425$; $p = 0.659$ / Block II: $F = 1.205$; $p = 0.302$ / Block III: $F = 1.077$; $p = 0.349$.

In dieser Darstellung ist bei Vergabe eines Incentives in allen drei Blöcken ein Anstieg der Bearbeitungsdauer und damit der intrinsischen Motivation erkennbar. Dies widerspricht auf dem ersten Blick den theoretischen Vorüberlegungen, da die intrinsische Motivation bei Incentivierung stabil bleiben sollte. Unter Beachtung der statistischen Signifikanz kann allerdings kein bedeutsamer Mittelwertunterschied festgestellt werden. Damit kann nur deskriptiv von einem leichten positiven Effekt eines Incentives auf die Bearbeitungszeit gesprochen werden.

7.1.4 Die Wiederbefragungsbereitschaft als Indikator der intrinsischen Motivation

Die Frage zur Bereitschaft für eine Wiederbefragung haben 17 Befragte verneint und weisen damit eine geringere intrinsische Motivation auf. Unter Kontrolle der Versuchsgruppen lässt sich dabei eine leicht erhöhte Wiederbefragungsbereitschaft für die Probanden mit einem Incentive von 20 Euro erkennen:

Tab. 13: Häufigkeitstabelle für die Wiederbefragungsbereitschaft.

		Kein Incentive	Incentive: 5€	Incentive: 20€	Gesamt
Wiederbefragungsbereitschaft	Ablehnung	7 (11.70%)	7 (11.70%)	3 (5.10%)	17 (9.50%)
	Zustimmung	53 (88.30%)	53 (88.30%)	56 (94.90%)	162 (90.50%)
	Gesamt	60 (100%)	60 (100%)	59 (100%)	179 (100%)

Quelle: eigene Daten.

Es sind keine statistisch signifikanten Effekte feststellbar, wobei dies jedoch der geringen Fallzahl geschuldet sein kann.

Da die eMail-Adressen für die Wiederbefragung erst einige Fragebogenseiten später erhoben wurden, hatten die Befragten die Möglichkeit ihre ursprüngliche Zustimmung für die Wiederbefragung zurückzuziehen und damit die Teilnahme an einer Folgebefragung nachträglich zu verweigern. Aufgrund dieser Möglichkeit verweigern zusätzlich 23 Befragte die Wiederbefragung.

gung und somit haben insgesamt 40 Befragte kein Interesse an einer unentgeltlichen Folgebefragung teilzunehmen.¹¹⁶ Auch für die neue Variable wird eine Analyse für die drei Versuchsgruppen vorgenommen:

Tab. 14: Häufigkeitstabelle für die Wiederbefragungsbereitschaft nach erneuter Rückfrage.

		Kein Incentive	Incentive: 5€	Incentive: 20€	Gesamt
Wiederbefragungsbereitschaft (final)	Ablehnung	17 (28.30%)	13 (21.70%)	10 (16.90%)	40 (22.30%)
	Zustimmung	43 (71.70%)	47 (78.30%)	49 (83.10%)	139 (77.70%)
	Gesamt	60 (100%)	60 (100%)	59 (100%)	179 (100%)

Quelle: eigene Daten.

In der Tabelle ist eine deutlichere Tendenz der Zunahme der Wiederbefragungsbereitschaft bei Incentivierung zu erkennen. Bei Berechnung von Chi-Quadrat wird jedoch erneut keine Signifikanz erreicht ($\chi^2 = 2.246$; $df = 2$; $p = 0.325$), was auch hier auf eine statistische Unabhängigkeit der intrinsischen Motivation vom Incentive hinweist.

7.1.5 Die Worthäufigkeit als Indikator für die intrinsische Motivation

Wie bereits im Kapitel zur Operationalisierung dargestellt, werden auch die Worte, welche zur Beantwortung einer abschließenden offenen Zusatzfrage genutzt werden, gezählt und analysiert. Die Worthäufigkeiten nehmen bei Incentivierung tendenziell zu, wobei in der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro der Effekt geringer ausfällt als in der Versuchsgruppe mit einem Incentive in Höhe von 5 Euro:

¹¹⁶ Es ist allerdings unklar, wieso die Befragten an dieser Stelle die ursprüngliche Zusage zurücknehmen. Dies kann beispielhaft zum einen an einer gesunkenen intrinsischen Motivation oder zum anderen an einer Verweigerung zur Nennung der eMail-Adresse aus Gründen der Anonymität.

Tab. 15: Mittelwerte für die Anzahl der Worte.

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	4.05	60	8.09
Incentive 5€	5.70	60	13.90
Incentive 20€	5.27	59	9.58
Gesamtsumme	5.01	179	10.78

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.376$; $p = 0.687$.

Zur Prüfung der statistischen Signifikanz wurde eine Varianzanalyse berechnet, welche jedoch keine signifikanten Unterschiede zwischen den drei Versuchsgruppen aufzeigen konnte.

Alternativ wird nun auch die Anzahl der gebrauchten Zeichen untersucht: im Vergleich zur Versuchsgruppe ohne Incentive werden in der Gruppe der Befragten mit einem Incentive von 5 Euro werden durchschnittlich ca. 9.5 Zeichen und in der Gruppe mit einem Incentive von 20 Euro ca. 7 Zeichen mehr zur Beantwortung der Frage genutzt.

Tab. 16: Mittelwerte für die Anzahl an Zeichen.

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	26.08	60	52.98
Incentive 5€	35.65	60	86.51
Incentive 20€	33.20	59	60.83
Gesamtsumme	31.64	179	68.07

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.317$; $p = 0.729$.

Auch in diesem Fall können keine signifikanten Unterschiede aufgedeckt werden. Dies kann über die hohe Streuung begründet werden. Daher wurden U-Tests durchgeführt, wobei jedoch auch hier keine signifikanten Unterschiede festgestellt werden konnten ($p > 0.2$). Die Insignifikanz der Ergebnisse kann stützt auch in diesem Fall die theoretischen Erwartungen zur Wirkung der Incentives (gemäß der Cognitive Evaluation Theory) auf die intrinsische Motivation.

7.1.6 Die Auswahl eines angemessenen Indikators der intrinsischen Motivation

Die Auswahl einer angemessenen Operationalisierung für die intrinsische Motivation soll aus theoretischen und empirischen Überlegungen erfolgen. Hierfür werden die einzelnen Indikatoren noch einmal vorgestellt und diskutiert:

a) Die Nutzbarkeit der Subskala der SIMS als Indikator der intrinsischen Motivation

Die Messung der intrinsischen Motivation über die Situational Intrinsic Motivation Scale basiert auf Selbsteinstufungen und kann deshalb anfällig für Akquieszenz oder soziale Erwünschtheit sein. Zur Vermeidung von Akquieszenz sind in der vollständigen Situational Motivation Scale (= 16-Items) einige Aussagen ungleich gepolt, so dass sich die Befragten nicht dauerhaft an einem Pol der Antwortskala verorten können. Bei Betrachtung der Daten kann darüber hinaus festgestellt werden, dass sich die Befragten bei den gleichgepolten Items nicht nur an einem Antwortpol aufhalten, sondern ein heterogenes Ankreuzverhalten aufweisen. Dies weist darauf hin, dass keine Akquieszenz vorliegt. Weitergehend wurden die Items der SIMS direkt zu Beginn der Befragung abgefragt, da hier keine bzw. nur sehr geringe Erschöpfungserscheinungen bei den Befragten vorliegen dürften und damit kein Satisficing erwartet wird. Um soziale Erwünschtheit auszuschließen, wurden die einzelnen Items der intrinsischen Motivation auf Korrelationen mit den Dimensionen der sozialen Erwünschtheit (gemessen über den BIDR) überprüft. Hierbei konnten keine signifikanten Zusammenhänge aufgedeckt werden. Für eine Nutzung der Situational Intrinsic Motivation Scale als Indikator für die intrinsische Motivation kann argumentiert werden, dass prinzipiell vier Items zur Bestimmung des Faktors für die intrinsische Motivation in den Strukturglei-

chungsmodellen verwendet werden können¹¹⁷ und damit der Messfehler kontrollierbar wird. Die Selbsteinstufungen auf Basis der SIMS erscheinen damit insgesamt als Indikator für die intrinsische Motivation angemessen.

b) Die Nutzbarkeit der Bearbeitungszeit als Indikator der intrinsischen Motivation

Der Vorteil der Verwendung der Bearbeitungszeit als Indikator für die intrinsische Motivation liegt in der Vielzahl an vorgenommenen Messungen. Dadurch sind zum einen Analysen über verschiedene Bearbeitungsbereiche des Fragebogens möglich und zum anderen ist der Messfehler in Strukturgleichungsmodellen kontrollierbar. Es ist jedoch fragwürdig, inwiefern die Bearbeitungszeit tatsächlich die intrinsische Motivation der Befragten widerspiegelt und nicht von anderen individuellen Eigenschaften moderiert wird. Diese Bedenken lassen sich auch empirisch fundieren: Im Rahmen der Befragung wurde die selbst eingestufte Fähigkeit im Umgang mit Computern erhoben. Es wird hierbei erwartet, dass erfahrene Nutzer sicherer im Umgang mit einer Online-Umfrage sind und daher auch weniger Bearbeitungszeit benötigen. Diese Annahme wird durch eine negative Korrelation zwischen der Fähigkeit der Computernutzung und der Bearbeitungszeit gestützt ($r = -0.170$; $p = 0.023$). Da Unklarheit über die (auch theoretischen) Abhängigkeitsstrukturen besteht, wird die Bearbeitungszeit nicht als Indikator für die intrinsische Motivation herangezogen.

¹¹⁷ Der Wert von Cronbachs Alpha liegt für die vier Items bei 0.682.

c) *Die Nutzbarkeit der Wiederbefragungsbereitschaft als Indikator für intrinsische Motivation*

Die Wiederbefragungsbereitschaft der Befragten kann als Indikator für die intrinsische Motivation betrachtet werden, da eine erneute, unentgeltliche Teilnahme an einer Folgebefragung für ein starkes intrinsisches Interesse sprechen kann. Es muss jedoch beachtet werden, dass eine Zustimmung zu einer potentiellen späteren Befragung unverbindlich ist und damit keine Aussage über die tatsächliche Bereitschaft zulässt. Die Zustimmung könnte somit aufgrund eines sozial erwünschten Antwortverhaltens gegeben werden.¹¹⁸ Darüber hinaus kann die Position der Messung der Wiederbefragungsbereitschaft im Fragebogen als problematisch betrachtet werden, da diese erst kurz vor Ende der Befragung erfolgt und damit nach den Messungen der Indikatoren der Antwortqualität. Dies könnte später in den multivariaten Analysen zu Unsicherheiten in der kausalen Interpretation führen, da die Messung der unabhängigen Variablen nicht vor den abhängigen Variablen erfolgen sollte.¹¹⁹ Ebenfalls ist kritisch anzumerken, dass für die Wiederbefragungsbereitschaft auf einer einzigen Messung beruht und damit der Messfehler nicht kontrolliert werden kann. Die Messung der Wiederbefragungsbereitschaft könnte trotz alledem zur Berechnung der Zusammenhänge in einem Strukturgleichungsmodell genutzt werden, jedoch wird dann eine Abwesenheit, bzw. Unabhängigkeit von Messfehlern unterstellt.

¹¹⁸ So zeigt sich beispielhaft bei den Befragten mit einem Incentive in Höhe von 5 Euro ein fast signifikanter ($p = 0.102$) positiver Effekt bei der Wiederbefragungsbereitschaft (final) und dem Impression Management.

¹¹⁹ Eine zentrale Annahme in der Konzeption von Kausalität liegt darin, dass die Ursache vor der Wirkung auftritt. Da die intrinsische Motivation als erklärende Variable für die Antwortqualität unterstellt wird, kann aufgrund der Messreihenfolge (erst die abhängige, dann die unabhängige Variable) kein sauberer Wirkschluss mehr getroffen werden. Nur unter der Annahme, dass die intrinsische Motivation im Befragungsprozess stabil und damit konstant bleibt, könnte die Wiederbefragungsbereitschaft trotz der späten Messung genutzt werden.

Aufgrund der oben genannten Einschränkungen soll die Wiederbefragungsbereitschaft nicht als Indikator für die intrinsische Motivation herangezogen werden.

d) Die Nutzbarkeit der Anzahl an Worten als Indikator für die intrinsische Motivation

Die Anzahl der Worte, bzw. der Zeichen in der letzten offenen Frage wird nun ebenfalls als Indikator für die intrinsische Motivation diskutiert. Es muss hierbei darauf hingewiesen werden, dass Unklarheit darüber besteht, ob und inwiefern die Befragten die offene Frage als freiwillig interpretieren und sie damit dem Kriterium der intrinsischen Motivation entspricht. Darüber hinaus kann auch an dieser Stelle kritisch argumentiert werden, dass zum einen der Messfehler in den folgenden multivariaten Analysen aufgrund der einmaligen Messung nicht kontrolliert werden kann und zum anderen auch in diesem Fall die Messung erst am Ende des Fragebogens erfolgt und damit nach den Messungen für die Indikatoren der Antwortqualität. Aufgrund dieser Überlegungen wird die Anzahl der Worte, bzw. der Zeichen der letzten offenen Frage ebenfalls nicht in den multivariaten Analysen herangezogen.

Diese Überlegungen führen zu dem Schluss, dass ausschließlich die Selbsteinstufungen auf Basis der SIMS als Indikator für die intrinsische Motivation genutzt werden.

7.2 Indikatoren der Reziprozitätshypothese

Die Reziprozitätshypothese wird über zwei Komponenten geprüft. Zum einen wird die Wirkung der verinnerlichten Reziprozitätsnorm und zum anderen die Veränderung der Bewertung des Versuchsleiters betrachtet.

- a) Die verinnerlichte Reziprozitätsnorm basiert auf drei Items zur Messung der positiven Reziprozität. Zuerst werden daher nun die Zusammenhänge zwischen den einzelnen Items analysiert:

Tab. 17: Korrelationen zwischen den Items zur Messung der verinnerlichteten Reziprozitätsnorm

Wenn mir jemand einen Gefallen tut, bin ich bereit, dies zu erwidern.	1		
Ich strenge mich besonders an, um jemanden zu helfen, der mir früher schon mal geholfen hat.	0.409 (p = 0.000) n = 178	1	
Ich bin bereit, Kosten auf mich zu nehmen, um jemanden zu helfen, der mir früher einmal geholfen hat.	0.410 (p = 0.000) n = 178	0.654 (p = 0.000) n = 179	1
	Wenn mir jemand einen Gefallen tut, bin ich bereit, dies zu erwidern.	Ich strenge mich besonders an, um jemanden zu helfen, der mir früher schon mal geholfen hat.	Ich bin bereit, Kosten auf mich zu nehmen, um jemanden zu helfen, der mir früher einmal geholfen hat.

Quelle: eigene Daten.

Bei Betrachtung der Korrelationen kann ein starker Zusammenhang zwischen dem zweiten und dem dritten Item wahrgenommen werden ($r > 0.6$), wobei aber auch der Zusammenhang zum ersten Item insgesamt recht stark erscheint ($r > 0.4$). Aufgrund der hohen Korrelationen werden alle drei Items bei der Konstruktion eines Mittelwertindex berücksichtigt.¹²⁰ Dieser kann nun als globale Messung der verinnerlichteten Reziprozitätsnorm herangezogen werden. Mit Hilfe dieses Index werden nun die Mittelwerte zu den einzelnen Versuchsgruppen berichtet:

¹²⁰ Dieses Vorgehen wird auch nach Durchführung einer Reliabilitätsauch (Cronbachs Alpha von 0.745) gestützt.

Tab. 18: Mittelwerte für den Mittelwertindex der verinnerlichten Reziprozitätsnorm, aufgegliedert für die drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	6.03	60	0.79
Incentive 5€	6.05	60	0.74
Incentive 20€	6.18	59	0.77
Gesamtsumme	6.09	179	0.77

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.713$; $p = 0.492$.

Allgemein sind sehr hohe Mittelwerte zu erkennen (>6), da der Maximalwert der Skala für eine sehr hohe verinnerlichte Reziprozitätsnorm bei dem Wert „7“ liegt. Es ist hierbei unklar, ob die spezielle Befragungssituation dieser Studie mit den Studierenden für die hohen Werte verantwortlich ist, oder die Reziprozitätsnorm im Allgemeinen - wie Gouldner (1960) es unterstellt - sehr stark verinnerlicht ist. Die Mittelwertunterschiede zwischen den drei Versuchsgruppen erscheinen nicht hoch und weisen keine statistische Signifikanz auf ($p > 0.100$). Dies ist insofern erwartungskonform, als dass eine Verinnerlichung einer universellen Norm nicht durch eine spontane, un konditionale Incentivierung beeinflusst werden dürfte, sondern eher von langfristigen, sozialisierenden Faktoren abhängt.

- b) Die Bewertung des Versuchsleiters basiert auf zwei Items: der Freundlichkeit und der Hilfsbereitschaft. Diese Items weisen eine hohe, signifikante Korrelation auf ($r = 0.596$; $p = 0.000$)¹²¹, wobei die Kodierungen so vorgenommen wurden, dass hohe Werte auf eine positive Bewertung schließen lassen. Aus diesen zwei Variablen wurde ein Mittelwertindex berechnet. Aus diesem Index werden nun folgend die Mittelwerte für die drei Versuchsgruppen ermittelt:

¹²¹ Der Wert von Cronbachs Alpha beträgt 0.740.

Tab. 19: Mittelwerte für den Mittelwertindex der Bewertung des Versuchsleiters, aufgegliedert für die drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	4.60	60	0.49
Incentive 5€	4.73	59	0.46
Incentive 20€	4.75	60	0.46
Gesamtsumme	4.69	179	0.47

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 1.790$; $p = 0.170$.

Der Mittelwert der Versuchsgruppe ohne Incentive (4.60) weist bei einer 7stufigen Antwortskala auf eine tendenziell positive Bewertung des Versuchsleiters hin, wobei in den Versuchsgruppen mit Incentive leicht höhere Mittelwerte festgestellt werden können. Der Versuchsleiter wird damit aufgrund der Vergabe eines Incentives etwas positiver eingestuft. Statistisch signifikant ist hierbei jedoch nur der Unterschied zwischen der Versuchsgruppe ohne Incentive und der Versuchsgruppe mit einem Incentive in Höhe von 20€ ($p = 0.082$). Die Versuchsgruppe mit einem Incentive in Höhe von 5€ liegt im Vergleich zur Versuchsgruppe ohne Incentive nur knapp über der festgesetzten maximal akzeptierten Irrtumswahrscheinlichkeit von 10% ($p = 0.136$). Unter Berücksichtigung des Ergebnisses der einfaktoriellen Varianzanalyse kann jedoch kein signifikanter Unterschied über die drei Versuchsgruppen festgestellt werden ($F = 1.790$; $p = 0.170$). Die Ergebnisse entsprechen damit deskriptiv den vorher formulierten Erwartungen.

7.3 Der Indikator für die identifizierte Regulation (extrinsische Motivation)

Die identifizierte Regulation, und damit die extrinsische Motivation wurde ebenfalls über Selbsteinstufungen auf Basis der SIMS gemessen. Die extrinsische Motivation umfasst damit die Wirkung verschiedener internalisierter Normen, wie z.B. der Reziprozitätsnorm. Für eine

erste Analyse werden auch hier zuerst die Korrelationen zwischen den einzelnen Items betrachtet:

Tab. 20: Korrelationen zwischen den Items der extrinsischen Motivation

2. Weil ich es zu meinem eigenen Wohl tue.	1			
6. Weil ich die Teilnahme an Umfragen für gut für mich halte.	0.388 (p = 0.000) n = 175	1		
10. Aus persönlichem Entschluss.	0.025 (p = 0.745) n = 175	0.194 (p = 0.010) n = 174	1	
14. Weil ich glaube, dass eine Teilnahme wichtig für mich ist.	0.479 (p = 0.000) n = 173	0.554 (p = 0.000) n = 172	0.262 (p = 0.001) n = 172	1
	Weil ich es zu meinem eigenen Wohl tue.	Weil ich die Teilnahme an Umfragen für gut für mich halte.	Aus persönlichem Entschluss.	Weil ich glaube, dass eine Teilnahme wichtig für mich ist.

Quelle: eigene Daten.

Es ist zu erkennen, dass die Korrelationen zu dem dritten Item nicht nur tendenziell geringer ausfallen, sondern auch - mit Blick auf das erste Item - keine Signifikanz mehr aufweisen. Dies lässt sich über die Frageformulierung begründen, da das erste, zweite und vierte Item als Entscheidungsbegründungen interpretiert werden können, während das dritte Item hingegen Verantwortlichkeiten für die Teilnahme in den Vordergrund stellt. Dies wird auch über den Kommentar eines Befragten deutlich:

„Antwortmöglichkeit "Aus persönlichem Entschluss" ist eher unpräzise. Ist gemeint, weil ich mich dazu entschlossen habe? Das wäre etwas tautologisch, denn wenn ich nicht zwangsweise muss, habe ich mich ja immer dazu entschlossen. Die Betonung auf persönlich klingt ein wenig nach "ich persönlich finde das eine gute Sache, oder fühle mich dabei gut", insgesamt aber ist der Satz in diesem Kontext eher uneindeutig.“ (Ifdn = 400)

Es wurde deswegen auch hier eine Reliabilitätsanalyse gerechnet, wobei ein Auslassen des dritten Items zu einer Verbesserung der Reliabilität führt (ein Anstieg von Cronbachs Alpha von 0.650 auf 0.712). Das dritte Item wird daher in folgenden Analysen ausgelassen und nicht mehr berücksichtigt.

Auf Grundlage dieser Entscheidung wurde ein Mittelwertindex berechnet, welcher nur das erste, zweite und vierte Item berücksichtigt. Die Ergebnisse für die drei Versuchsgruppen sind in der folgenden Tabelle zusammengefasst:

Tab. 21: Mittelwerte für die stark zusammenhängenden Items der extrinsischen Motivation, aufgegliedert für die drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	2.81	58	1.24
Incentive 5€	3.20	58	1.18
Incentive 20€	3.46	60	1.46
Gesamtsumme	3.16	176	1.32

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 3.805$; $p = 0.024$.

Es kann festgestellt werden, dass mit steigender Höhe des Incentives die identifizierte Regulation und damit die extrinsische Motivation steigt. Die Differenzen zwischen der Versuchsgruppe ohne Incentive mit den Versuchsgruppen mit Incentive sind hierbei auch (annähernd)

signifikant (5€: $p = 0.105$; 20€: $p = 0.007$). Die Differenz zwischen den Versuchsgruppen mit Incentive weist jedoch keine Signifikanz auf ($p = 0.274$). Damit spricht der Befund erwartungsgemäß für eine positive Wirkung der Incentives auf die extrinsische Motivation.

8. Erste Analysen zu den abhängigen Variablen

Zur Messung der Antwortqualität sind Indikatoren für die einzelnen definitorischen Komponenten *durchdacht, (situational) wahrheitsgemäß, vollständig und anweisungsbefolgend* erhoben worden. Auch diese sollen nun auf Zusammenhänge untereinander und innerhalb der drei Versuchsgruppen geprüft werden. Es soll dabei analog zu Kapitel 7 ergänzend eine empirische Auswahl an geeigneten Operationalisierungen für die später folgenden multivariaten Analysen vorgenommen werden.

8.1 Analyse der Indikatoren für ein durchdachtes Bearbeiten des Fragebogens:

8.1.1 Akquieszenz und Status Quo-Effekt

Als erste Indikatoren für ein durchdachtes Bearbeiten des Fragebogens werden die Akquieszenz und der Status Quo-Effekt betrachtet.

a) Akquieszenz

Gemäß der Operationalisierung aus Kapitel 6.2.1 wurden zur Messung der Akquieszenz die Ankreuzmuster zu den Bewertungsfragen der verschiedenen Drogen analysiert. Es weisen hierbei nur neun Befragte Akquieszenz auf und damit kann von einem eher differenzierten Ankreuzmuster bei den restlichen Befragten ausgegangen werden. Dies weist damit prinzipiell auf ein geringes Satisficing bei den Befragten hin.

Unter Kontrolle der Versuchsgruppen kann kein signifikanter Zusammenhang festgestellt werden, wobei dies durch die geringe Fallzahl begründet werden kann. Als Erweiterung der Messung werden daher auch die beiden Skalen zur Politikerbewertung herangezogen. Mit Bezug auf das Ankreuzverhalten konnte bei sechs Befragten Akquieszenz identifiziert werden. Die drei Messungen der Akquieszenz zeigen jedoch keine hohe interne Konsistenz auf (Cronbachs Alpha = 0.463), so dass eine Zusammenführung in einem additiven Index empirisch nicht gestützt ist.¹²²

Aufgrund der sehr geringen Streuung der ursprünglichen Messung und den Unstimmigkeiten bezüglich der Konsistenz zu den erweiterten Messungen, soll Akquieszenz in dieser Studie nicht als Indikator für ein durchdachtes Bearbeiten des Fragebogens in den folgenden multivariaten Analysen genutzt werden.

b) Status Quo-Effekt

Insgesamt wählten 32.2% der Befragten die Kategorie, welche dem Status Quo entspricht. Unter Betrachtung der drei Versuchsgruppen ergeben sich keine signifikanten Unterschiede bezüglich des Status Quo-Effekts ($\chi^2 = 0.699$; $df = 2$; $p = 0.705$). Die Häufigkeitsverteilung innerhalb der drei Versuchsgruppen ist folgend tabellarisch dargestellt:

¹²² Die geringe interne Konsistenz kann auf inhaltlicher Ebene erklärt werden: Die Bewertungen der Drogen-Fragen werden auf einer Zustimmungsskala gemessen, während die Bewertungen der Politiker auf einer bipolaren Skala erfasst werden.

Tab. 22: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen dem Status Quo-Effekt und der Incentivierung

		Kein Incentive	Incentive: 5 Euro	Incentive: 20 Euro	Gesamt
Status Quo-Effekt	Status Quo gewählt (0)	17 (28.3%)	21 (35.0%)	20 (33.9%)	58 (32.4%)
	Nicht Status Quo gewählt (1)	43 (71.7%)	39 (65.0%)	39 (66.1%)	121 (67.5%)
	Gesamt	60 (100%)	60 (100%)	59 (100%)	179 (100%)

Quelle: eigene Daten.

Der Status Quo-Effekt erscheint aufgrund der theoretischen Herleitung und Konzeption als angemessener Indikator für ein durchdachtes Bearbeiten des Fragebogens.

Werden die beiden Messungen für Satisficing (Akquieszenz und Status Quo-Effekt) im Zusammenhang betrachtet, so weist die Korrelation zwar das erwartete Vorzeichen auf ($r = -0.114$), ist allerdings nicht signifikant von Null verschieden ($p = 0.129$).

8.1.2 Die genutzte Erinnerungsstrategie zur Beantwortung von Häufigkeitsfragen

Zur Beantwortung der Häufigkeitsfrage: „Wie oft haben Sie in den vergangenen 4 Wochen Sport getrieben?“ können verschiedene Erinnerungsstrategien genutzt werden. Im Kapitel zur Operationalisierung wurden vier Fragen vorgestellt, welche unterschiedliche Strategien zur Erinnerung messen sollen:

Tab. 23: Häufigkeitsverteilung von Erinnerungsstrategien zur Beantwortung einer Häufigkeitsfrage

	Antworten		Prozent der Fälle
	Häufigkeit	Prozent	
Ich habe an jedes einzelne Mal gedacht, wo ich Sport getrieben habe.	110	35.30%	61.50%
Ich habe daran gedacht wie oft ich normalerweise Sport treibe.	84	26.90%	46.90%
Ich habe an die verschiedenen Sportarten gedacht, die ich getrieben habe.	69	22.10%	38.50%
Ich habe es auf Basis meines generellen Eindrucks geschätzt.	49	15.70%	27.40%
Gesamtsumme	312	100.00%	174.30%

Quelle: eigene Daten. Es waren Mehrfachnennungen möglich.

Insgesamt gaben 35 Befragte an ausschließlich die Erinnerungsstrategie „Ich habe an jedes einzelne Mal gedacht, wo ich Sport getrieben habe.“ genutzt zu haben.¹²³ Diese wird gemäß der Operationalisierung als Anzeichen für ein durchdachtes Beantworten der Häufigkeitsfragen interpretiert und steht damit für eine hohe Antwortqualität. Zuerst soll nun geprüft werden, ob und inwiefern sich die ausschließliche Nutzung der Erinnerungsstrategie unter Berücksichtigung der drei Versuchsgruppen ändert. Hierfür wurde eine Mehrfeldertabelle erstellt, wobei die Versuchsgruppe als Spaltenvariable festgesetzt wurde:

Tab. 24: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen der Erinnerungsstrategie und der Incentivierung

		Kein Incentive	Incentive: 5 Euro	Incentive: 20 Euro	Gesamt
kognitiv aufwendige Erinnerungsstrategie	Aufwendige Strategie nicht gewählt (0)	53 (88.3%)	48 (80.0%)	44 (73.3%)	145 (80.6%)
	Aufwendige Strategie gewählt (1)	7 (11.7%)	12 (20%)	16 (26.7%)	35 (19.4%)
	Gesamt	60 (100%)	60 (100%)	60 (100%)	180 (100%)

Quelle: eigene Daten.

¹²³ Es wurde eine binäre Variable erstellt, wobei der Wert 1 vergeben wird, wenn die Befragten ausschließlich die erste Erinnerungsstrategie gewählt haben. Bei Wahl dieser Erinnerungsstrategie wird folglich ein höher kognitiver Aufwand bewältigt, was wiederum auf eine höhere Antwortqualität hinweist.

Nach Berechnung von Chi-Quadrat kann kein signifikanter Zusammenhang ($\chi^2 = 4.327$; $df = 2$; $p = 0.115$) festgestellt werden. Ein hohes Incentive geht demnach nur deskriptiv mit der Wahl einer kognitiv aufwendigen Erinnerungsstrategie einher. Dies könnte darauf hinweisen, dass die Antwortqualität bei Vergabe eines Incentives steigen kann.¹²⁴

Die ausschließliche Wahl der ersten Erinnerungsstrategie weist keine Korrelation zu dem Status Quo oder der Akquieszenz auf.¹²⁵ Nur bei Kontrolle der Versuchsgruppen kann ein einziger signifikanter Effekt zwischen dem Status Quo-Effekt und der Erinnerungsstrategie in der Versuchsgruppe ohne Incentive festgestellt werden ($r = 0.229$; $p = 0.079$). Dieser Effekt entspricht den theoretischen Vorüberlegungen, da ein Verneinen des Status Quo mit der Wahl einer komplexen Erinnerungsstrategie einhergeht. Das nur einer der Indikatoren in einer spezifischen Versuchsgruppe signifikante Korrelationen aufweist, widerspricht jedoch der ursprünglichen Einordnung der Erinnerungsstrategie als Symptom von Satisficing und wird daher nicht als Indikator für ein durchdachtes Bearbeiten des Fragebogens herangezogen.¹²⁶

8.1.3 Indikatoren für ein durchdachtes Bearbeiten des Fragebogens: Konsistenz

Die Konsistenz wurde über zwei Reichweiten im Fragebogen gemessen: zum einen nach kurzer zeitlicher Abfolge, d.h. die Konsistenzmessungen erfolgten auf einer einzigen Fragebogenseite und zum anderen nach langer zeitlicher Abfolge, d.h. die Konsistenzmessung erfolgte über mehrere Fragebogenseiten hinweg.

¹²⁴ Wenn die strikte Kodiervorgabe der ausschließlichen Wahl der ersten Erinnerungsstrategie aufgelöst wird und auch die Befragten berücksichtigt werden, welche neben der komplexen Erinnerungsstrategie noch mindestens eine weitere, beliebige wählten, dann ist der Zusammenhangsrichtung noch immer erwartungsgemäß, aber nicht signifikant ($\chi^2 = 2.431$; $df = 2$; $p = 0.297$).

¹²⁵ Siehe Anhang, S. 206.

¹²⁶ Alternativ kann argumentiert werden, dass die Messung der Akquieszenz und des Status Quo-Effekt fehlerhaft sind und deswegen keine erwarteten korrelativen Beziehungen bestehen. Da die Operationalisierung der genutzten Erinnerungsstrategie als Symptom von Satisficing noch nicht erprobt ist, wird an dieser Stelle die dazugehörige Messung angezweifelt.

1) Konsistenz bei kurzer zeitlicher Abfolge

Hierzu gehören vier Messungen zu je zwei Frageeinheiten. Das erste Messpaar umfasst die Fragen wie viele Stunden die Befragten durchschnittlich wach sind bzw. schlafen. Es wird erwartet, dass bei einem sorgfältigen Bearbeiten die Summe der beiden Durchschnittswerte „24“ lautet und damit der gesamte Tageszyklus abgedeckt wird.¹²⁷ Das zweite Messpaar umfasste die Fragen wie viel Prozent des aktuellen Studiums schon absolviert wurden bzw. noch fehlen.¹²⁸ Hier wird erwartet, dass die Summe der Prozentzahlen 100% ergibt. Werden nun die Konsistenzwerte berechnet, so ergeben sich folgende Verteilungen:

Tab. 25: Häufigkeitstabelle über die zwei Konsistenzmessungen bei kurzer zeitlicher Abfolge

	Konsistenz: 24 h Tag	Konsistenz: 100% Studium
Nein (= 0)	44 (24.7%)	19 (11.1%)
Ja (= 1)	134 (75.3%)	152 (88.9%)
Gesamtsumme	178 (100%)	171 (100%)

Quelle: eigene Daten.

Zwischen den beiden Messungen für Konsistenz liegt kein statistisch signifikanter Zusammenhang vor ($r = -0.075$; $p = 0.335$). Darüber hinaus weist die Korrelation ein negatives Vorzeichen auf, so dass die zwei Messungen tendenziell einander eher widersprechen: Ein konsistentes Beantworten der ersten Frage tritt eher mit einem inkonsistenten Beantworten der zweiten Frage - und umgekehrt - auf.¹²⁹ Dies kann, neben Tippfehlern bei der Eingabe und einer allgemeinen Unkonzentriertheit auch daran lie-

¹²⁷ Die Korrelation zwischen dem ersten Messpaar beträgt $r = -0.666$ ($p = 0.000$) und kann damit als hoch angesehen werden. Dies weist auf eine starke Konsistenz zwischen den Messungen hin.

¹²⁸ Die Korrelation zwischen dem zweiten Messpaar beträgt $r = -0.893$ ($p = 0.000$) und kann damit als sehr hoch angesehen werden. Dies weist auf eine sehr starke Konsistenz zwischen den Messungen hin.

¹²⁹ Aus diesem Grund wird aus den beiden Konsistenzmessungen kein Index gebildet.

gen, dass die Fragen zu global gestellt wurden und die Befragten daraufhin nicht sinnvoll antworten konnten. Dies wird auch durch die hinterlassenen Kommentare der Befragten deutlich:

„Ich fand die Beantwortung der aller vier Fragen sehr schwierig, da es bei mir sehr hohe Unterschiede zwischen einem normalen Wochentag und dem Wochenende gibt. Das Schlafdefizit, das ich unter der Woche habe, kompensiere ich am Wochenende durch das besonders lange ausschlafen. Frage ist jetzt, was die Durchschnittzeit angeben soll. Wie lange ich während der Unizeit schlafe? Oder insgesamt mit dem Wochenende?“ (lfdn = 362)

„Die Frage nach der durchschnittlichen Schlafdauer könnte man evtl. aufteilen in "unter der Woche" und "am Wochenende", oder sogar in "während des Semesters" und "in den Semesterferien". Wenn ich drei Tage lang chronisch übermüdet bin und dann drei Tage lang ausschlafe gleicht sich das ja wieder aus.“ (lfdn = 400)

2) Konsistenz bei langer zeitlicher Abfolge

Zur Messung der Konsistenz über weite Distanz wurde ein Fragenkatalog zur Bewertung von Politikern zu Beginn der Befragung gestellt und am Ende der Befragung wiederholt. Bei der Bewertung der Sympathie von Politikern kann bei einem durchdachten

Beantworten davon ausgegangen werden, dass die Einstellungen innerhalb der Befragung stabil bleiben und sich folglich bei einer Messwiederholung am Ende des Fragebogens replizieren lassen. Zur Vermeidung von Kontexteffekten, wurden im Verlauf der Befragung nur das politische Interesse und die zukünftige Wahlabsicht abgefragt. Diese Fragen stehen durchschnittlich 30 Minuten vor den Wiederholungsfragen.

Zur Messung der Konsistenz sollten zehn verschiedene Politiker, auf einer Sympathie-Skala von 1 (= sehr unsympathisch) bis 7 (= sehr sympathisch) bewertet werden. Die Abweichungen aller Erst- und der Zweitmessungen wurden als absolute Differenz definiert und folgend aufsummiert:

Tab. 26: Häufigkeitsverteilung der Summe der Abweichungen aller Politikerbewertungen zu zwei Messzeitpunkten¹³⁰

Abweichung	Häufigkeit	Gültige Prozent	Kumulative Prozente
0	26	22.2	22.2
1	24	20.5	42.7
2	23	19.7	62.4
3	12	10.3	72.6
4	10	8.5	81.2
5	10	8.5	89.7
6	6	5.1	94.9
7	1	0.9	95.7
9	3	2.6	98.3
10	2	1.7	100.0
Gesamtsumme	117	100.0	

Quelle: eigene Daten.

Es ist ersichtlich, dass 22.2% aller gültigen Fälle exakt dieselben Angaben machten und damit ein sehr hohes Maß an Konsistenz aufweisen. 50% der Befragten weisen eine

¹³⁰ 63 Befragte können in der Analyse nicht berücksichtigt werden, da sie nicht alle Fragen beantwortet haben und infolge dessen als Missing Value behandelt werden. Die Bewertung des fiktiven Raphael Zastel wird bei der Erstellung des Summenindex nicht berücksichtigt, da eine Instabilität auf eine Verbesserung der Antwortqualität zurückgeführt werden kann (begründeter Übersprung). Wird der Summenindex mit Berücksichtigung von Raphael Zastel berechnet, so sinkt die gültige Fallzahl auf 96, wobei die Quartilwerte erhalten bleiben.

Abweichung von zwei Skalenpunkten und 75% von 4 Skalenpunkten auf. Dies bedeutet, dass 75% der zugrunde liegenden Probanden bei 10 Bewertungen, welche jeweils eine siebenstufige Antwortskala aufweisen insgesamt nur vier Skalenpunkte abweichen. Dies weist augenscheinlich auf eine hohe Konsistenz hin. Der oben dargestellte Befund wird durch die Test-Retest-Korrelationen gestützt: die niedrigste Korrelation lag mit 0.865 ($p = 0.000$) bei der Sympathieeinstufung zu Peter Altmaier, wobei die höchste Korrelation mit 0.962 ($p = 0.000$) bei der Bewertung von Angela Merkel gemessen wurde.¹³¹

Werden nun die einzelnen Konsistenzmessungen kurzer und weiter Distanz miteinander in Beziehung gesetzt, so ergibt sich folgendes Bild:

Tab. 27: Korrelationen zwischen den Messungen für Konsistenz bei kurzer und langer zeitlicher Abfolge.

Konsistenz: 24 h Tag	1		
Konsistenz: 100% Studium	-0.075 ($p = 0.335$) n = 169	1	
Konsistenz: Politikerbewertung	-0.093 ($p = 0.368$) n = 96	-0.027 ($p = 0.799$) n = 92	1
	Konsistenz: 24 h Tag	Konsistenz: 100% Studium	Konsistenz: Politikerbewertung

Quelle: eigene Daten.

¹³¹ Es muss allerdings angemerkt werden, dass aufgrund der Konstruktion der Variable nur Befragte berücksichtigt werden, welche keine Fragen übersprungen haben und damit eine höhere Antwortqualität aufweisen. Übersprungen wurden Fragen besonders bei den Politikern Alexander Dobrindt, Heiko Maas, Andrea Nahles und Peter Altmaier. Die Befragten geben hierbei in den Kommentaren an, dass sie diese Politiker nicht kennen, bzw. nicht einschätzen können. Die Konsistenz kann daher das Ergebnis eines nachträglichen Selektionseffektes sein, welcher die Befragten mit einer niedrigen Antwortqualität ausschloss. Es muss aber deutlich darauf hingewiesen werden, dass auch die Befragten mit begründeten Übersprungen ausgeschlossen werden, da sie nicht alle Fragen beantwortet haben. Dies bedeutet, dass die Gruppe der Missing Values keine Homogenität bezüglich der Antwortqualität aufweist.

Es kann kein einheitliches Muster unter den einzelnen Konsistenzmessungen festgestellt werden: die Korrelation zu den Messungen mit kurzer zeitlicher Abfolge weist ein gegenläufiges Vorzeichen auf und mit Bezug auf die Konsistenzmessung mit langer zeitlicher Abfolge können auch keine konsistenten Strukturen beobachtet werden. Bei Kontrolle der drei Versuchsgruppen kann ebenfalls kein (signifikanter) Zusammenhalt zwischen den einzelnen Messungen festgestellt werden. Aufgrund des mangelnden internen Zusammenhangs der einzelnen Messungen untereinander, wird die Konsistenz, gemessen über die Indikatoren mit kurzer und langer zeitlicher Abfolge nicht als Maß für die Antwortqualität herangezogen.

8.1.4 Anzahl an Worten in offenen Fragen

Die Anzahl der Worte in offenen Fragen wurde ebenfalls als Indikator für Antwortqualität herangezogen, wobei ein „Mehr“ an Worten auch als ein „Mehr“ an kognitiven Aufwand interpretiert werden kann. Zur Messung wurden zwei offene Fragen ausgewählt: die erste offene Frage („Bitte überlegen Sie sich mindestens zwei Möglichkeiten, die Sie sehen um die Antwortqualität von Befragten bei Umfragen zu verbessern.“) wurde durchschnittlich mit 20.3 Worten, bzw. 145.6 Zeichen beantwortet. Die zweite Frage („Bitte erläutern Sie daher die Gründe für Ihre Teilnahme.“) erreichte durchschnittlich 29.5 Worte und 193.2 Zeichen. Erweitert kann auch die Anzahl der Worte in den Kommentaren in die Analysen einbezogen werden. Hierbei wurden durchschnittlich 65.4 Worte und 420.6 Zeichen geschrieben.

Zuerst werden nun die Zusammenhänge zwischen der Anzahl der Worte zu den einzelnen offenen Fragen, bzw. den Kommentaren betrachtet:

Tab. 28: Korrelationen der Worte zwischen den drei offenen Fragen

Verbesserung Antwortqualität	1		
Motivationsgründe für Studienteilnahme	0.289 (p = 0.000) n = 179	1	
Worte in den Kommentaren	0.109 (p = 0.148) n = 179	0.130 (p = 0.083) n = 179	1
	Verbesserung Antwortqualität	Motivationsgründe für Studienteilnahme	Worte in den Kommentaren

Quelle: eigene Daten.

Die Anzahl der Worte korrelieren mit mittlerer Zusammenhangsstärke, wobei zwischen der ersten offenen Frage und der Anzahl an Worten in den Kommentaren kein signifikanter Zusammenhang nachgewiesen werden kann. Die Vorzeichen entsprechen den Erwartungen: Die Wortanzahl in der ersten offenen Frage hängt positiv mit der Wortanzahl in der zweiten offenen Frage zusammen, ebenso die Anzahl der Worte in den Kommentaren. Es kann an dieser Stelle jedoch der Einwand geäußert werden, dass nicht die Worte, sondern vielmehr die Anzahl der genutzten Anschläge entscheidend für die Messung eines Aufwands sind. Daher werden nun folgend die Zusammenhänge für die Anzahl an Zeichen betrachtet:

Tab. 29: Korrelationen der Zeichen zwischen den drei offenen Fragen

Verbesserung Antwortqualität	1		
Motivationsgründe für Studienteilnahme	0.282 (p = 0.000) n = 179	1	
Zeichen in den Kommentaren	0.099 (p = 0.188) n = 179	0.136 (p = 0.069) n = 179	1
	Verbesserung Antwortqualität	Motivationsgründe für Studienteilnahme	Zeichen in den Kommentaren

Quelle: eigene Daten.

Die Messungen der Anzahl der Worte und der Zeichen erscheinen äquivalent, da die Zusammenhangsstärken und Signifikanzen stabil wirken. Auch wenn die Korrelationen nicht sehr

hoch sind, werden zwei additive Indices für die Gesamtzahl an Worten und die Gesamtzahl an Zeichen berechnet. Die Anzahl der Worte bzw. der Zeichen in den Kommentaren wird allerdings nicht in die Indices eingerechnet, da sie nicht als Ergebnis einer direkten Frage geschrieben wurden, sondern als Notwendigkeit verstanden werden um Fehler bzw. Verbesserungen im gesamten Fragebogen aufzuzeigen.

Als nächstes wird der Zusammenhang zwischen dem gebildeten Summenindex der Anzahl an Worten und der Höhe des Incentives untersucht:

Tab. 30: Mittelwerte des Gesamtindex für die Anzahl der Worte, aufgliedert nach den drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	41.67	60	27.59
Incentive 5€	55.37	60	44.62
Incentive 20€	52.75	59	27.36
Gesamtsumme	49.91	179	34.52

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 2.710$; $p = 0.069$.

Bei Vergabe eines Incentives kann stets ein signifikanter Anstieg der Worthäufigkeiten gegenüber der Versuchsgruppe ohne Incentive festgestellt werden (5€: $p = 0.030$; 20€: $p = 0.079$). Damit liegt ein positiver Effekt der Incentives auf die Antwortqualität vor. Dies kann dadurch erklärt werden, dass sich die Befragten aufgrund des Incentives reziprok verhalten und mehr Worte in den offenen Fragen formulieren. Bei genauerer Betrachtung der durchschnittlichen Worthäufigkeiten kann ein etwas geringerer Effekt in der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro beobachtet werden. Der Mittelwert der zweiten Versuchsgruppe ist jedoch nicht signifikant vom Mittelwert der dritten Versuchsgruppe verschieden und die Differenz muss daher vorsichtig interpretiert werden.

Es werden nun auch die Mittelwerte für die Anzahl der Worte in den Kommentaren bei Berücksichtigung der Versuchsgruppen berechnet:

Tab. 31: Mittelwerte der Anzahl der Worte in den Kommentaren, aufgliedert nach den drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	49.27	60	75.69
Incentive 5€	77.78	60	137.25
Incentive 20€	69.37	59	134.15
Gesamtsumme	65.37	179	118.96

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.905$; $p = 0.407$.

Trotz der augenscheinlichen Variation der Mittelwerte, können keine statistisch signifikanten Unterschiede festgestellt werden. Dies lässt sich über die hohe Streuung der Variablen erklären:

Tab. 32: Metrische Verteilungsinformationen zu der Anzahl der Worte in den Kommentaren

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Standardabweichung</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Häufigkeit</i>
Kein Incentive	49.27	75.69	0.00	469.00	60
Incentive 5€	77.78	137.25	0.00	934.00	60
Incentive 20€	69.37	134.15	0.00	973.00	59

Quelle: eigene Daten.

Es gibt Befragte die bis zu 973 Worte über die Kommentarfunktion geschrieben haben. Aufgrund der hohen Streuung ist es schwierig zu signifikanten Ergebnissen zu gelangen, da auch die Standardfehler aus der Streuung der Variablen berechnet werden. Als nächster Schritt soll sich daher zur Prüfung eines möglichen Unterschieds der Mann-Whitney U-Test herangezogen. Dieser weist auch keine signifikanten Effekte aus, wobei der Unterschied zwischen der Versuchsgruppe ohne Incentive und der Versuchsgruppe mit einem Incentive in Höhe von 5 Euro nur leicht über der maximal akzeptierten Irrtumswahrscheinlichkeit von 10 % liegt ($p = 0.132$).

Die Anzahl an Worten bzw. Zeichen in den Kommentaren soll nicht als Indikator für ein durchdachtes Bearbeiten des Fragebogens gewählt werden, da die Messung auch weitere Facetten

der Antwortqualität umfasst. So kann beispielhaft ein Überspringen im Fragebogen durch hinterlassene Kommentare begründet werden oder das Verweigern des Befolgens von Anweisungen in den Kommentaren erklärt werden. Die Anzahl der Worte in den offenen Fragen erscheinen dagegen eher angemessen.

8.1.5 Zusammenfassung der Indikatoren eines durchdachten Bearbeitens eines Fragebogens

In den vorherigen Analysen wurden verschiedene Indikatoren für ein durchdachtes Bearbeiten des Fragebogens vorgestellt und diskutiert. Da nicht alle Indikatoren in den multivariaten Analysen genutzt werden können, soll nun eine theoretisch und empirisch begründete Auswahl getroffen werden. In den vorherigen Kurzanalysen wurden bereits einige Indikatoren als weniger verlässlich deklariert, so dass folgend lediglich der Status Quo-Effekt und die Anzahl an geschriebenen Worten gegeneinander abgewogen werden sollen.

Der Status Quo-Effekt wurde von Krosnick als Symptom für ein starkes Satisficing definiert, da der benötigte kognitive Aufwand beim vollständigen Durchlaufen des Antwortprozesses als sehr hoch angesehen wird und damit Phasen des kognitiven Antwortprozesses (vgl. Tourangeau & Rasinski (1988)) übersprungen werden können.

Problematisch ist jedoch, dass die Wahl des Status Quo eine tatsächliche Meinung darstellen kann, welche nach reiflicher Überlegung gebildet wurde. Dies bedeutet, dass nur schwer unterschieden werden kann, ob eine Satisficing-Strategie oder ein komplett durchlaufener kognitiver Antwortprozess zugrunde liegt. Aus diesem Grund wurde die Messung so konzipiert, dass die Befragten bei Nicht-Wahl des Status Quo zusätzlich eine schriftliche Begründung für ihr Antwortverhalten abgeben sollten. Damit wurde der Aufwand zur Beantwortung weiter

erhöht und die Wahl der Kategorie des Status Quo tritt deutlicher als Möglichkeit zur Aufwandsminimierung in den Vordergrund. Aufgrund dessen lässt sich zwar noch immer nicht unterscheiden, ob Satisficing oder tatsächliche Meinungsbildung vorliegt, aber die Interpretation im Sinne des Status Quo-Effektes erscheint dadurch etwas sicherer.

Werden die Zusammenhänge des Status Quo-Effekt mit den anderen Indikatoren für ein durchdachtes Bearbeiten berechnet, so kann festgestellt werden, dass die Korrelationen insgesamt recht schwach sind, aber tendenziell die erwarteten Vorzeichen aufweisen.¹³²

Die Anzahl an geschriebenen Worten wird oft mit einem höheren kognitiven Aufwand in Verbindung gebracht, da mit steigender Wortzahl auch eine steigende Informationsverarbeitung vermutet wird. Die (implizite) Interpretation lautet daher: je mehr Worte formuliert werden, desto reichhaltiger ist die Antwort. Demzufolge liegt mit steigender Wortzahl auch eine Steigerung der Antwortqualität vor. Diese Annahme überdeckt jedoch die Möglichkeit, dass auch wenige Worte als Anzeichen für eine erhöhte Antwortqualität interpretiert werden können. Dies ist gegeben, wenn die Antworten sehr präzise und auf den Punkt gerichtet sind. Aus dieser Perspektive würde eine Zunahme an Worten gegen eine hohe Antwortqualität sprechen. Bei Betrachtung der Korrelationen können auch hier für die Zusammenhänge zwischen der Anzahl an Worten und den anderen Indikatoren für ein durchdachtes Bearbeiten des Fragebogens tendenziell die erwarteten Vorzeichen beobachtet werden.¹³³ Der Zusammenhang zu einer Konsistenzmessung kurzer Reichweite weist hierbei sogar ein signifikantes Niveau auf.

¹³² Siehe Anhang, S. 206.

¹³³ Siehe Anhang, S. 206.

Auf Basis der oben dargestellten Vor- und Nachteile wird nun ein Indikator ausgewählt: Da zum einen eine deutliche theoretische Fundierung für den Zusammenhang zwischen den Worthäufigkeiten und der Antwortqualität fehlt und zum anderen die Häufigkeiten nur unter starken Unsicherheiten zu interpretieren sind, erscheint der Status Quo-Effekt trotz seiner oben aufgezeigten Schwächen als angemessener und soll damit in der multivariaten Analyse als Indikator für ein durchdachtes Bearbeiten berücksichtigt werden.

8.2 Analyse der Indikatoren für ein wahrheitsgemäßes Bearbeiten des Fragebogens

8.2.1 Die Äußerung von Pseudo-Opinions (Falschangaben)

Im Fragebogen sind sechs Fragen enthalten, die die Abgabe von Pseudo-Opinions aufdecken sollen. Zwei Fragen beziehen sich auf die Bewertung der nicht existenten Droge LA-42, zwei weitere Fragen auf die Bewertung des fiktiven Politikers Raphael Zastel, eine Frage mit einer Bewertung von nicht angezeigten Deckblättern und die letzte Frage auf die Kenntnis der Droge LA-42. Zuerst sollen die Zusammenhänge zwischen den sechs Messungen geprüft werden:

Tab. 33: Korrelationen zwischen den Messungen der Pseudo-Opinions

Bewertung LA-42 I	1					
Bewertung LA-42 II	0.915 (p = 0.000) n = 180	1				
Bewertung Politiker I	0.667 (p = 0.000) n = 180	0.635 (p = 0.000) n = 180	1			
Bewertung Politiker II	0.559 (p = 0.000) n = 179	0.529 (p = 0.000) n = 179	0.671 (p = 0.000) n = 179	1		
Bewertung Deckblatt	0.293 (p = 0.000) n = 180	0.305 (p = 0.000) n = 180	0.289 (p = 0.000) n = 180	0.266 (p = 0.000) n = 179	1	
Kenntnis LA-42	0.226 (p = 0.002) n = 180	0.235 (p = 0.002) n = 180	0.100 (p = 0.182) n = 180	0.124 (p = 0.099) n = 179	0.175 (p = 0.019) n = 180	1
	Bewertung LA-42 I	Bewertung LA-42 II	Bewertung Politiker I	Bewertung Politiker II	Bewertung Deckblatt	Kenntnis LA-42

Quelle: eigene Daten.

Die berechneten Korrelationen weisen ein einheitliches Vorzeichen auf, sind jedoch nicht immer statistisch signifikant. Die Falschangaben zur Kenntnis der Droge LA-42 und der Deckblattbewertung weisen, im Vergleich zu den anderen Variablen eher schwächere Korrelationen auf. Die Zusammenhänge zwischen der Falschangabe zur Kenntnis von LA-42 und den Bewertungen des fiktiven Politikers sind sogar so gering, dass diese nicht signifikant sind, bzw. nur gerade noch unter der maximal akzeptierten Irrtumswahrscheinlichkeit von 10% liegen.

Die schwächeren Zusammenhänge zur Falschangabe der Kenntnis der Droge LA-42 können dadurch erklärt werden, dass aufgrund einer Falschangabe die Filterfunktion zu einem hohen (zu erwartenden) Bearbeitungsaufwand führt. Bei den anderen Messungen wird der Bearbeitungsaufwand im Fragebogen bei einer Falschantwort hingegen tendenziell verringert, da dann keine Begründung für einen Übersprung in den Kommentaren verfasst werden muss.

Die geringen Zusammenhänge zur Bewertung der nicht dargestellten Deckblätter lassen sich wiederum dadurch erklären, dass im Vergleich zu den anderen vier Items der Pseudo-Bewertungen eine abweichende Darstellungsform gewählt wurde und die Befragten deshalb im Antwortverhalten variieren. Zum anderen ist es denkbar, dass die Bewertung von nicht dargestellten Deckblättern eine deutlich erkennbare Falschantwort ist. Dies wird dadurch unterstützt, dass neben der Deckblattbewertung keine weiteren Items auf Fragebogenseite abgefragt wurden.

Auf Grundlage der Korrelationen wird nun ein Summenindex für die Häufigkeit an Pseudo-Opinions erstellt. Trotz der geringeren Zusammenhänge zur Deckblattbewertung wird die Messung für den Summenindex berücksichtigt, während die Falschangabe zur Kenntnis von LA-42 aufgrund der berichteten inhaltlichen Gegensätzlichkeiten nicht in die Berechnung eingeht. Die Häufigkeiten des Summenindex werden in der folgenden Tabelle zusammengefasst:

Tab. 34: Häufigkeiten der Pseudo-Opinions

Pseudo-Opinion	Häufigkeit	Gültige Prozent	Kumulative Prozente
0	36	20.1	20.1
1	16	8.9	29.1
2	17	9.5	38.5
3	15	8.4	46.9
4	67	37.4	84.4
5	28	15.6	100
Gesamtsumme	179	100	

Quelle: eigene Daten.

36 Befragte haben keine Bewertung über einen fiktiven Sachverhalt abgegeben und weisen damit eine hohe Antwortqualität auf, jedoch haben 50% der Befragten bei fünf möglichen Falschantworten bereits vier Falschantworten gegeben. Dies weist auf eine tendenziell eher geringe Antwortqualität hin.

Im nächsten Schritt werden die Mittelwerte des Summenindex für die drei Versuchsgruppen analysiert. Es kann ein leichter Anstieg der Falschangaben mit steigendem Incentive beobachtet werden:

Tab. 35: Die Mittelwerte der Falschangaben, aufgliedert nach den drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	2.67	60	1.91
Incentive 5€	2.73	60	1.83
Incentive 20€	3.03	59	1.66
Gesamtsumme	2.81	179	1.80

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 0.698$; $p = 0.499$.

Die Mittelwertunterschiede sind nicht signifikant. Es ist jedoch überraschend, dass der höchste Mittelwert in der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro vorzufinden ist. Der Anstieg ist zwar statistisch nicht signifikant ($p = 0.268$), könnte aber darauf hindeuten, dass ein sehr hohes Incentive die Antwortqualität verschlechtert. Genauere Analysen werden diesbezüglich in Kapitel 9.1 durchgeführt.

Die Messungen der Pseudo-Opinions erscheinen als Indikator für ein wahrheitsgemäßes Bearbeiten des Fragebogens insgesamt angemessen.

8.2.2 Soziale Erwünschtheit

Gemäß der Ausführungen in Kapitel 6.2.3 soll die soziale Erwünschtheit als Indikator für ein wahrheitsgemäßes Bearbeiten in Betracht gezogen werden. Hierfür wurden das Messkonzept des Balanced Inventory of Social Desirability (BIDR) genutzt.

Der BIDR misst zwei Dimensionen der sozialen Erwünschtheit: die Selbsttäuschung (SDE) und die Fremdtäuschung (IM), wobei die Fremdtäuschung in dieser Studie als primärer Indikator

für ein wahrheitsgemäßes Bearbeiten interpretiert wird, da diese auf bewussten Entscheidungen beruht und folglich von der Vergabe eines Incentives abhängig sein kann. Die Selbsttäuschung ist gemäß Paulhus (1984) auf unbewusste Mechanismen zurückzuführen und sollte sich daher, wie ein Persönlichkeitsmerkmal, stabil gegenüber einem Incentives verhalten. Die Dimensionen der Selbst- und Fremdtäuschung gelten als unabhängig voneinander und dürfen daher nicht korrelieren. Dieses Strukturmerkmal ist in den Daten vorhanden, da die geringe Korrelation zwischen der Selbst- und Fremdtäuschung ($r = 0.082$) keine Signifikanz aufweist ($p = 0.275$).¹³⁴ Es wird folgend zuerst geprüft, ob die Mittelwerte der Fremd- und Selbsttäuschung in den drei Versuchsgruppen verschieden sind:

Tab. 36: Die Mittelwerte der Fremd- und Selbsttäuschung, aufgegliedert auf die drei Versuchsgruppen

<i>Versuchsgruppe</i>	Fremdtäuschung: IM			Selbsttäuschung: SDE		
	<i>Mittelwert</i>	<i>Standardabweichung</i>	<i>Häufigkeit</i>	<i>Mittelwert</i>	<i>Standardabweichung</i>	<i>Häufigkeit</i>
Kein Incentive	4.19	1.33	60	4.71	1.00	60
Incentive 5€	4.33	1.39	60	4.53	1.04	60
Incentive 20€	4.51	1.42	60	4.43	1.13	60
Gesamtsumme	4.34	1.38	180	4.56	1.06	180

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab für IM: $F = 0.816$; $p = 0.444$ und für SDE: $F = 1.026$; $p = 0.360$.

Die berechneten Mittelwerte der Fremd- und Selbsttäuschung sind zwar nicht signifikant verschieden, dennoch lassen sich überraschende Tendenzen ablesen: So steigt die Fremdtäuschung bei Vergabe eines Incentives leicht an, wohingegen die Mittelwerte der Selbsttäuschung leicht abnehmen. Wird der Anstieg der Fremdtäuschung als ein Anstieg der sozial erwünschten Antworten interpretiert werden, so entspricht dies dem bereits vorher dargestellten Befund zu den Falschangaben: Es werden bei Vergabe von Incentives tendenziell mehr

¹³⁴ Zur Berechnung der Korrelation wurden Mittelwertindices für die beiden Dimensionen gebildet.

Falschantworten. Die Abnahme der Selbsttäuschung könnte darauf hinweisen, dass den Befragten ihr Antwortverhalten bewusst wird und damit eher als Ergebnis einer Entscheidung interpretiert werden kann. Es muss hierbei jedoch darauf hingewiesen werden, dass die Selbsttäuschung als unbewusster Mechanismus konzipiert ist, welcher Stabilität gegenüber der Gabe eines Incentives aufweisen sollte. Die Abnahme soll daher über die Messungen erklärt werden: Das Item „Ich bin mir oft unsicher in meinem Urteil“ weist bei Kontrolle der drei Versuchsgruppen relativ starke Mittelwertunterschiede auf:¹³⁵

Tab. 37: Die Mittelwerte der Antworten zu dem rekodierten Item „Ich bin mir oft unsicher in meinem Urteil“, aufgegliedert nach den drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Häufigkeit</i>	<i>Standardabweichung</i>
Kein Incentive	4.90	60	1.43
Incentive 5€	4.43	60	1.66
Incentive 20€	4.47	60	1.59
Gesamtsumme	4.60	180	1.57

Quelle: eigene Daten. Die einfaktorielle Varianzanalyse ergab: $F = 1.662$; $p = 0.193$.

Diese Unterschiede sind leicht über der maximal angesetzten Irrtumswahrscheinlichkeit von 10% und daher als nicht signifikant zu betrachten (5€: $p = 0.104$; 20€: $p = 0.131$). Die deskriptive Abnahme könnte jedoch dadurch erklärt werden, dass sich die Befragten aufgrund des Incentive intensiver mit der Befragung auseinandersetzen, deswegen häufiger in Abwägungsprozesse geraten und somit mehr Unsicherheit in der Urteilsfindung erleben. Dieses Erleben führt dazu, dass bei Betrachtung der Selbsttäuschung eine Abnahme festgestellt werden kann.

¹³⁵ Das zugrunde liegende Item wurde für die Analysen so rekodiert, dass hohe Werte für eine Selbsttäuschung sprechen. Die berichteten Mittelwerte des Items müssen daher aufgrund der neuen Kodierung in Richtung „Ich bin mir oft sicher in meinem Urteil“ interpretiert werden.

8.2.3 Zusammenfassung der Indikatoren eines (situational) wahrheitsgemäßen Bearbeitens des Fragebogens

Die Messungen der Pseudo-Opinions werden nun auf Zusammenhänge mit der sozialen Erwünschtheit geprüft. Es wird hierbei erwartet, dass die Falschangaben positiv mit der Fremdtäuschung zusammenhängen, da zum einen die Abgabe einer Antwort als sozial erwünscht wahrgenommen werden kann und zum anderen eigenes Unwissenheit überdeckt wird. Diese vermuteten Zusammenhänge werden nun für die drei Versuchsgruppen untersucht:

Tab. 38: Korrelationen zwischen den Pseudo-Opinions und Dimensionen der sozialen Erwünschtheit¹³⁶

Kein Incentive	IM	0.143 (p = 0.275) n = 60	0.112 (p = 0.395) n = 60	0.014 (p = 0.917) n = 60	0.087 (p = 0.509) n = 60	0.117 (p = 0.372) n = 60
	SDE	0.262 (p = 0.043) n = 60	0.306 (p = 0.018) n = 60	0.158 (p = 0.229) n = 60	0.334 (p = 0.009) n = 60	0.218 (p = 0.094) n = 60
Incentive: 5 Euro	IM	0.131 (p = 0.320) n = 60	0.087 (p = 0.509) n = 60	0.088 (p = 0.504) n = 60	-0.101 (p = 0.444) n = 60	-0.082 (p = 0.534) n = 60
	SDE	0.191 (p = 0.144) n = 60	0.139 (p = 0.291) n = 60	-0.038 (p = 0.772) n = 60	-0.040 (p = 0.763) n = 60	0.085 (p = 0.520) n = 60
Incentive: 20 Euro	IM	0.156 (p = 0.234) n = 60	0.077 (p = 0.557) n = 60	-0.034 (p = 0.794) n = 60	-0.043 (p = 0.746) n = 59	0.208 (p = 0.111) n = 60
	SDE	0.169 (p = 0.198) n = 60	0.104 (p = 0.431) n = 60	0.178 (p = 0.175) n = 60	0.265 (p = 0.043) n = 59	0.307 (p = 0.017) n = 60
		Bewertung LA-42 I	Bewertung LA-42 II	Bewertung Politiker I	Bewertung Politiker II	Bewertung Deckblatt

Quelle: eigene Daten.

¹³⁶ Die Variable zur Kenntnis der Droge LA-42 ist nicht in dieser Tabelle berücksichtigt, da sie aus oben bereits erläuterten Gründen nicht in den multivariaten Analysen herangezogen wird.

Die Korrelationen der Fremdtäuschung und den Pseudo-Opinions weisen in der Versuchsgruppe ohne Incentive die erwarteten Vorzeichen auf, wobei jedoch keine statistische Signifikanz der Ergebnisse vorliegt. In den Versuchsgruppen mit einem Incentive sind hingegen auch gegenläufige Effekte zu beobachten, was den ursprünglich unterstellten Zusammenhang zu widerlegen scheint. Damit fehlt eine statistische Fundierung der Annahme, dass die beobachtete Steigerung der Falschangaben eine bewusste Strategie der Probanden darstellt. Die negativen Zusammenhänge in der Versuchsgruppe mit 5 Euro könnten ein Indiz dafür sein, dass sich die Befragten ehrlicher gegenüber den BIDR-Items äußern, aber aus unbewussteren Gründen häufiger Falschangaben machen. Dies zeigt sich in den deutlichen Zusammenhängen mit der Selbsttäuschung. Dies weist darauf hin, dass sich die Befragten ein Unwissen gegenüber den gestellten Fragen nicht eingestehen wollen und daher zu Falschangaben neigen.

Als Indikator für ein wahrheitsgemäßes Bearbeiten des Fragebogens soll die Anzahl an Pseudo-Opinions, bzw. Falschantworten herangezogen werden. Die Begründung liegt darin, dass die Falschangaben als direkte Messung der Antwortqualität interpretiert werden können, wohingegen die Selbsteinstufungen eher als indirekte Messung angesehen werden und darüber hinaus auch instabiler (z.B. aufgrund von Skaleneffekten) erscheinen.

8.3 Analyse der Indikatoren für ein vollständiges Bearbeiten des Fragebogens

8.3.1 Das Überspringen von Fragebogenfragen

Im Rahmen dieser Befragung wurden von den geschlossenen Fragen bis zu 37 übersprungen und damit nicht beantwortet.¹³⁷ Unter Berücksichtigung der drei Versuchsgruppen kann eine

¹³⁷ Die Fragen zu den Pseudo-Opinions wurden hierbei nicht berücksichtigt, da diese so konzipiert sind, dass die Befragten bei sorgfältiger Bearbeitung einen Übersprung vornehmen müssen.

Abnahme der Übersprünge erkannt werden: Befragte ohne Incentive übersprangen im Durchschnitt 4.9 Fragen, bei einem Incentive in Höhe von 5 Euro durchschnittlich 4.7 Fragen und bei Vergabe von 20 Euro durchschnittlich 3.1 Fragen. Die Unterschiede sind jedoch nicht signifikant.¹³⁸ Es soll darüber hinaus darauf aufmerksam gemacht werden, dass die Reduktion der Übersprünge nicht zwangsläufig eine Verbesserung der Antwortqualität bedeutet, da die Befragten mehr Falschangaben im Fragebogen tätigen können oder aufwandsreduzierende Antwortmuster nutzen. Dies findet sich auch in den Kommentaren der Befragten wieder:

„Kenne die Droge LA-42 nicht, klingt aber sehr chemisch und daher ungesund.“ (Ifdn = 498)

Anm. des Autors:
Die Person hat die Droge LA-42 trotz Unkenntnis zweimal bewertet.

*„Habe immer das mittlere angeklickt, wenn ich absolut keine Ahnung hatte.“
(Ifdn = 565)*

¹³⁸ Die einfaktorielle Varianzanalyse ergab: $F = 1.474$; $p = 0.232$.

„gesundheitslich = körperliche Gesundheit
Wenn Auswahl in Mitte>>> keine Aussage machbar, da wenig Wissen“
(Ifdn = 928)

Darüber hinaus wurde im Kapitel zur Operationalisierung (6.2.3) aufgezeigt, dass nicht jeder Übersprung auch das Minimieren von Aufwand zum Ziel haben muss. Damit die Übersprünge erläutert werden können, wurde während der Befragung eine durchgängige Kommentarfunktion angeboten. Damit konnten die Befragten einen Übersprung begründen und für den Forscher ist damit eher ersichtlich, ob eine Strategie zur Minimierung des Aufwands verfolgt wird. Die Kommentarfunktion wurde durchschnittlich 2.1 Mal zur Begründung von Übersprüngen genutzt. Unter Kontrolle der drei Versuchsgruppen zeigt sich eine leichte Abnahme der begründeten Übersprünge: in der Gruppe ohne Incentive wurde die Kommentarfunktion zur Begründung von Übersprüngen durchschnittlich 2.3 Mal genutzt, in der ersten Experimentalgruppe 2.2 Mal, bzw. 1.8 Mal in der zweiten Experimentalgruppe. Die Unterschiede sind nicht signifikant, aber weisen auf eine tendenzielle Abnahme der Begründungen mit steigendem Incentive hin.¹³⁹ Wird hierbei beachtet, dass die Befragten der dritten Versuchsgruppe insgesamt weniger Übersprünge durchführen, so löst sich jedoch der vermeintlich negative Effekt auf, da diese folglich auch nur eine geringe Anzahl an Übersprüngen begründen können.

¹³⁹ Die einfaktorielle Varianzanalyse ergab: $F = 0.293$; $p = 0.747$.

Folgend soll nun die Anzahl der nicht kommentierten Übersprünge betrachtet werden, da bei diesen der Aufwand der Begründung gescheut wurde und daher insgesamt von einer eher schlechten Antwortqualität ausgegangen werden kann. Hierfür wird von der Variable mit der Gesamtzahl an Übersprünge die Variable mit den begründeten Übersprünge abgezogen:

Tab. 39: Die Häufigkeit an Übersprünge ohne Begründung über einen Kommentar

Übersprünge	Häufigkeit	Gültige Prozent	Kumulative Prozente
0	67	37.8	37.8
1	51	28.3	66.1
2	23	12.8	78.9
3	16	8.9	87.8
4	4	2.2	90.0
5	4	2.2	92.2
6	4	2.2	94.4
9	1	.6	95.0
11	1	.6	95.6
12	1	.6	96.1
15	2	1.1	97.2
19	2	1.1	98.3
20	1	.6	98.9
24	1	.6	99.4
37	1	.6	100.0
Gesamtsumme	179	100.0	

Quelle: eigene Daten.

67 Befragte haben, wenn überhaupt, nur mit einer Begründung in den Kommentaren Fragen übersprungen. 50% der Befragten haben eine Frage und 75% zwei Fragen ohne Begründung übersprungen. Dies weist insgesamt auf eine recht hohe Antwortqualität hin, auch wenn neun Probanden mehr als zehn Fragen ohne Begründung übersprungen haben. Unter Berücksichtigung der drei Versuchsgruppen ergeben sich folgende Mittelwerte:

Tab. 40: Mittelwerte zu den nicht begründeten Übersprüngen, aufgliedert nach den drei Versuchsgruppen

<i>Versuchsgruppe</i>	<i>Mittelwert</i>	<i>Standard- abweichung</i>	<i>Häufigkeit</i>
Kein Incentive	2.55	4.77	60
Incentive 5€	2.50	5.87	60
Incentive 20€	1.41	1.86	59
Gesamtsumme	2.14	4.50	180

Quelle: eigene Daten. Eine Person hat die Befragung abgebrochen und kann daher nicht berücksichtigt werden. Die einfaktorielle Varianzanalyse ergab: $F = 1.219$; $p = 0.298$.

Es ist zu erkennen, dass die Befragten, welche ein Incentive in Höhe von 20€ erhielten, deutlich weniger Fragen übersprangen wie die Befragten aus den beiden anderen Versuchsgruppen. Die Abnahme an nicht-begründeten Übersprüngen könnte durch reziprokes Verhalten erklärt werden. Dem widerspricht allerdings die Beobachtung, dass die Befragten mit einem Incentive in Höhe von 5 Euro keine bzw. eine sehr geringe Abnahme an Übersprüngen aufweisen. Demgemäß müssten Schwellenwerte bezüglich der Höhe des Incentive angenommen werden, welche erst durch Überschreiten Reziprozität aktiviert wird, wobei 5 Euro als zu niedrig wahrgenommen werden. Alternativ kann vermutet werden, dass Befragte die Befragungssituation so einschätzen, dass die Forscher eine Antwort gegenüber einem Überspringen bevorzugen. Dies könnte dazu führen, dass die Anzahl an Übersprüngen abnimmt, allerdings die Anzahl an falschen bzw. willkürlichen Angaben zunimmt. Diese Vermutung soll in den späteren multivariaten Analysen unter Kontrolle der Falschangaben überprüft werden.

8.3.2 Das Ausweichverhalten bei Filterfragen

Im Folgenden wird nun geprüft, inwiefern Verneinungen von Filterfragen als Strategie zur Minimierung des Bearbeitungsaufwands genutzt werden. Insgesamt geben 166 Befragte an, jede

der vorgestellten Drogen zu kennen.¹⁴⁰ Dies bedeutet, dass 14 Befragte mindestens einmal das Wissen über eine der acht genannten Drogen verneinen. Die Verteilungen für die drei Versuchsgruppen werden in der folgenden Tabelle zusammengefasst:

Tab. 41: Mehrfeldertabelle zur Prüfung des Zusammenhangs zwischen dem Filterverhalten und der Incentivierung

		Kein Incentive	Incentive: 5 Euro	Incentive: 20 Euro	Gesamt
Alle Filterfragen positiv beantwortet	Ja (0)	57 (95.0%)	55 (91.7%)	54 (90.0%)	166 (92.2%)
	Nein (1)	3 (5.0%)	5 (8.3%)	6 (10.0%)	14 (7.8%)
	Gesamt	60 (100%)	60 (100%)	60 (100%)	180 (100%)

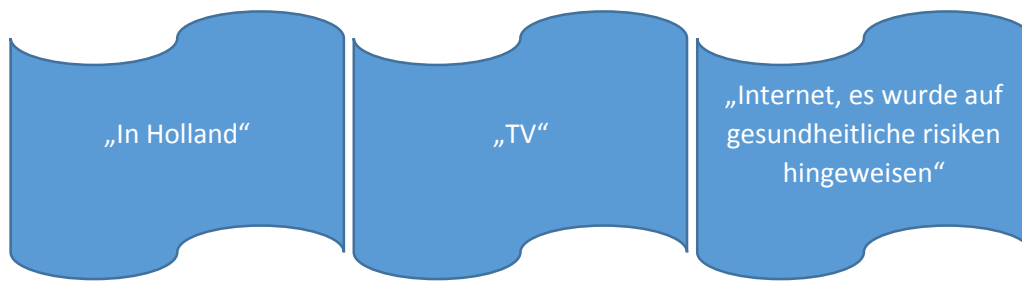
Quelle: eigene Daten.

Es kann hierbei keine signifikante Beziehung zwischen der Vergabe eines Incentives und der Verneinung von Filterfragen festgestellt werden ($p = 0.297$). Es muss an dieser Stelle darauf hingewiesen werden, dass ein Verneinen auf tatsächlicher Unkenntnis beruhen kann und folglich keine Strategie zur Verminderung von Bearbeitungsaufwand sein muss. Es werden daher ergänzend die Kommentare zu den vorher gestellten Drogen-Bewertungsfragen hinzugezogen, um zu prüfen, ob die Befragten in den Vorfragen ein Unwissen über die Drogen äußerten. Unter Berücksichtigung der Kommentare kann nun festgestellt werden, dass von den vierzehn Befragten nur drei vorher anmerkten nicht alle Drogen zu kennen. Demzufolge haben 11 Befragte in den Vorfragen falsch geantwortet (und dann bei den Filterfragen korrekt die Unkenntnis angegeben) oder bei den Filterfragen bewusst falsche Angaben gemacht. Es ist damit unklar, ob der Aufwand zur Bearbeitung der Filterfragen in dieser Befragung ausreichend hoch

¹⁴⁰ Die Droge LA-42 wird in dieser Analyse nicht berücksichtigt, da sie nicht existiert und folglich verneint werden muss.

genug angesetzt war, um ein Vermeidungsverhalten zu provozieren. Dies ist besonders kritisch zu hinterfragen, da 16 Befragte die Filterfragen zur nicht existenten Droge LA-42 bejahten und sich damit bewusst für einen erhöhten Aufwand entschieden. Ob dies aus Gründen der Konsistenz oder des Verwechselns mit einer wirklich existierenden Droge geschieht kann leider nicht geklärt werden. Auf die offene Folgefrage des Filters, wo und wie die Befragten zum ersten Mal etwas von der Droge mitbekommen haben, werden vorwiegend nur sehr kurze und vage Angaben gemacht, wie zum Beispiel:

Abb. 20: Drei ausgewählte Antworten zu der offenen Frage: „Bitte beschreiben Sie, wo und wie Sie zum ersten Mal von der Droge LA-42 etwas mitbekommen haben.“



Quelle: eigene Darstellung.

Aufgrund der oben dargestellten Probleme in der Interpretation der Ergebnisse wird das Ausweichverhalten bei Filterfragen nicht als Indikator für ein vollständiges Bearbeiten herangezogen.

8.3.3 Zusammenfassung für die Indikatoren eines vollständigen Bearbeitens des Fragebogens

Obwohl die Nutzbarkeit des Filterverhaltens bereits verneint wurde, soll geprüft werden, ob ein konsistenter Zusammenhang mit dem nicht-begründeten Überspringen von Fragen besteht. Hierfür wurde der Zusammenhang zwischen den beiden Indikatoren berechnet: die Korrelation weist auf einen gegenläufigen Zusammenhang hin ($r = -0.041$), d.h. das Überspringe mit einem aufwandsfördernden Filterverhalten zusammenhängen. Dieser Zusammenhang ist

allerdings statistisch nicht signifikant ($p = 0.588$) und auch unter Kontrolle der drei Versuchsgruppen kann kein konsistenter Zusammenhang aufgedeckt werden. Aufgrund der in Kapitel 6.2.3 erläuterten Probleme mit der Interpretierbarkeit des Filterverhaltens soll das nicht-begründete Überspringen von Fragen als Indikator für ein vollständiges Bearbeiten des Fragebogens herangezogen werden.

8.4 Indikatoren für ein anweisungsbefolgendes Bearbeiten des Fragebogens

Zur Prüfung des Befolgens von Anweisungen wurden zwei Messungen vorgenommen. Die erste Messung umfasste das Befolgen der Anweisungen zur Berechnung eines Glückszahlkoeffizienten unter Verwendung des Windows-Taschenrechners. Das Programm des Taschenrechners wurde hierfür bereits vor Beginn der Befragung gestartet und im Hintergrund gehalten. Die Befragten konnten dann über die Task-Leiste das Programm einfach anwählen und nutzen. Mithilfe von Para-Daten wird nun geprüft, inwiefern die Probanden der Anweisung gefolgt sind: Nur 6 Befragte haben die Fragebogenseite nicht verlassen und haben demgemäß nicht den Windows-Taschenrechner genutzt.¹⁴¹

Die zweite Messung zur Erfassung eines anweisungsbefolgenden Bearbeitens erfolgt erneut über ein Verlassen des Fragebogens. Diesmal wurde die Anweisung gegeben Google Maps® zu nutzen, um die exakte Entfernung (zu Fuß, kürzester Weg) vom Erhebungslabor bis zum Wohnsitz der Befragten zu ermitteln. Diesmal haben 14 Befragte die Anweisung nicht befolgt und folglich die Entfernung geschätzt oder ausgedacht.¹⁴²

Zwischen den beiden Messungen liegt ein mittelstarker Zusammenhang vor ($r = 0.408$), welcher mit einem empirischen Signifikanzniveau von $p = 0.000$ höchstsignifikant ist. Das positive

¹⁴¹ Aufgrund der geringen Fallzahl wird die Aufgliederung in die drei Versuchsgruppen nicht vorgenommen.

¹⁴² Es kann an dieser Stelle argumentiert werden, dass die Befragten aufgrund von Fahrrad-Tachometern oder Schrittzählern beim Joggen die Entfernung bekannt ist und deswegen auf die Nutzung von Google Maps® verzichtet wurde.

Vorzeichen der Korrelation ist gemäß der Erwartung: wer der Anweisung zur Mathematikaufgabe folgt, der folgt auch eher der Anweisung zur Messung der Entfernung.

Aufgrund des Zusammenhangs werden die beiden Variablen zu einem Summenindex zusammengefasst. 162 Befragte haben demnach die Anweisungen immer befolgt, 12 Befragte haben die Anweisungen nur einmal und 4 Personen haben sie nie befolgt. Dieser Index wird als Indikator für ein anweisungsbefolgendes Bearbeiten des Fragebogens in den multivariaten Datenanalysen weiter genutzt.

8.5 Der Zusammenhang zwischen den ausgewählten Indikatoren der Antwortqualität

Es wird nun untersucht, inwiefern die vier ausgewählten Indikatoren (= Status Quo-Effekt, Falschangaben, nicht begründete Übersprünge und ein anweisungsbefolgendes Bearbeiten des Fragebogens) zusammenwirken und damit die Definition der Antwortqualität empirisch stützen. Hierfür wurden explorative Faktoranalysen mithilfe von SPSS berechnet. Die Ergebnisse sind in der folgenden Tabelle eingetragen:

Tab. 42: Faktorladungen der explorativen Faktorenanalyse für die ausgewählten Indikatoren der Antwortqualität

	<i>Mustermatrix (Oblimin-Rotation)</i>		<i>Rotierte Komponentenmatrix (Varimax-Rotation)</i>	
	Faktor I	Faktor II	Faktor I	Faktor II
<i>Status Quo</i>	0.669	-0.056	0.669	0.067
<i>Falschangaben</i>	-0.774	-0.081	-0.773	0.068
<i>Überspringen</i>	0.347	-0.737	0.355	0.743
<i>Anweisungen</i>	0.295	0.810	0.286	-0.804

Quelle: eigene Daten. Die Fallzahl beträgt n = 176. Die Faktoren basieren auf einem Eigenwertkriterium von ≥ 1 . Es wurde für die Berechnungen in der linken Tabellenhälfte die Varimax-Rotation und für die rechte Hälfte die Oblimin-Rotation durchgeführt.

Auf Basis der explorativen Faktoranalyse lassen sich zwei Faktoren unterscheiden. Der erste Faktor umfasst die Komponenten für ein durchdachtes und wahrheitsgemäßes Bearbeiten des

Fragebogens. Dieses Zusammenspiel ist insofern plausibel, als dass bei diesen beiden Komponenten inhaltliche Aspekte des Beantwortungsprozesses angesprochen sind. Der Faktor behandelt damit eine inhaltliche, materielle Ebene. Der zweite Faktor setzt sich aus den Komponenten für ein vollständiges und anweisungsbefolgendes Bearbeiten zusammen. Auch dies erscheint plausibel, da ein vollständiges Bearbeiten auch als (unausgesprochene) Anweisung verstanden werden kann. Dieser Faktor bezieht sich damit folglich auf eine eher formelle Ebene, also auf die Rahmenbedingungen in einer Befragung. Diese Zwei-Faktoren-Lösung entspricht den theoretischen Vorannahmen aus Kapitel 2.4 und stützt damit die entwickelte Definition der Antwortqualität. Die aufgedeckten Zusammenhänge der einzelnen Komponenten werden in den folgenden multivariaten Analysen berücksichtigt und in den Strukturgleichungsmodellen entsprechend modelliert.

9. Hypothesenprüfung

Im Rahmen der multivariaten Analysen werden mithilfe von Mplus simultane Gruppenvergleiche von Strukturgleichungsmodellen zur Überprüfung der Hypothesen durchgeführt. Hierbei werden die vorher ausgewählten Indikatoren für die Antwortqualität¹⁴³ in ein Abhängigkeitsverhältnis mit den Indikatoren der intrinsischen Motivation, der Reziprozitätshypothese und der extrinsischen Motivation gesetzt.¹⁴⁴

¹⁴³ Die Indikatoren können aufgrund der internen Konsistenz (siehe Kapitel 7.1) nicht zu einem Qualitätsindex zusammengefasst werden und gehen daher als Einzelmessungen in die Analysen ein.

¹⁴⁴ In den folgenden Berechnungen wurde ein Algorithmus zur Schätzung von MAR-Ausfällen genutzt. Darüber hinaus wird aufgrund der Dichotomie der abhängigen Variable „Status Quo“ diese als kategorial definiert. Dem folgend wird der WLSM-Schätzer für die Berechnungen genutzt und die Theta-Parametrisierung angewendet. Die angeführten Namen für die Faktoren wurden aufgrund der einfacheren Interpretation in den Graphiken angepasst.

9.1 Die Prüfung der Hypothesen zur Wirkung der intrinsischen Motivation und der Reziprozität auf die Antwortqualität

Das Strukturmodell zur Prüfung der Hypothesen 1a, 1b, 2a, 2b und 2c wird im Sinne der im vierten Kapitel formulierten theoretischen Vorüberlegungen konstruiert. Die Hypothesen werden noch einmal wiederholt und die erwarteten Zusammenhänge zur Verdeutlichung auch noch graphisch dargestellt:

H1a: Der Erhalt eines un konditionalen Incentives hat keinen Einfluss auf die Höhe der intrinsischen Motivation.

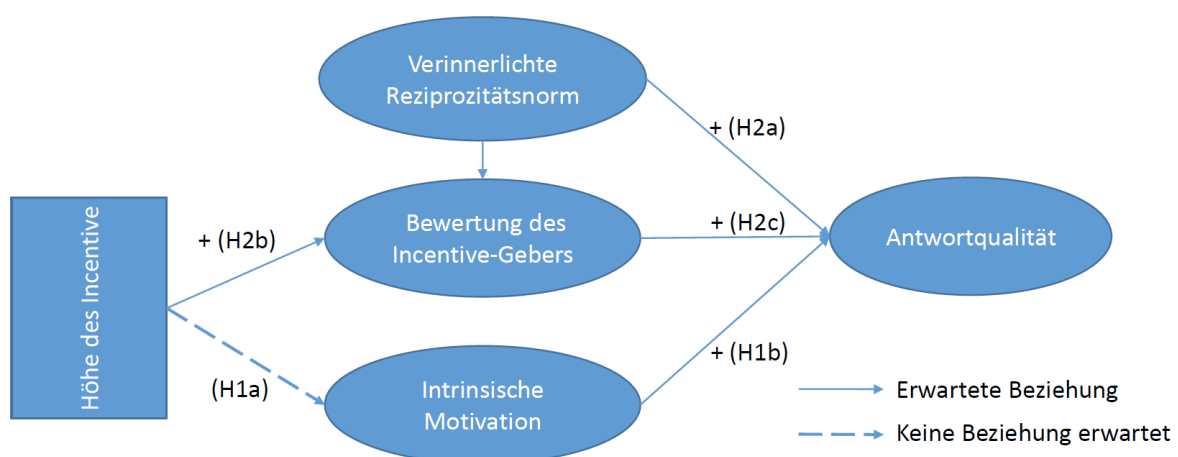
H1b: Je höher die intrinsische Motivation, desto höher die Antwortqualität.

H2a: Je höher das Incentive, desto stärker wirkt die verinnerlichte Reziprozitätsnorm positiv auf die Antwortqualität.

H2b: Je höher das Incentive, desto positiver wird der Belohnungsgeber wahrgenommen.

H2c: Je positiver der Belohnungsgeber wahrgenommen wird, desto höher die Antwortqualität.

Abb. 21: Die Hypothesen bezüglich der Wirkung der intrinsischen Motivation und der Reziprozität auf die Antwortqualität¹⁴⁵

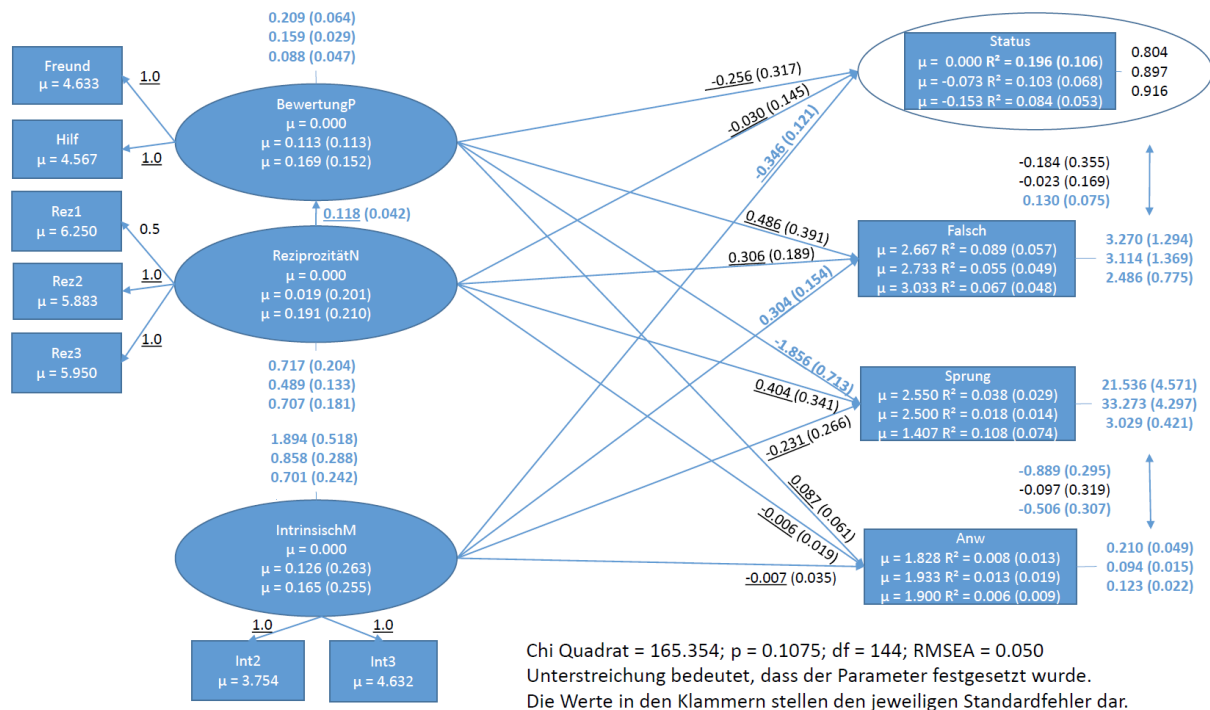


Quelle: eigene Darstellung.

¹⁴⁵ Die Antwortqualität wird in der Darstellung gemäß den theoretischen Überlegungen als Gesamtkonstrukt betrachtet und nicht in die vier Facetten aufgegliedert.

Im Rahmen der Modellierung der Strukturgleichungsmodelle wird eine gerichtete Beziehung zwischen dem Faktor der verinnerlichten Reziprozitätsnorm auf den Faktor zur Wahrnehmung der Person zugelassen, wohingegen deren Beziehungen zur intrinsischen Motivation stets auf Null festgesetzt werden. Gemäß der empirischen Befunde werden ebenfalls Kovarianzen zwischen dem Status Quo und den Falschangaben, sowie dem nicht begründeten Überspringen von Fragebogenfragen und dem Befolgen von Anweisungen im Modell zugelassen. Da die Hypothesen 1a, 1b, 2b und 2c einen linearen (additiven) Zusammenhang postulieren, wird zuerst ein Modell gerechnet, bei welchem die Effekte über alle drei Versuchsgruppen konstant gehalten werden:

Abb. 22: Graphische Darstellung des Strukturgleichungsmodells bei restriktiver Modellierung der Zusammenhänge der Faktoren auf die Indikatoren der Antwortqualität, zur Prüfung der Hypothesen 1a – 2c¹⁴⁶



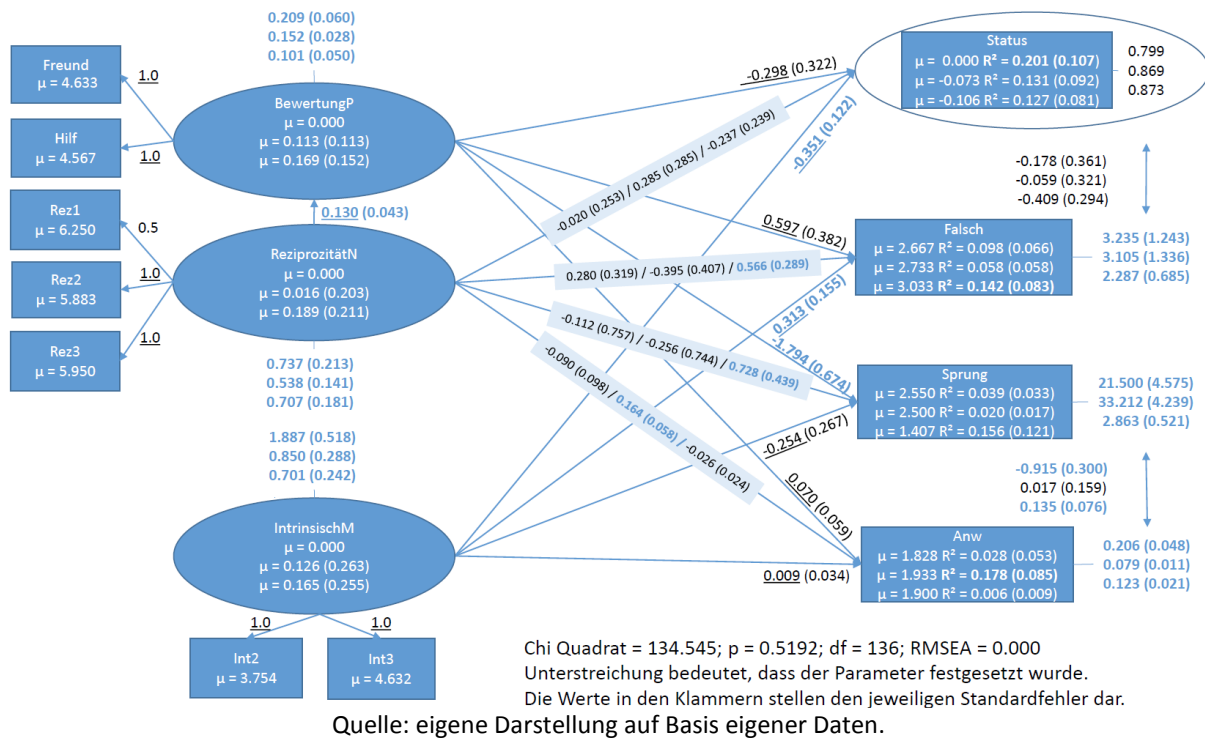
Quelle: eigene Darstellung auf Basis eigener Daten.

¹⁴⁶ Die Messung des Status Quo ist mit einem Kreis umrahmt, da Mplus diese als latente Variable interpretiert, welche dichotom gemessen. Daher wird der Mittelwert in der ersten Gruppe auf 0 festgesetzt. Fett gedruckte Werte weisen auf signifikante Beziehungen hin, wobei hier aufgrund der geringen Fallzahl eine maximale Irrtumswahrscheinlichkeit von 10% akzeptiert wird.

Bei Betrachtung der Mittelwerte der Facetten der Antwortqualität kann folgendes festgestellt werden: der Status Quo-Effekt weist in den Versuchsgruppen mit einem Incentive ein negatives Vorzeichen auf, was auf eine verstärkte Wahl des Status Quo, also einer Verschlechterung der Antwortqualität hindeutet. Die Anzahl an Falschangaben steigt bei Vergabe eines Incentives, was ebenfalls als negativer Effekt von Incentives auf die Antwortqualität interpretiert werden kann. Dem Gegenüber sinkt die Anzahl an unbegründeten Übersprüngen bei Vergabe eines Incentives und deutet damit auf eine Verbesserung der Antwortqualität hin. Zudem kann ein Anstieg des anweisungsbefolgenden Bearbeitens beobachtet werden. Diese Veränderungen sollen nun folgend über die gebildeten Faktoren erklärt werden. Bevor jedoch Zusammenhänge interpretiert werden, soll zuerst die Güte der Modelle betrachtet werden: der Chi²-Wert erscheint mit 165.354 recht hoch, jedoch liegt das empirische Signifikanzniveau – aufgrund der Freiheitsgrade (df = 144) – bei $p = 0.1075$, was auf eine gerade noch hinreichende Passung des theoretischen Modells schließen lässt.¹⁴⁷ Da Hypothese 2a eine Interaktion impliziert wird daher im zweiten Schritt die Beziehungen der Reziprozitätsnorm auf die Indikatoren der Antwortqualität über die Versuchsgruppen freigegeben und können damit in der Höhe variieren:

¹⁴⁷ Es ist hervorzuheben, dass eine geringe Anzahl an Fällen oft zu einer Überschätzung des Modellfits führen kann. Daher erscheint das empirische Signifikanzniveau unzureichend.

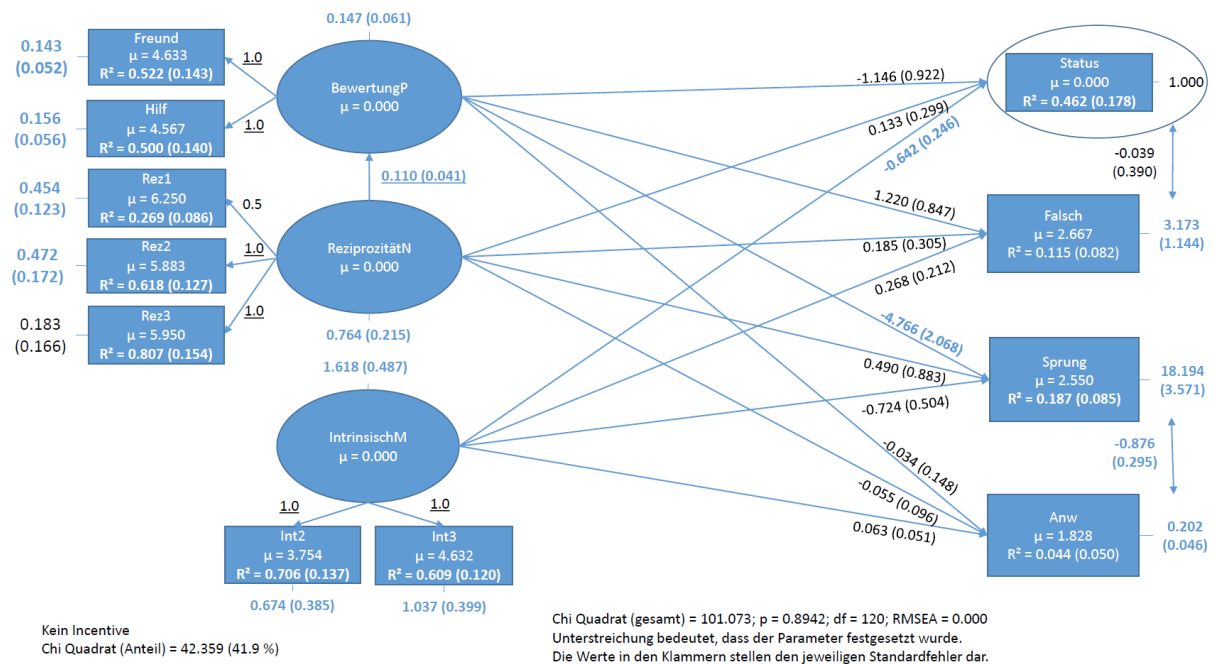
Abb. 23: Graphische Darstellung des Strukturgleichungsmodells bei Freigabe des restriktiven Zusammenhangs der verinnerlichteten Reziprozitätsnorm, zur Prüfung der Hypothesen 1a – 2c



Das theoretische Modell scheint nun die empirischen Zusammenhänge besser zu erfassen ($p = 0.5192$) und auch der RMSEA (= 0.000) und weist auf eine bessere Passung der Daten hin. Zur statistischen Absicherung wurde ein Santorra-Bentler χ^2 -Differenztest berechnet, welcher die wahrgenommene Verbesserung der Modellpassung stützt ($\chi^2 = 17.74$; $df = 8$; $p = 0.008$). Die signifikante Verbesserung des Fits gibt Anlass zur Prüfung, ob auch die Zusammenhänge der beiden anderen Faktoren nicht-additiv sind. Aus diesem Grund wurde erneut ein Satorra-Bentler χ^2 -Differenztest berechnet, diesmal im Vergleich zu einem liberalen Modell, bei welchem keine Gleichheitsrestriktionen für die Zusammenhänge der Faktoren auf die Indikatoren der Antwortqualität gesetzt werden.

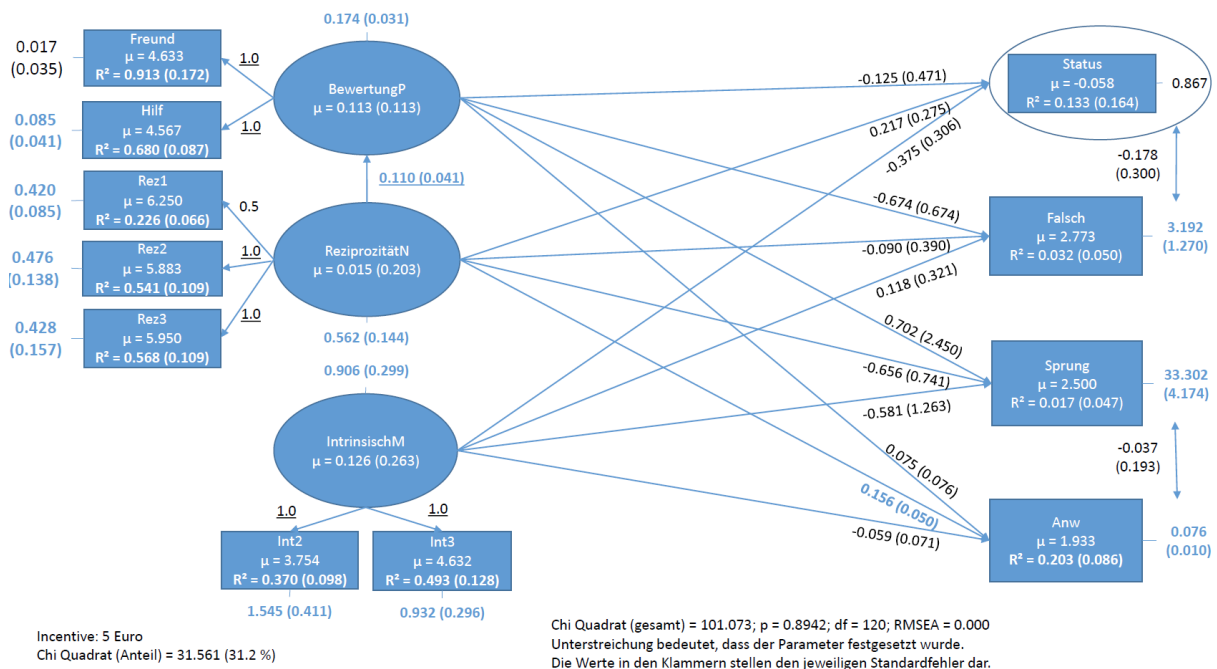
Die Ergebnisse der Fit-Werte sind in den folgenden Strukturgraphiken eingefügt.

Abb. 24: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe ohne Incentive, zur Prüfung der Hypothesen 1a – 2c¹⁴⁸



Quelle: eigene Darstellung auf Basis eigener Daten.

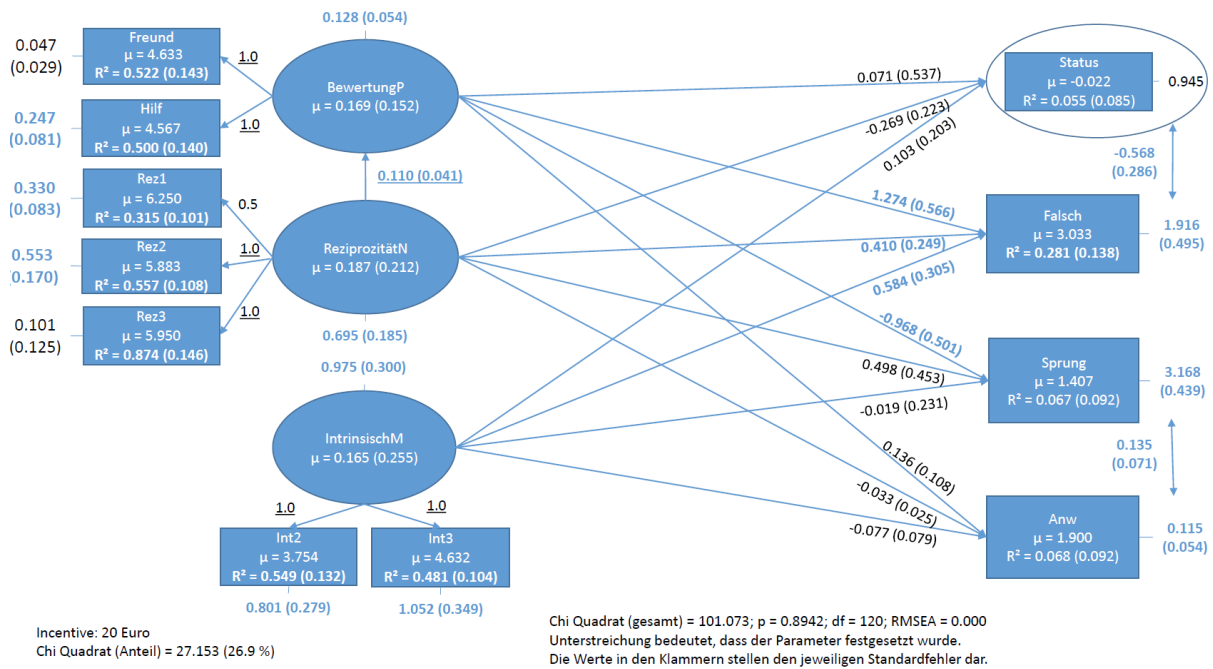
Abb. 25: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe mit einem Incentive in Höhe von 5€, zur Prüfung der Hypothesen 1a – 2c



Quelle: eigene Darstellung auf Basis eigener Daten.

¹⁴⁸ Die Messung des Status Quo ist mit einem Kreis umrahmt, da Mplus diese als latente Variable interpretiert, welche dichotom gemessen. Daher wird der Mittelwert in der ersten Gruppe auf 0 festgesetzt. Fett gedruckte Werte weisen auf signifikante Beziehungen hin, wobei hier aufgrund der geringen Fallzahl eine maximale Irrtumswahrscheinlichkeit von 10% akzeptiert wird. Die Werte der standardisierten Koeffizienten sind im Anhang abgebildet, S. 207 – 208.

Abb. 26: Graphische Darstellung des Strukturgleichungsmodells für die Versuchsgruppe mit einem Incentive in Höhe von 20€, zur Prüfung der Hypothesen 1a – 2c



Quelle: eigene Darstellung auf Basis eigener Daten.

Im liberalen Modell zeigt sich im Vergleich zu den vorhergehenden Modelle erneut eine Verbesserung der Passung ($\chi^2 = 101.073$; $df = 120$). Dieses Ergebnis wird auch durch das Resultat des χ^2 -Differenztests gestützt ($\chi^2 = 24.66$, $df = 16$, $p = 0.019$). Aufgrund dieser signifikanten Modellverbesserung werden daher im Folgenden die unstandardisierten Regressionskoeffizienten des liberalen Modells für die Interpretationen herangezogen.

Zur Prüfung der Hypothese 1a wird nun untersucht, ob sich der Mittelwert der intrinsischen Motivation trotz Incentivierung stabil verhält und folglich nicht ändert. Bei Betrachtung der Mittelwerte in den drei Versuchsgruppen kann ein leichter Anstieg der intrinsischen Motivation bei Gabe von 5 Euro bzw. 20 Euro festgestellt werden. Dies könnte darauf hinweisen, dass ein Incentive entgegen der ursprünglichen Erwartung die intrinsische Motivation verbessert

und die Befragten damit positiv beeinflusst.¹⁴⁹ Es ist jedoch anzumerken, dass die gemessenen Anstiege nicht signifikant sind (5€: $p = 0.631$; 20€: $p = 0.516$) und damit eine statistische Absicherung zur Aussage einer Steigerung der intrinsischen Motivation fehlt. Aufgrund dessen muss die Hypothese 1a beibehalten werden: es gibt keine Veränderung der intrinsischen Motivation bei Gabe eines Incentives.

Zur Prüfung der Hypothese 2b werden nun die Mittelwerte des Faktors zur Bewertung des Versuchsleiters betrachtet. Es wird hierbei vermutet, dass die Gabe eines Incentive die Bewertung positiv beeinflusst. Tatsächlich lässt sich eine Steigerung in den Mittelwerten beobachten, wobei jedoch auch hier darauf hingewiesen werden muss, dass die Anstiege nicht signifikant sind (5€: $p = 0.314$; 20€: $p = 0.267$). Somit wird die Hypothese 2b zwar deskriptiv gestützt, kann aber unter Beachtung der statistischen Signifikanz nicht akzeptiert werden.

Zur Prüfung der Zusammenhangshypothesen 1b, 2a und 2c werden nun die unstandardisierten Regressionsgewichte betrachtet. Zur besseren Interpretation werden diese für alle drei Versuchsgruppen in einer einzigen Koeffiziententabelle zusammengefasst, wobei rote Werte auf eine Verschlechterung der Antwortqualität und grüne Werte auf eine Verbesserung der Antwortqualität hinweisen:

¹⁴⁹ Alternativ könnte der (deskriptive) Anstieg der intrinsischen Motivation auch aus den Daten des Strukturgleichungsmodell erklärt werden: die Varianz des Faktors für die intrinsische Motivation sinkt bei Vergabe eines Incentives. Dieser Effekt ist zwar nicht signifikant (5€: $p = 0.2126$; 20€: $p = 0.2604$), könnte aber so interpretiert werden, dass die Befragten entgegen der ursprünglichen Annahmen doch ein akquieszentes Verhalten aufweisen und daher den gestellten Aussagen einheitlicher zustimmen. Der Anstieg könnte damit als Anzeichen für eine Verschlechterung der Antwortqualität interpretiert werden.

Tab. 43: Unstandardisierten Regressionsgewichte des Strukturgleichungsmodells, aufgegliedert nach den drei Versuchsgruppen (Hypothesen 1a – 2c)

	Bewertung d. Person			Reziprozitätsnorm			Intrinsische Motivation		
	0 Euro	5 Euro	20 Euro	0 Euro	5 Euro	20 Euro	0 Euro	5 Euro	20 Euro
Status Quo	-1.146 (0.922)	-0.125 (0.471)	0.071 (0.537)	0.133 (0.299)	0.217 (0.275)	-0.269 (0.223)	-0.642 (0.246)	-0.375 (0.306)	0.103 (0.203)
Falschangaben	1.220 (0.847)	-0.674 (0.674)	1.274 (0.566)	0.185 (0.305)	-0.090 (0.390)	0.410 (0.249)	0.268 (0.212)	0.118 (0.321)	0.584 (0.305)
Überspringen	-4.766 (2.068)	0.702 (2.450)	-0.968 (0.501)	0.490 (0.883)	-0.656 (0.741)	0.498 (0.453)	-0.742 (0.504)	-0.581 (1.263)	-0.019 (0.231)
Anweisungen	-0.034 (0.148)	0.075 (0.076)	0.136 (0.108)	-0.055 (0.096)	0.156 (0.050)	-0.033 (0.025)	0.063 (0.051)	-0.059 (0.071)	-0.077 (0.079)

Quelle: eigene Daten. Fett gedruckte Werte weisen eine empirische Signifikanz von $p < 0.10$ auf. Die Werte in den Klammern entsprechen den Standardfehlern.

a) Die Interpretation der Ergebnisse zum Status Quo-Effekt

Zu Beginn werden Wald-Tests durchgeführt, um zu prüfen, ob zwischen den unstandardisierten Regressionsgewichten der drei Versuchsgruppen für die jeweiligen Faktoren signifikante Unterschiede bestehen. Für die Faktoren der Personenbewertung und der Reziprozitätsnorm können keine signifikanten Unterschiede aufgedeckt werden. Dies ist insofern interessant, da für die beiden Faktoren ein positiver Effekt für die Versuchsgruppen mit einem Incentive erwartet wurde. Beim Faktor der intrinsischen Motivation kann ein signifikanter Unterschied zwischen der Versuchsgruppe ohne Incentive und der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro festgestellt werden ($p = 0.0194$). Hierbei wird der negative (signifikante) Effekt in der Versuchsgruppe ohne Incentive (-0.642) aufgehoben, so dass dies als positive Wirkung des Incentives interpretiert werden kann. Es muss jedoch darauf hingewiesen werden, dass der Wert des unstandardisierten Regressionsgewichts in der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro nicht signifikant von Null verschieden ist und damit auf eine statistische Unabhängigkeit der intrinsischen Motivation auf den Status Quo-Effekt hinweist.

Ein ähnlicher Effekt lässt sich bei der Bewertung des Versuchsleiters feststellen, da auch hier der ursprünglich negative Effekt mit steigendem Incentive aufgelöst wird. Lediglich bei der verinnerlichten Reziprozitätsnorm kann eine Verschlechterung der Antwortqualität bei Vergabe eines Incentives in Höhe von 20 Euro wahrgenommen werden. Diese Ergebnisse sind jedoch vorsichtig zu interpretieren, da wie bereits oben angeführt keine signifikanten Unterschiede der unstandardisierten Regressionsgewichte festgestellt werden konnten.

b) Interpretation der Ergebnisse zu den Pseudo-Opinions (Falschangaben)

Auch hier wird zuerst geprüft, ob sich die unstandardisierten Regressionsgewichte über die drei Versuchsgruppen hinweg für die jeweiligen Faktoren signifikant unterscheiden. Die einzigen signifikanten Unterschiede werden bei der Bewertung des Versuchsleiters aufgedeckt werden. Die Versuchsgruppe mit einem Incentive in Höhe von 5 Euro unterscheidet sich hierbei von den beiden anderen Versuchsgruppen (0€: $p = 0.0801$; 20€: $p = 0.0258$). Dies deutet darauf hin, dass ein Incentive von 5 Euro die Antwortqualität verbessert, da weniger Falschangaben von den Befragten gemacht werden. Ein ähnliches Muster zeigt sich auch bei der verinnerlichten Reziprozitätsnorm, da auch hier bei Vergabe von 5 Euro eine Reduktion der Falschangaben festgestellt werden kann. Nur im Rahmen der intrinsischen Motivation kann diese positive Wirkung nicht festgestellt werden.

Es ist darüber hinaus hervorzuheben, dass bei einem Incentive in Höhe von 20 Euro die Anzahl an Falschangaben signifikant erhöht wird (Bewertung des Versuchsleiters: $p = 0.024$; verinnerlichte Reziprozitätsnorm: $p = 0.100$; intrinsische Motivation: $p = 0.056$). Dies spricht insgesamt für eine Verschlechterung der Antwortqualität bei Gabe eines Incentives in Höhe von 20 Euro.

c) Die Interpretation der Ergebnisse zu den nicht begründeten Übersprüngen

Die Bewertung des Versuchsleiters weist erneut als einziger Faktor signifikante Unterschiede zwischen den unstandardisierten Regressionsgewichten der drei Versuchsgruppen auf. Hierbei überspringen die Befragten in der Versuchsgruppe ohne Incentive deutlich seltener Fragen als die Befragten der beiden anderen Versuchsgruppen. Die Versuchsgruppe mit einem Incentive in Höhe von 5 Euro weist dabei einen negativen Effekt auf die Antwortqualität auf, da mehr nicht begründete Übersprünge vorgenommen werden. Die Probanden, welche 20 Euro erhielten weisen hingegen wieder einen positiven Effekt auf, d.h. es werden wieder weniger Fragen übersprungen. Dieses Muster lässt sich in den beiden anderen Faktoren nicht aufdecken. Im Rahmen der Reziprozitätsnorm weisen die Probanden mit einem Incentive von 5 Euro eine positive Antwortqualität auf, während 20 Euro wiederum eine Verschlechterung bewirken. Diese Entwicklung findet sich auch tendenziell bei der intrinsischen Motivation, da auch hier die Probanden mit einem Incentive in Höhe von 20 Euro nicht mehr die positiven Effekte der anderen Versuchsgruppen auf die Antwortqualität aufweisen.

Die genannten Befunde sprechen erneut für eine tendenziell positive Wirkung eines Incentives in Höhe von 5 Euro. Ein Incentive von 20 Euro weist hingegen eine komplexere Struktur auf: der einzige signifikante Effekt bewirkt eine Verbesserung der Antwortqualität, wohingegen die nicht signifikanten Regressionsgewichte auf eine Verschlechterung hinweisen.

d) Die Interpretation der Ergebnisse zu dem Befolgen von Anweisungen im Fragebogen

Auch hier wird zuerst geprüft, ob sich die unstandardisierten Regressionsgewichte der drei Versuchsgruppen bei den jeweiligen Faktoren signifikant voneinander unterscheiden. In

diesem Fall weist die verinnerlichte Reziprozitätsnorm signifikant unterschiedliche Wirkstrukturen auf: die Versuchsgruppe mit einem Incentive in Höhe von 5 Euro weist im Vergleich zu den beiden anderen Versuchsgruppen eine signifikant erhöhte Antwortqualität auf, da mehr Anweisungen im Fragebogen befolgt werden. Dieses Muster kann auch bei der Bewertung des Versuchsleiters beobachtet werden, wobei ein Incentive von 20 Euro eine leicht höhere positive Wirkung auf die Antwortqualität aufweist. Unter Betrachtung der Wirkungen im Rahmen der intrinsischen Motivation zeigt sich eine andere Entwicklung: der positive Effekt in der Versuchsgruppe ohne Incentive kann in den Versuchsgruppen mit einem Incentive nicht mehr nachgewiesen werden, es liegt vielmehr eine Verschlechterung der Antwortqualität vor. Diese Unterschiede sind allerdings nicht signifikant, wobei auch hier das empirische Signifikanzniveau nur leicht über der maximal akzeptierten Irrtumswahrscheinlichkeit von 10 % liegt (5€: $p = 0.162$; 20€ = 0.136).

Aufgrund dessen kann nicht von einer eindeutigen Verbesserung oder Verschlechterung der Antwortqualität bei Vergabe eines Incentives gesprochen werden. Der einzige signifikante Effekt ist positiv und bewirkt damit eine Steigerung der Antwortqualität bei einem Incentive in Höhe von 5 Euro. Die intrinsische Motivation weist hingegen einen negativen Effekt auf, welcher bei einem Incentive von 20 Euro noch höher erscheint. Ein Incentive von 20 Euro führt bei einer positiven Bewertung des Versuchsleiters wiederum zu einer Steigerung der Antwortqualität, da in dieser Versuchsgruppe mehr Arbeitsanweisungen befolgt werden.

Zusammengefasst lässt sich für die Effekte der Incentives auf die Antwortqualität nur eine grobe Einteilung vornehmen. Bei Betrachtung der Ergebnisse zur Bewertung des Versuchsleiters scheint ein Incentive die Antwortqualität tendenziell zu verbessern. Dies entspricht den theoretischen Annahmen, da die schenkende Person positiver wahrgenommen wird und damit Reziprozität gefördert wird. Jedoch lässt sich kein eindeutiges Wirkmuster feststellen, da das Incentive in Höhe von 5 Euro tendenziell andere Indikatoren der Antwortqualität positiv beeinflusst als das Incentive von 20 Euro. Darüber hinaus ist hervorzuheben, dass auch bei den Befragten ohne Incentive tendenziell negative Effekte auf die Antwortqualität festgestellt werden können. Dies bedeutet, dass mit einer steigenden positiven Bewertung der Person eine geringere Antwortqualität einhergeht. Dies kann über die Interpretation der Rolle der Befragten erklärt werden. Die Befragten können die Befragungssituation so interpretieren, dass jede abgegebene Antwort für die forschende Person wichtig und sinnvoll ist. Hierbei unterstellen die Befragten, dass der Fragebogen so qualitativ hochwertig konzipiert wurde und damit jede Antwort eine sinnvolle Antwort darstellt. Diese Interpretation der Ergebnisse wird dadurch gestützt, dass die Anzahl an Falschangaben mit positiver Bewertung der Person steigt, während die nicht begründeten Übersprünge signifikant sinken.

Die verinnerlichte Reziprozitätsnorm weist hingegen ein deutliches Wirkmuster auf: Ein Incentive von 5 Euro verbessert durchgängig die Antwortqualität, wohingegen ein Incentive von 20 Euro entgegen der theoretischen Vorannahmen eine Verschlechterung der Antwortqualität bewirkt. Dies kann dadurch erklärt werden, dass 5 Euro als freundliche Geste des guten Willens angesehen werden und damit ein reziprokes Verhalten fördern, während 20 Euro eher als Bezahlung wahrgenommen werden und damit einem reziproken Verhalten entgegenwir-

ken.¹⁵⁰ Die bevorzugte Interpretation lautet, dass ein Incentive in Höhe von 20 Euro zu Reaktanz bei den Befragten führen kann, da sich diese dadurch unter Druck gesetzt fühlen: „It is reasonable to assume, then, that if a person’s behavioral freedom is reduced or threatened with reduction, he will become motivationally aroused“ (Brehm (1966), S. 2). Diese Annahme kann durch Befunde über die Bewertung der Incentives durch die Befragten gestützt werden: Die Befragten der Versuchsgruppe mit einem Incentive in Höhe von 5 Euro gaben auf einer 7er-Skala zur Bewertung des Incentives (1 = negativ; 7 = positiv) durchschnittlich einen Wert von 4.56 an, wohingegen die Befragten mit einem Incentive in Höhe von 20 Euro einen leicht niedrigeren durchschnittlichen Wert mit 4.49 angaben. Hier wäre ein tendenziell höherer Mittelwert der Bewertung des Incentives zu erwarten, da 20 Euro einen deutlich höheren Wert und damit Nutzen für die Befragten aufweisen. Gleichzeitig wird das 20€-Incentive für die Teilnahme an einer Befragung als tendenziell zu hoch eingestuft (5.54) und bleibt den Befragten während des Befragungsprozesses eher präsent. Eine Interpretation der negativen Effekte bei Gabe von 20 Euro als Ergebnis von Reaktanz erscheinen damit plausibel.

Die Ergebnisse bezüglich der Wirkung der intrinsischen Motivation auf die Indikatoren der Antwortqualität weisen tendenziell auf einen negativen Effekt von Incentives hin. Interessant sind hierbei zwei Aspekte: zum einen kann auch in der Versuchsgruppe ohne Incentive nicht unterstellt werden, dass mit steigender intrinsischer Motivation zugleich eine Steigerung der Antwortqualität einhergeht. Dies widerspricht den Annahmen von Deci & Ryan, da nicht die erwarteten Effekte („enhanced performance, persistence, and creativity“ (Ryan & Deci (2000)),

¹⁵⁰ Der Fragebogen umfasste die Fragen, ob das Incentive von 5€, bzw. 20€ eher als Bezahlung (= 1) oder als Dankeschön (= 7) verstanden wird. Die Befragten in der Versuchsgruppe mit einem Incentive in Höhe von 5€ wählten durchschnittlich einen Wert von 5.82 während die Befragten mit einem Incentive in Höhe von 20€ durchschnittlich einen leicht niedrigeren Wert von 5.43 angaben. Auf Grundlage dieser Daten kann das Argument der unterschiedlichen Interpretation der Incentives nur schwerlich beibehalten erhalten werden.

S. 69)) auftreten. Zum anderen konnte bei Incentivierung kein Verlust an intrinsischer Motivation festgestellt werden, welcher diese negativen Effekte erklären könnte. Dies bedeutet, dass sich aufgrund der Vergabe eines Incentives nicht nur die Höhe der intrinsischen Motivation, sondern auch die Wirkweise bei den Befragten ändern kann. Es könnte hierbei entgegenargumentiert werden, dass die initiale intrinsische Motivation während der Bearbeitung des Fragebogens gesunken ist und daher bei den später folgenden Messungen der Antwortqualität negative Effekte aufweist. Dem kann allerdings entgegengesetzt werden, dass, gemäß Deci & Ryan (1985) ein Anstieg an intrinsischer Motivation dennoch prinzipiell zu einer besseren Arbeitsweise führen sollte. Alternativ – und dies ist die bevorzugte Erklärung – kann, wie bei den Befunden zur Bewertung des Versuchsleiters, die Interpretation der Befragungssituation durch die Befragten als Begründung herangezogen werden. Dies erscheint insofern plausibel, da auch in diesem Fall die Falschangaben steigen, während die nicht-begründeten Übersprünge sinken.¹⁵¹ Auch für die beiden anderen Versuchsgruppen mit einem Incentive kann dieses Erklärungsmuster herangezogen werden.

Werden die oben dargestellten Befunde auf die Hypothesen bezogen, so weist dies darauf hin, dass weder die positive Bewertung des Versuchsleiters, noch die verinnerlichte Reziprozitätsnorm sowie die intrinsische Motivation durchgängig positive Effekte auf die Antwortqualität aufweisen. Die Hypothesen 1b, 2a und 2c können nicht akzeptiert werden, da nicht einmal mehr als die Hälfte der unstandardisierten Regressionsgewichte das erwartete Vorzeichen aufweisen und damit für eine hohe Antwortqualität sprechen.

¹⁵¹ Es kann auch argumentiert werden, dass die initiale intrinsische Motivation während der Bearbeitung des Fragebogens gesunken ist und daher bei den später folgenden Messungen der Antwortqualität negative Effekte aufweist. Dem kann allerdings entgegengesetzt werden, dass, gemäß Deci & Ryan (1985) ein Anstieg an intrinsischer Motivation dennoch prinzipiell zu einer besseren Arbeitsweise führen sollte.

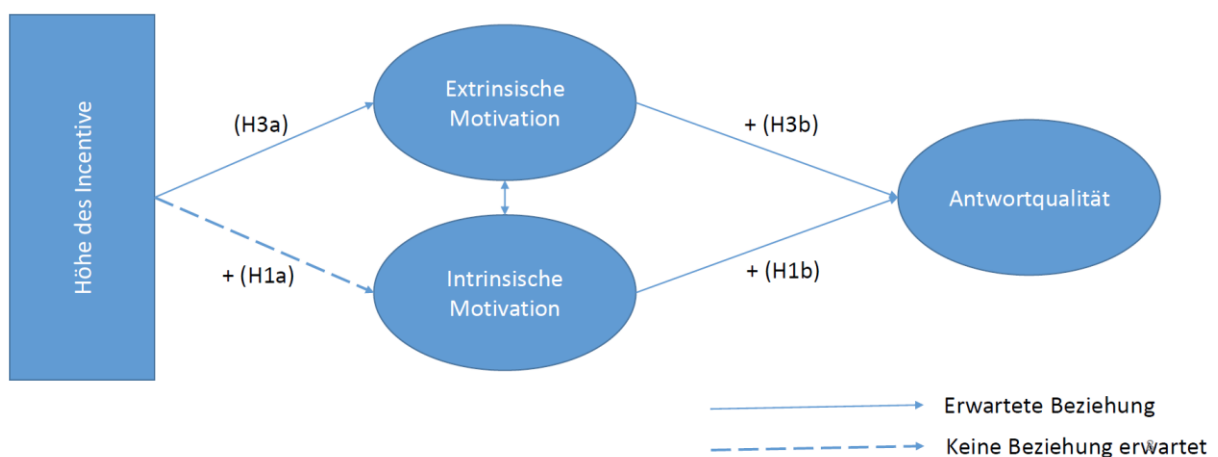
9.2 Prüfung der Hypothesen zur extrinsischen Motivation

Zur Prüfung der Hypothesen 3a und 3b wurde ein angepasstes Strukturgleichungsmodell erstellt.¹⁵² Hierbei wird auf die Prüfung der Hypothesen zur intrinsischen Motivation nicht mehr eingegangen, da diese bereits in Kapitel 9.1 geprüft wurden. Die erwarteten Hypothesen werden noch einmal berichtet und die Wirkrichtungen können der folgenden Graphik entnommen werden:

H3a: Je höher das unkonditionale Incentive, desto höher die extrinsische Motivation.

H3b: Je höher die extrinsische Motivation, desto höher die Antwortqualität.

Abb. 27: Die Hypothesen bezüglich der Wirkung der intrinsischen Motivation und der extrinsischen Motivation auf die Antwortqualität

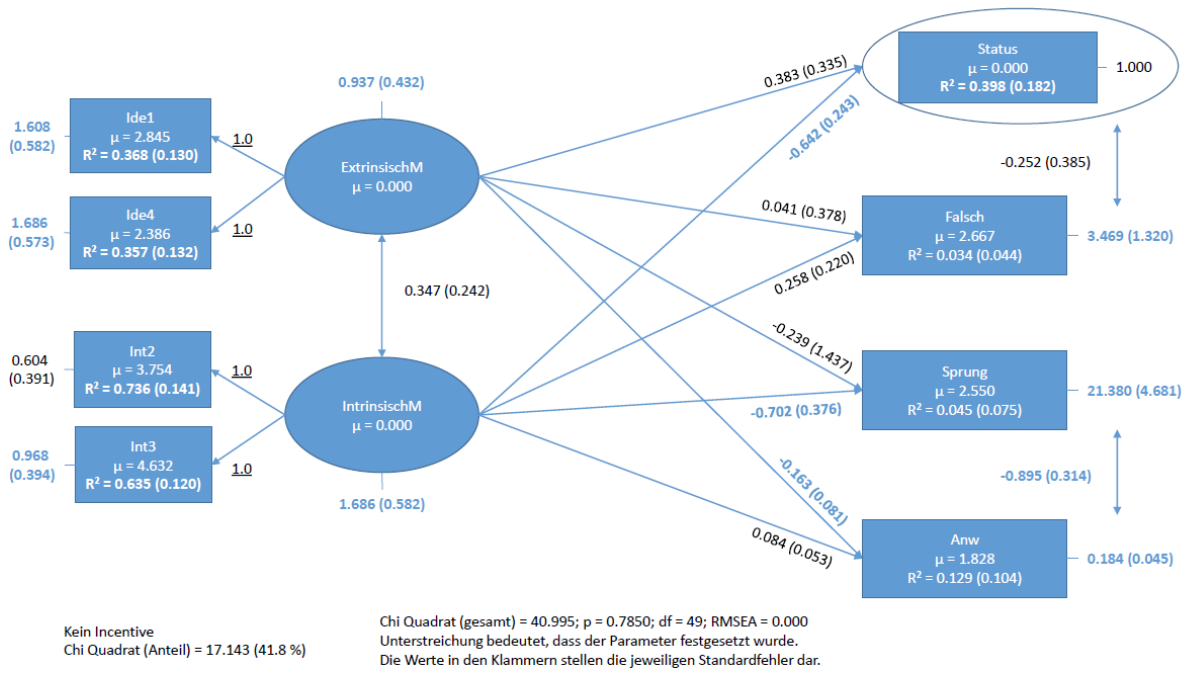


Quelle: eigene Darstellung.

Zwischen der extrinsischen und der intrinsischen Motivation wird, wie in Kapitel 4.1 dargelegt, eine ungerichtete Beziehung freigegeben. Es werden nun folgend die Ergebnisse der unstandardisierten Regressionsgewichte in Strukturgraphiken wiedergegeben, wobei – wie in den Strukturgleichungsmodellen zuvor – die restriktiven Gleichheitsbeziehungen zwischen den Versuchsgruppen aufgelöst wurden:

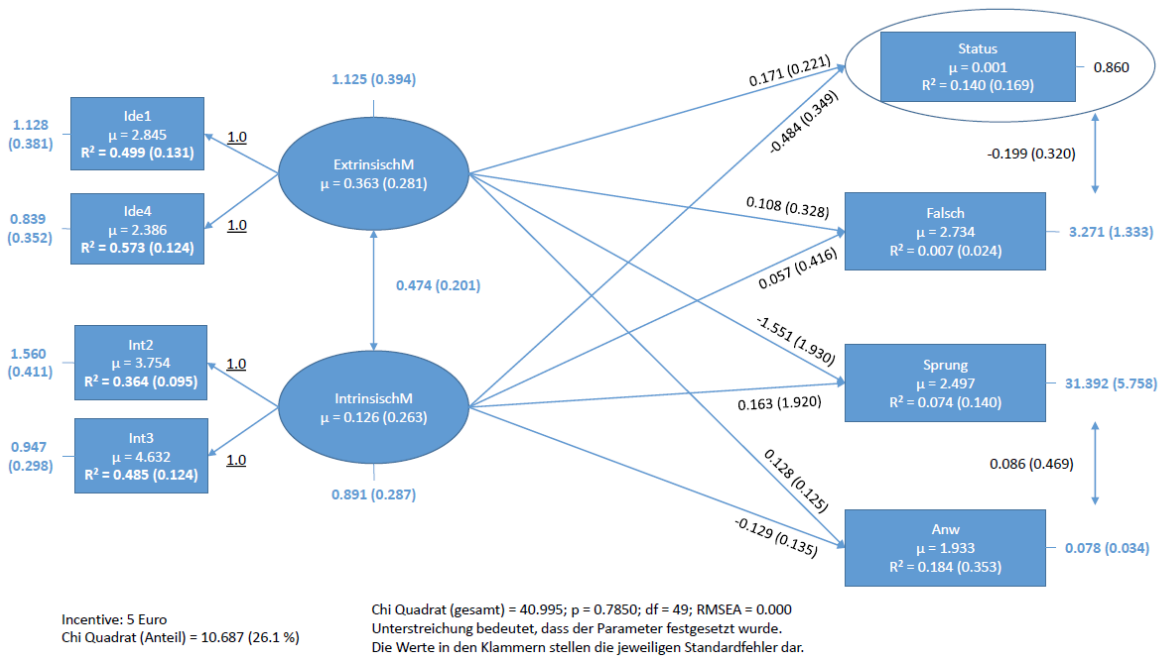
¹⁵² Die angeführten Namen für die Faktoren wurden zur einfacheren Interpretation in den Graphiken angepasst.

Abb. 28: Strukturgleichungsmodell für die Versuchsgruppe ohne Incentive, zur Prüfung der Hypothesen 3a und 3b¹⁵³



Quelle: eigene Darstellung auf Basis eigener Daten.

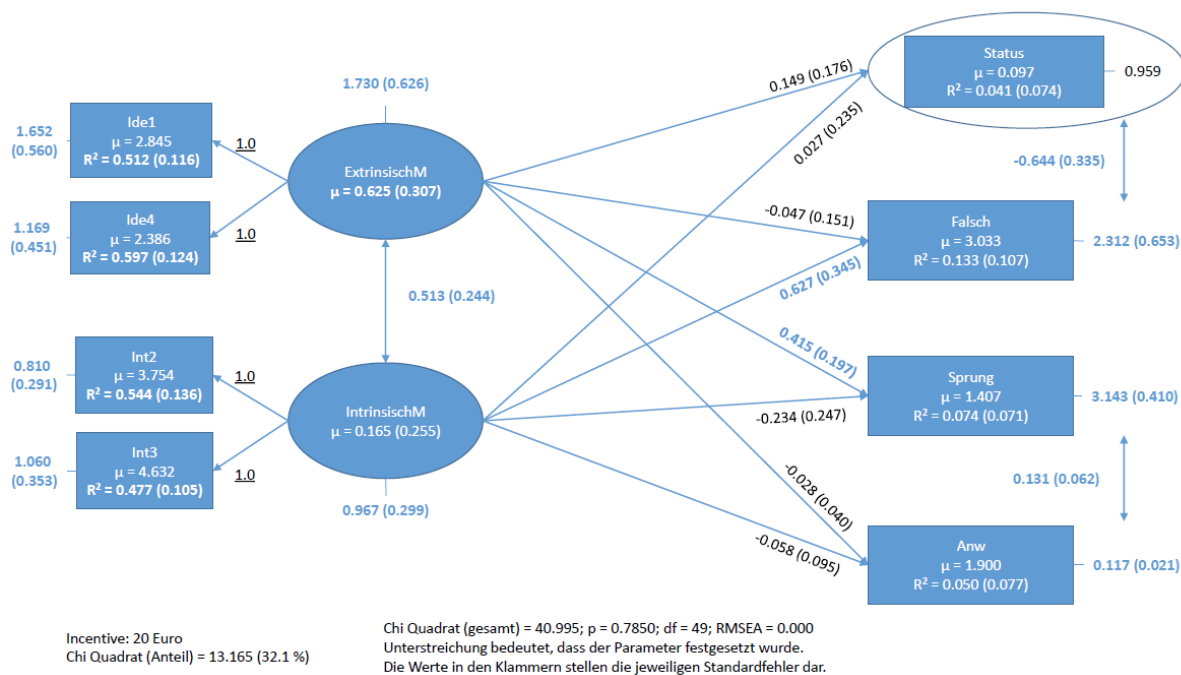
Abb. 29: Strukturgleichungsmodell für die Versuchsgruppe mit einem Incentive in Höhe von 5€, zur Prüfung der Hypothesen 3a und 3b



Quelle: eigene Darstellung auf Basis eigener Daten.

¹⁵³ Fett gedruckte Werte weisen auf signifikante Beziehungen hin, wobei hier aufgrund der geringen Fallzahl eine maximale Irrtumswahrscheinlichkeit von 10% akzeptiert wird. Die Werte der standardisierten Koeffizienten sind im Anhang, S. 209 – 210.

Abb. 30: Strukturgleichungsmodell für die Versuchsgruppe mit einem Incentive in Höhe von 20€, zur Prüfung der Hypothesen 3a und 3b



Quelle: eigene Darstellung auf Basis eigener Daten.

Die Mittelwert der extrinsischen Motivation steigt bei Vergabe von 5 Euro an ($p = 0.197$) und erreicht in der Versuchsgruppe mit einem Incentive von 20 Euro auch statistische Signifikanz ($p = 0.042$).¹⁵⁴ Dies bedeutet, dass die extrinsische Motivation durch die Gabe eines unbedingten Incentives erhöht werden kann. Das ist gemäß der Erwartung, da die extrinsische Motivation ein jegliches normbezogenes Verhalten umfasst und damit auch die Wirkung der Reziprozitätsnorm. Bei strenger Testung kann die Hypothese 3a jedoch nicht akzeptiert werden, da nur bei einem Incentive von 20 Euro ein signifikanter Anstieg erreicht wird.¹⁵⁵ Interes-

¹⁵⁴ Die identifizierte Regulation basiert auf zwei Items der SIMS. Das zweite und dritte Item der Skala mussten aufgrund unzureichender Korrelationen oder Mehrdimensionalität ausgeschlossen werden.

¹⁵⁵ Unter Berücksichtigung der geringen Fallzahl innerhalb einer Gruppe ($n = 60$) könnte für eine Akzeptanz der Hypothese argumentiert werden. Da allerdings unklar ist, ob die statistischen Zusammenhänge bei einer höheren Fallzahl auch die gewünschten Signifikanzen aufweisen, wird die Hypothese 3a nicht akzeptiert.

sant ist in diesem Kontext auf die Varianz des Faktors zur extrinsischen Motivation zu verweisen, da auch diese bei Vergabe eines Incentives steigt. So ist der Wert der Varianz in der Versuchsgruppe mit dem Incentive in Höhe von 20 Euro im Vergleich zur Versuchsgruppe ohne Incentive fast doppelt so hoch. Auch wenn diese Abweichungen nicht signifikant sind, so erscheint das in Kapitel 9.1 genutzte Argument der Akquieszenz zur Erklärung der Verringerung der Varianz der intrinsischen Motivation doch als eher unplausibel, da die Items zur Messung der extrinsischen Motivation zur selben Skala wie die Items der intrinsischen Motivation gehören und somit in der gleichen Item-Matrix abgefragt wurden.

Die Hypothese 3b, also die Vermutung das die extrinsische Motivation einen positiven Effekt auf die Antwortqualität aufweist, soll ebenfalls unter Verwendung der unstandardisierten Regressionsgewichte geprüft werden. Dafür werden auch hier die Regressionsgewichte für die jeweiligen drei Versuchsgruppen in einer einzigen Tabelle zusammengefasst, wobei erneut rote Werte für eine Verschlechterung und grüne Werte für eine Verbesserung der Antwortqualität stehen:

Tab. 44: Unstandardisierte Regressionsgewichte des Strukturgleichungsmodells, aufgliedert nach den drei Versuchsgruppen (Hypothesen 3a – 3b)

	Extrinsische Motivation			Intrinsische Motivation		
	0 Euro	5 Euro	20 Euro	0 Euro	5 Euro	20 Euro
Status Quo	0.383 (0.335)	0.171 (0.221)	0.149 (0.176)	-0.642 (0.243)	-0.484 (0.349)	0.027 (0.235)
Falschangaben	0.041 (0.378)	0.108 (0.328)	-0.047 (0.151)	0.258 (0.220)	0.057 (0.416)	0.627 (0.345)
Überspringen	-0.239 (1.437)	-1.551 (1.930)	0.415 (0.197)	-0.702 (0.376)	0.163 (1.920)	-0.234 (0.247)
Anweisungen	-0.163 (0.081)	0.128 (0.125)	-0.028 (0.040)	0.084 (0.053)	-0.129 (0.135)	-0.058 (0.095)

Quelle: eigene Daten. Fett gedruckte Werte weisen eine empirische Signifikanz von $p < 0.10$ auf. Die Werte in den Klammern entsprechen den Standardfehlern.

a) Die Interpretation der Ergebnisse zum Status Quo-Effekt

Zuerst werden Wald-Tests berechnet, um herauszufinden, inwiefern sich die unstandardisierten Regressionsgewichte der jeweiligen Faktoren untereinander signifikant unterscheiden. Hierbei kann nur für die intrinsische Motivation ein statistisch bedeutsamer Unterschied festgestellt werden: die Versuchsgruppe der Befragten ohne Incentive weist im Vergleich mit der Versuchsgruppe mit einem Incentive in Höhe von 20 Euro eine signifikante, negative Wirkung auf die Antwortqualität auf ($p = 0.0312$). Dies entspricht dem Muster, welches bereits in Kapitel 9.1 aufgezeigt wurde. Die extrinsische Motivation hat hingegen einen durchgängig positiven Einfluss auf die Antwortqualität, wobei dies vorsichtig interpretiert werden sollte, da keiner der Effekte signifikant von Null verschieden ist. Es ist hierbei hervorzuheben, dass die extrinsische Motivation über die Versuchsgruppen hinweg steigt, aber auf Ebene der Antwortqualität keine einheitlichen Steigerungen aufweist.

b) Die Interpretation der Ergebnisse zu Pseudo-Opinions (Falschantworten)

Die Ergebnisse der Wald-Tests ergeben, dass die unstandardisierten Regressionsgewichte der extrinsischen Motivation und der intrinsischen Motivation zwischen den Versuchsgruppen keine signifikanten Unterschiede aufweisen. Die Vorzeichen der Regressionsgewichte deuten darauf hin, dass über alle Versuchsgruppen hinweg die Anzahl an Falschangaben mit steigender identifizierter Regulation und intrinsischer Motivation zunimmt. Auch hier bewirkt ein höheres Incentive keinen Anstieg der Antwortqualität. Mit Blick auf die intrinsische Motivation kann bei Gabe eines Incentives in Höhe von 20 Euro ein signifikanter negativer Effekt auf die Antwortqualität festgestellt werden. Dies könnte als Indiz

dafür gesehen werden, dass auf zu hohe Incentives verzichtet werden sollte, da die extrinsische Motivation keine Steigerung der Antwortqualität bewirkt und negative Effekte der intrinsischen Motivation auftreten.

c) Die Interpretation der Ergebnisse zu nicht-begründeten Übersprünge

Auch in diesem Fall können keine signifikanten Unterschiede zwischen den einzelnen unstandardisierten Regressionsgewichten der jeweiligen Faktoren im Bezug auf die drei Versuchsgruppen festgestellt werden. Ein Incentive in Höhe von 5 Euro weist im Rahmen der extrinsischen Motivation einen positiven Effekt auf die Antwortqualität auf, auch wenn dieser nicht statistisch signifikant ist. Bei der Versuchsgruppe mit einem Incentive von 20 Euro kann hingegen ein signifikanter, negativer Effekt auf die Antwortqualität festgestellt werden. Dies weist darauf hin, dass die Gabe eines Incentives die extrinsische Motivation zwar steigert, aber nicht zu einer sorgfältigeren Bearbeitung führt.

d) Die Interpretation der Ergebnisse zu dem Befolgen von Anweisungen im Fragebogen

Die extrinsische Motivation weist einen signifikanten, negativen Effekt in der Versuchsgruppe ohne Incentive auf. Dies bedeutet, dass die Befragten die gestellten Arbeitsanweisungen seltener befolgen. Wird ein Incentive von 5 Euro ausgegeben, dann kann hingegen ein positiver Effekt beobachtet werden, d.h. die Befragten weisen eine erhöhte Antwortqualität auf, da sie mehr Arbeitsanweisungen befolgen. Dies könnte darauf hinweisen, dass die Befragten erst durch einen externen Anreiz den erhöhten Aufwand akzeptieren und damit einer extrinsischen Motivation bedürfen. Dem widerspricht jedoch, dass bei ei-

ner Gabe von 20 Euro der Effekt auf die Antwortqualität erneut negativ ist und damit Anweisungen weniger befolgt werden. Dies ist hervorzuheben, da eine höhere extrinsische Motivation folglich nicht zu einer höheren Antwortqualität führt.

Zusammengefasst kann festgestellt werden, dass die extrinsische Motivation bei Vergabe eines Incentives von 5 Euro tendenziell positive Effekte auf die Antwortqualität aufweist. Es kann jedoch erneut festgestellt werden, dass ein Absinken der nicht-begründeten Übersprünge mit einem Anstieg an Falschangaben verbunden ist. Dies spricht wieder dafür, dass die Befragten ein jegliches Beantworten von Fragebogenfragen als positive Handlung zugunsten des Forschers interpretieren und folglich häufiger Falschangaben machen.¹⁵⁶

Bei Einem Incentive von 20 Euro können zumeist negative Effekte festgestellt werden, welche analog zu den Befunden zur verinnerlichten Reziprozitätsnorm im Kapitel 9.1 interpretiert werden können. Dies ist möglich, da die extrinsische Motivation – gemäß der Darstellung von Deci & Ryan (1985) – die Wirkung der Reziprozitätsnorm umfasst. Somit können auch hier die negativen Effekte als Ergebnis von Reaktanz interpretiert werden.¹⁵⁷ Es muss an dieser Stelle darauf hingewiesen werden, dass neben der Reziprozitätsnorm noch weitere Normen oder Situationsbewertungen Einfluss auf die Ergebnisse der extrinsischen Motivation haben können.

¹⁵⁶ Dies kann im Rahmen des Framing-Ansatzes von Esser (1999) verstanden werden, welcher die Definition der Situation deutlich hervorhebt. Stocké (2002; 2004) weist auf die Bedeutung der Definition der Situation in der Umfrageforschung hin.

¹⁵⁷ Da die Effekte der intrinsischen Motivation den bereits berichteten Ergebnissen in Kapitel 9.1. entsprechen soll auf eine erneute Interpretation verzichtet werden.

Die Hypothese 3b kann mit Blick auf die Ergebnisse nicht akzeptiert werden, da lediglich sechs der zwölf Koeffizienten in die erwartete Richtung weisen und zum anderen die einzigen signifikanten Regressionsgewichte ein gegenläufiges Vorzeichen aufweisen und damit für eine Verschlechterung der Antwortqualität sprechen.

9.3 Zusammenfassung der Ergebnisse

Im vierten Kapitel wurden Hypothesen zur Wirkung von Incentives auf die Antwortqualität aus der Cognitive Evaluation Theory und der Reziprozitätshypothese abgeleitet. Diesen theoretischen Ansätzen folgend wurde stets ein positiver Effekt auf die Antwortqualität vorhergesagt.

Zur Prüfung der Hypothesen wurden Indikatoren der Antwortqualität nach theoretischen und empirischen Argumentationen ausgewählt und in Strukturgleichungsmodellen als abhängig modelliert. Entgegen der theoretischen Erwartungen wurde keine durchgängige Verbesserung der Antwortqualität in den Versuchsgruppen mit einem Incentive festgestellt. Es konnte sogar oftmals eine Verschlechterung der Antwortqualität beobachtet werden. Aus diesem Grund konnten die Zusammenhangshypothesen (1b, 2a, 2c und 3b) nicht bestätigt werden. Dabei ist hervorzuheben, dass in der Versuchsgruppe ohne Incentive ebenfalls tendenziell negative Zusammenhänge der intrinsischen Motivation auf die Facetten der Antwortqualität festgestellt werden konnten. Dies ist insofern überraschend, da die Teilnehmer an dieser Studie als prinzipiell höher motiviert eingestuft wurden. Eine Erklärung lässt sich auf die Definition der Befragtenrolle zurückführen. Es wird vermutet, dass die Befragten die Forscher in ihren Studien unterstützen wollen, aber aufgrund von Fehlinterpretationen über die Ziele und Erwartungen der Forscher zu einem unerwünschten Antwortverhalten tendieren. Ein solches unerwünschtes Antwortverhalten kann eine vermehrte Anzahl von Falschangaben bedeuten,

da die Befragten unterstellen könnten, dass jede Antwort für den Forscher wichtig und sinnvoll ist.¹⁵⁸ Ein solches Verhalten findet sich in den Ergebnissen dieser Studie wieder: es werden bei steigender intrinsischer Motivation oder Reziprozität vermehrt Falschangaben gemacht, wobei zugleich die Anzahl an unbegründeten Übersprüngen sinkt. Zugespitzt kann aus dieser Erklärung heraus die Vermutung formuliert werden, dass mit steigender intrinsischer Motivation, bzw. Reziprozität höchstens der Wille der Befragten für eine verbesserte Antwortqualität steigt.

Es konnten, neben diesen unerwarteten und widersprüchlichen Ergebnissen auch Hypothesen bestätigt werden: So steigt die extrinsische Motivation bei Vergabe eines Incentives an, wobei der erwartete positive Effekt auf die Antwortqualität - wie bereits berichtet - jedoch ausbleibt. Die intrinsische Motivation verhält sich bei Incentivierung ebenfalls erwartungsgemäß, da keine signifikanten Zu- oder Abnahmen festgestellt werden konnten.

Die Ergebnisse der Hypothesenprüfungen werden noch einmal in der folgenden Tabelle zusammengefasst:

¹⁵⁸ Dies ist insofern plausibel, da bei einem Fragebogen eine sinnvolle Konstruktion unter Verwendung von sinnvollen und notwendigen Fragen erwartet werden kann.

Tab. 45: Übersicht über die Ergebnisse der Hypothesenprüfungen¹⁵⁹

Hypothese	Akzeptiert
<i>H1a: Der Erhalt eines un konditionalen Incentives hat keinen Einfluss auf die Höhe der intrinsischen Motivation.</i>	✓
<i>H1b: Je höher die intrinsische Motivation, desto höher die Antwortqualität.</i>	✗
<i>H2a: Je höher das un konditionale Incentive, desto stärker wirkt die verinnerlichte Reziprozitätsnorm positiv auf die Antwortqualität.</i>	✗
<i>H2b: Je höher das un konditionale Incentive, desto positiver wird der Belohnungsgeber wahrgenommen.</i>	✗
<i>H2c: Je positiver der Belohnungsgeber wahrgenommen wird, desto höher die Antwortqualität.</i>	✗
<i>H3a: Je höher das un konditionale Incentive, desto höher die identifizierte Motivation.</i>	(✓)
<i>H3b: Je höher die identifizierte Motivation, desto höher die Antwortqualität.</i>	✗

Soll abschließend die Wirkung der Incentives auf die Antwortqualität beurteilt werden, so könnte von einem leichten positiven Effekt eines Incentives in Höhe von 5 Euro gesprochen werden, da die Antwortqualität oftmals stieg, oder zumindest negative Effekte auf die Antwortqualität – im Vergleich zur Versuchsgruppe ohne Incentive – abgeschwächt werden konnten. Bei einer Incentivierung von 20 Euro werden hingegen die negativen Effekte auf die Antwortqualität deutlicher. Hier kann neben der Definition der Befragtenrolle auch die Reaktanz als Erklärung herangezogen werden, da sich die Befragten aufgrund des hohen Incentives z.B. aufgrund eines hohen Reziprozitätsdrucks in ihrer Verhaltensfreiheit eingeschränkt fühlen.

¹⁵⁹ Die Hypothese H2b wird aufgrund fehlender statistischer Signifikanz abgelehnt, auch wenn deskriptiv ein Anstieg festgestellt werden konnte.

10. Diskussion der internen und externen Validität der Ergebnisse

Vor einer abschließenden Bewertung soll die Reichweite der im neunten Kapitel berichteten Ergebnisse diskutiert werden, da sich aus dem methodischen Vorgehen in dieser Studie Limitationen ergeben können.

So kann argumentiert werden, dass aufgrund der unterschiedlichen Rekrutierungsstrategien die interne Validität eingeschränkt ist, da dadurch unterschiedliche Stimuli bei den Befragten gesetzt werden könnten. Selbstverständlich wäre ein einheitliches Rekrutierungsverfahren optimal gewesen, dies konnte jedoch aus den folgenden Gründen nicht umgesetzt werden: a) Die Erhebung begann am 13.04.2015 und damit zu Beginn der Vorlesungszeit. Es wurde hierbei als wichtig betrachtet, dass die Erhebungsphase zum Ende der Vorlesungszeit abgeschlossen ist, da die Studierenden neben den Klausurvorbereitungen auch in evtl. folgende Hausarbeits-, Praktika- oder Urlaubsphasen übergehen und damit für eine Teilnahme an der Studie nicht mehr zur Verfügung stehen. Die angestrebte Rücklaufquote für die Minimalfallzahl konnte für dieses Zeitfenster jedoch nur unter Erweiterung der Rekrutierungsverfahren eingehalten werden. b) Mit einer längeren Feldphase stieg zudem das Risiko, dass die Befragten, trotz eines aufklärenden Nachgesprächs – mit Bitte um Verschwiegenheit – die Information über ein Incentive weitergeben und damit das Experiment scheitert. Auch aus diesem Grund wurden die neuen Rekrutierungsstrategien eingebunden, da damit die Rücklaufgeschwindigkeit erhöht werden sollte und folglich das Risiko einer Verbreitung der Information über Incentives sinkt. Zur Kontrolle der Wirkung der verschiedenen Rekrutierungsstrategien sollten die Probanden im Fragebogen angeben über welche Rekrutierungsstrategien sie auf die Befragung aufmerksam wurden. Hierbei konnten keine systematischen Zusammenhänge mit den Indikatoren der multivariaten Analyse festgestellt werden.

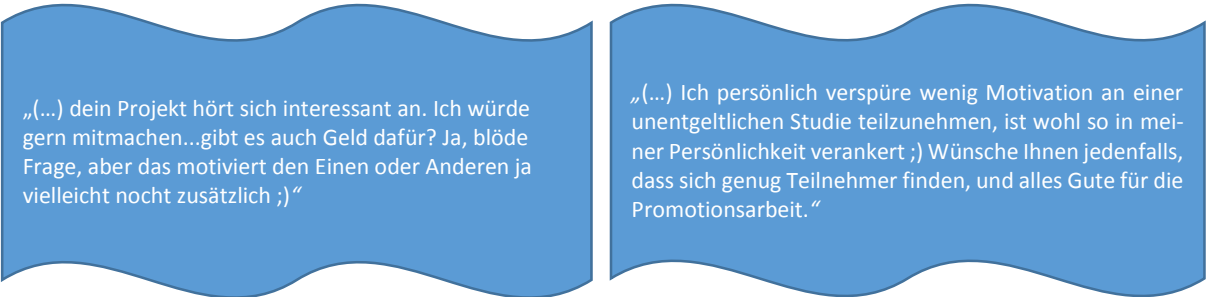
Die befürchtete Korrumpierung des Designs aufgrund der Kenntnis der Incentives konnte ebenfalls nicht festgestellt werden. Dies spiegelt sich zum einen in der durchgängig geringen Rücklaufquote wider. Hier wäre bei Kenntnis der Incentives ein wahrnehmbarer Anstieg zu erwarten gewesen. Zum anderen wurde im Rahmen der Befragung erhoben, ob und inwiefern bei den Probanden Vorinformationen bezüglich der Studie vorlagen. Aus diesen Angaben ist ebenfalls nicht ersichtlich, dass die Befragten vor der Teilnahme Kenntnis von einem Incentive hatten.

Es kann darüber hinaus auch diskutiert werden, dass aufgrund der Laborsituation die externe Validität eingeschränkt ist. Die Argumentation lautet hierbei, dass die Befragten auf die Erhebungssituation reagieren und sich in Folge dessen nicht natürlich Verhalten. Für eine solche Argumentation spricht z.B. die sehr geringe Abbruchquote von ca. 1%, welche in Online-Befragungen (wenn überhaupt) nur sehr selten erreicht wird. Dem kann allerdings erwidert werden, dass im Gegenzug die interne Validität in Laborexperimenten zumeist recht hoch ist, da äußere Einflüsse während der Erhebung kontrolliert werden können. Zum anderen wäre dann die für dieses Design notwendige Vergabe der un konditionalen Incentives problematisch: In postalischen Befragungen kann das monetäre Incentive zwar problemlos beigelegt werden, aber dafür können viele Zusatzinformationen (z.B. Bearbeitungszeiten, Pausen beim Bearbeiten des Fragebogens) nicht erfasst werden. Bei Online-Befragungen können diese Zusatzinformationen zwar erhoben werden, aber hier ist es wiederum sehr fragwürdig, inwiefern das un konditionale Incentive (operationalisiert z.B. über digitale Gutscheine oder Bonuspunkte für Online-Shops) noch die gleiche Wirkung entfaltet.

Die Laborsituation könnte weitergehend dadurch problematisiert werden, da das Erhebungslabor über eine Verbindungstür an den Besprechungsraum des Forschers angebunden war. Somit lag eine räumliche Nähe des Forschers zu den Probanden während des Befragungsprozesses vor. Dies könnte einen Verlust des Anonymitätsempfindens bewirkt haben und damit die Validität beeinträchtigen. Dem wurde versucht entgegenzuwirken, indem zum einen den Probanden zu Beginn der Befragung deutlich Anonymität zugesichert wurde. Im Fragebogen wurde dennoch absichernd erhoben, inwiefern die Befragten während der Teilnahme an der Studie die Anonymität einstufen. Hierbei geben die Probanden auf einer 7stufigen Antwortskala (7 = sehr hohe Anonymität) einen durchschnittlichen Wert von 5.21 an. Damit kann tendenziell von einer gewährten Anonymität gesprochen werden. Zum anderen wurden die Probanden gebeten das Erhebungslabor nach Beendigung der Befragung durch einen separaten Ausgang zu verlassen, welcher nicht direkt mit dem Besprechungsraum des Forschers verbunden ist. Dadurch sollte ihnen die Möglichkeit eines vermeintlich unbeobachteten Verlassens der Befragungssituation gegeben werden. Unterstützt wurde dieses Vorgehen dadurch dass die später folgenden Abschlussgespräche nicht angekündigt wurden und den Teilnehmern das Gefühl einer möglichen Rechtfertigungsverpflichtung für das individuelle Antwort- oder Abbruchverhalten genommen werden sollte.

Mit Bezug auf die externe Validität kann auch angemerkt werden, dass die in dieser Studie genutzte Incentivierungs-Strategie (= unangekündigt und unkonditionale) nicht der alltäglichen Praxis entspricht und die Ergebnisse daher nicht für z.B. konditionale oder angekündigte Incentives verallgemeinert werden können. Dieser Argumentation kann jedoch erwidert werden, dass die Forschungsfrage nur unter dieser Experimentalbedingung durchgeführt werden

konnte. Dies liegt darin begründet, dass nur so für die drei Versuchsgruppen die gleichen Voraussetzungen geschaffen werden konnten und sie sich daher, aufgrund der Randomisierung nur im Grad der Incentivierung unterscheiden. Darüber hinaus kann aufgrund der unangekündigten Vergabe des Incentives bei den Probanden eine erhöhte intrinsische Motivation unterstellt werden, welche in Befragungen mit Anreizsystemen nicht so stark gegeben sein sollte. Eine solche geringe intrinsische Motivation zeigt sich beispielhaft in einigen eMail-Antworten bezüglich der Einladung zur Befragung:



„(...) dein Projekt hört sich interessant an. Ich würde gern mitmachen...gibt es auch Geld dafür? Ja, blöde Frage, aber das motiviert den Einen oder Anderen ja vielleicht noch zusätzlich ;)“

„(...) Ich persönlich verspüre wenig Motivation an einer unentgeltlichen Studie teilzunehmen, ist wohl so in meiner Persönlichkeit verankert ;) Wünsche Ihnen jedenfalls, dass sich genug Teilnehmer finden, und alles Gute für die Promotionsarbeit.“

Da die höher motivierten Befragten bereits negative Effekte auf die Antwortqualität aufweisen, erscheint es daher plausibel, dass bei weniger motivierten Befragten zumindest ähnliche negative Wirkstrukturen vorliegen. Im Zusammenhang mit der Incentivierungs-Strategie muss allerdings angemerkt werden, dass aufgrund fehlender zeitlicher und finanzieller Ressourcen leider kein Doppel-Blind-Design umgesetzt werden konnte. Dies wäre sehr wünschenswert gewesen, da damit mögliche erwartungsbezogene Einflüsse des Forschers auf die Befragten ausgeschlossen werden könnten.

Neben Aspekten des Erhebungsdesigns soll auch die Validität der genutzten Indikatoren angesprochen werden. In dieser Studie wurden diese zwar auf Basis theoretischer und empiri-

scher Überlegungen ausgewählt, aber dennoch bestehen Unsicherheiten in der Interpretation: Inwiefern sind die Messungen der Antwortqualität stabil? Müssen die ausgewählten Indikatoren in den multivariaten Analysen unterschiedlich gewichtet werden? Sind die Ergebnisse dieser Studie auch auf konditionale Incentives oder Designs mit vorher angekündigtem Incentive übertragbar? Diese offenen Fragen können jedoch nicht im Rahmen einer einzigen Studie geklärt werden. Diese Studie leistete jedoch ein erstes systematisches Aufdecken der heterogenen Abhängigkeitsstrukturen und den sich daraus ergebenden Konsequenzen. Diese zentralen Erkenntnisse wurden vorher noch nicht in diesem Umfang dargestellt und ermöglichen damit einen neuen Blick auf das gesamte Forschungsfeld. Zur Klärung der oben angedeuteten offenen Fragen sind jedoch noch weitere Forschungen von Nöten.

11. Fazit

Nach Darstellung der zentralen Ergebnisse in Kapitel 9.3 und der Diskussion der internen und externen Validität in Kapitel 10 wird in diesem Kapitel das Vorgehen in dieser Studie noch einmal zusammengefasst und das Gesamtfazit gezogen:

Zu Beginn wurde eine Definition der Antwortqualität entwickelt, welche zwischen vier Facetten unterscheidet: einem durchdachten Beantworten, einem (situational) wahrheitsgemäßen Beantworten, einem vollständigen Beantworten und einem anweisungsbefolgenden Beantworten. Im dritten Kapitel wurde der aktuelle Forschungsstand zu Incentives berichtet und dabei, mit besonderem Blick auf die Antwortqualität, auf Defizite beim aktuellen Forschungsstand aufmerksam gemacht. In diesem Kontext wurde auch das Ziel dieser Forschungsarbeit vorgestellt und im Forschungsfeld eingebettet. Darauf aufbauend wurden im vierten Kapitel

zwei theoretische Ansätze vorgestellt, welche zur Erklärung einer Wirkung von Incentives auf die Antwortqualität hinzugezogen werden können: die Cognitive Evaluation Theory (Deci & Ryan (1985)) und die Reziprozitätshypothese (Gouldner (1960)). Für die späteren Analysen wurden aus diesen Ansätzen sieben Kausalhypothesen abgeleitet, welche stets eine positive Wirkung von Incentives auf die Antwortqualität vorhersagten. Zum Verständnis des Aufbaus der Studie wurde im fünften Kapitel das experimentelle Design und der Befragungskontext berichtet. Nach Erläuterung der Befragungssituation wurden im sechsten Kapitel die verschiedenen Indikatoren der (intrinsischen und extrinsischen) Motivation, der Reziprozität und der vier Facetten der Antwortqualität vorgestellt und operationalisiert, wobei im siebten und achten Kapitel die Messqualität geprüft wurde. Darauf aufbauend konnten dann im neunten Kapitel die Hypothesenprüfungen vorgenommen werden.

Die zentralen Ergebnisse lassen hierbei folgende Schlüsse zu:

- 1. Incentives weisen heterogene Effekte auf die vier Facetten der Antwortqualität auf. Die Höhe des Incentives beeinflusst hierbei nicht nur die Stärke der Effekte, sondern auch deren Wirkrichtung.*
- 2. Ein Incentive in Höhe von 5 Euro weist tendenziell positive Effekte bezüglich der vier Facetten der Antwortqualität auf. Bei einem Incentive in Höhe von 20 Euro können dagegen eher negative Effekte beobachtet werden.*

Bei Betrachtung der Höhe des Incentives kann nur unter Vorbehalt von einem positiven Effekt von Incentives gesprochen werden. Die multivariaten Analysen deuten dabei darauf hin, dass

bei einem zu hochwertigen Incentive eine Verschlechterung der Antwortqualität die Folge sein kann. Dieser negative Effekt wurde im Rahmen dieser Studie über Reaktanz erklärt, da ein hohes Incentive einen (zu) starken Reziprozitätsdruck oder Verantwortungsdruck aufbauen kann. Die Ergebnisse sprechen also dafür, dass der Wert eines Incentive nicht übertrieben werden darf.

3. *Die negativen Effekte auf die Antwortqualität lassen sich aus der Definition der Rolle des Befragten ableiten. Hierbei wird vermutet, dass die Befragten die Forscher in ihren Studien unterstützen wollen, aber aufgrund von Fehlinterpretationen über die Ziele und Erwartungen der Forscher zu einem unerwünschten Antwortverhalten tendieren.*

Die widersprüchlichen Ergebnisse bezüglich der Antwortqualität könnten als Ergebnis einer Unfähigkeit der Befragten oder gar als Sabotageakte interpretiert werden. Dieser Schluss erscheint jedoch verkürzt und unangemessen. Daher wurde eine weitere Erklärung in Betracht gezogen: Unter der Prämisse, dass die Befragten den Forscher in seiner Arbeit unterstützen wollen, stellen sie ihm deswegen mehr Informationen für seine Auswertungen zur Verfügung. Diese Schlussfolgerung basiert auf dem Befund, dass die im Fragebogen erfassten Falschangaben fast durchgängig ansteigen. Diese Vermutung wird auch dadurch gestützt, dass parallel ein Absinken von Item-Nonresponse in den Daten beobachtet werden kann.

Eine Abnahme des Item-Nonresponse bei Incentivierung von Befragten wurde bereits in anderen Experimentalstudien festgestellt (Medway (2012); Singer et al. (1999); Berk et al. (1987)). In diesem Kontext wurde in Leitfäden (vgl. Pforr (2015); Stadtmüller & Porst (2005)) und Enzyklopädien (vgl. Boulianne (2008)) üblicherweise von positiven Wirkungen von Incentives auf die Antwortqualität gesprochen. Es erscheint jedoch unter Anbetracht der Ergebnisse

dieser Studie sehr unsicher, ob die vermeintlich positiven Befunde auch wirklich positiv zu interpretieren und nicht mit mehr Falschangaben verbunden sind.

Die Bedeutung der Wahrnehmung und Interpretation der Befragungssituation für die Antwortqualität wurde schon in Studien zur sozialen Erwünschtheit aufgezeigt (Skarbek-Kozietulska et al. (2012); Stocké (2004)). Die Definition der Befragungssituation scheint nun auch für die Wirkung von Incentives auf die Antwortqualität eine bedeutende Rolle zu spielen. Damit die zugrunde liegenden Mechanismen von Incentives deutlicher herausgearbeitet werden können, ist jedoch noch viel Forschung von Nöten.

12. Literaturverzeichnis

Achen, Christopher H. (1975): Mass Political Attitudes and the Survey Response. IN: American Political Science Review 69, S. 1218 – 1231.

Adloff, Frank / Mau, Steffen (2005): Vom Geben und Nehmen. Zur Soziologie der Reziprozität. Campus Verlag: Frankfurt am Main.

Armer, Michael/ Baldigo, Jeannin (1973): A Transitive Index to Test for Acquiescent Response Style. IN: Journal of Social Psychology 90, S. 185 – 196.

Arzheimer, Kai/ Klein, Markus (1998): Die Wirkung materieller Incentives auf den Rücklauf einer schriftlichen Panelbefragung. IN: ZA-Informationen 48, S. 6 – 31.

Bachleitner, Reinhard/ Weichbold, Martin/ Aschauer, Wolfgang (2010): Die Befragung im Kontext von Raum, Zeit und Befindlichkeit. Beiträge zu einer prozessorientierten Theorie der Umfrageforschung. Wiesbaden: VS-Verlag.

Barge, Scott/ Gehlbach, Hunter (2012): Using the Theory of Satisficing to Evaluate the Quality of Survey Data. IN: Research in Higher Education 53, S. 182 – 200.

Bem, Daryl J. (1972): Self-Perception Theory. Self-perception theory. IN: Berkowitz, Leonard (Hg.): Advances in Experimental Social Psychology 6. New York: Academic Press, S. 1 – 62.

Bergius, Rudolf (2014): Autotelisch. IN: Wirtz, Markus Antonius (Hg.): Dorsch – Lexikon der Psychologie (17. Auflage), S. 242. Bern: Verlag Hans Huber.

Berlin, Martha/ Mohadjer, Leyla/ Waksberg, Joseph/ Kolstad, Andrew/ Kirsch, Irwin/ Rock, Donald/ Yamamoto, Kentaro (1992): An Experiment in Monetary Incentives. IN: Proceedings of the Survey Research Methods Section of the American Statistical Association, S. 393 – 398.

Biemer, Paul P./ Groves, Robert M./ Lyberg, Lars E./ Mathiowetz, Nancy A./ Sudman, Seymour (1991): Measurement Errors in Surveys. New York/ Chichester/ Brisbane/ Toronto/ Singapore: Wiley & Sons.

Biemer, Paul P./Lyberg, Lars E. (2003): Introduction to Survey Quality. Hoboken (New Jersey): Wiley & Sons.

Blau, Peter Michael (1964): Exchange and Power in Social Life. New York: Wiley.

Bonke, Jens/ Fallesen, Peter (2010): The impact of incentives and interview methods on response quantity and quality in diary- and booklet-based surveys. IN: Survey Research Methods 4, S. 91 – 101.

Bosnjak, Michael/ Tuten, Tracy L. (2003): Prepaid and Promised Incentives in Web Surveys. IN: Social Science Computer Review 21, S. 208 – 217.

Boulianne, Shelley (2008): Incentives. IN: Lavrakas, Paul J. (Hg.): Encyclopedia of Survey Research Methods. Los Angeles/ London/ New Delhi/ Singapore/ Washington DC: SAGE Publication, S. 328 – 331.

Brehm, Jack W. (1966): A Theory of Psychological Reactance. New York/ London: Academic Press.

Bühner, Markus (2011): Einführung in die Test- und Fragebogenkonstruktion. München/ Harlow/ Amsterdam/ Madrid/ Boston/ San Francisco/ Don Mills/ Mexico City/ Sydney: Pearson.

Cameron, Judy (2001): Negative Effects of Reward on Intrinsic Motivation - A Limited Phenomenon: Comment on Deci, Koestner and Ryan (2001). IN: Review of Educational Research 71, S. 29 – 42.

Cameron, Judy/ Banko, Katherine M./ Pierce, W. David (2001): Pervasive negative effects of rewards on intrinsic motivation: The myth continues. IN: The Behavior Analyst 24, S. 1 – 44.

Cameron, Judy/ Pierce, W. David (1994): Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis. IN: Review of Educational Research 64, S. 363 – 423.

Cameron, Judy/ Pierce, W. David (2002): Rewards and Intrinsic Motivation. Westport (Connecticut)/ London: Bergin & Garvey.

Church, Allan H. (1993): Estimating the Effect of Incentives on Mail Survey Response Rates. A Meta-Analysis. IN: Public Opinion Quarterly 57, S. 62 – 79.

Coleman, James S. (1991): Grundlagen der Sozialtheorie: Handlungen und Handlungssysteme. München: Oldenbourg Verlag.

Conrad, Frederick G./ Brown, Norman R./ Cashman, Erin R. (1998): Strategies for estimating behavioral frequency in survey interviews. IN: Memory 6, S. 339 – 366.

Converse, Philip E. (1964): The Nature Of Belief Systems in Mass Publics. IN: Apter, David E. (Hg.): Ideology and Discontent. New York/ Toronto/ London: The Free Press of Glencoe, S. 206 – 261.

Cronbach, L. J. (1942): Studies of Acquiescence as a Factor in the True-False Test. IN: Journal of Educational Psychology 33, S. 401 – 415.

Cronbach, Lee J. (1946): Response Sets and Test Validity. IN: Educational and Psychological Measurement 6, S. 475 – 494.

Davern, Michael/ Rockwood, Todd H./ Sherrod, Randy/ Campbell, Stephen (2003): Prepaid Monetary Incentives and Data Quality in Face-to-Face Interviews. Data from the 1996 Survey of Income and Program Participation Incentive Experiment. IN: Public Opinion Quarterly 67, S. 139 – 147.

Deci, Edward L. (1971): Effects of Externally Mediated Rewards on Intrinsic Motivation. IN: Journal of Personality and Social Psychology 18, S. 105 – 115.

Deci, Edward L. (1975): Intrinsic Motivation. New York/ London: Plenum Press.

Deci, Edward L./ Ryan, Richard M. (1985): Intrinsic Motivation and Self-Determination in Human Behavior. New York/ London: Plenum Press.

Deci, Edward L./ Ryan, Richard M. (1993): Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. IN: Zeitschrift für Pädagogik 39, S. 223 – 238.

Deci, Edward L./Koestner, Richard/ Ryan, Richard M. (2001a): Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. IN: Review of Educational Research 71, S. 1 – 27.

Deci, Edward L./Koestner, Richard/ Ryan, Richard M. (2001b): The Pervasive Negative Effects of Rewards on Intrinsic Motivation: Response to Cameron. IN: Review of Educational Research 71, S. 43 – 51.

Deci, Edward L./Koestner, Richard/ Ryan, Richard M. (1999): A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. IN: Psychological Bulletin 125, S. 627 – 668.

DeCharms, R. (1968): Personal Causation: The internal affective determinants of behavior. New York: Academic Press.

Deming, W. Edwards (1944): On Errors in Surveys. IN: American Sociological Review 9, S. 359 – 369.

Deutscher Ethikrat (2012): Intersexualität. Stellungnahme. Erreichbar unter: <http://www.ethikrat.org/dateien/pdf/stellungnahme-intersexualitaet.pdf> [letzter Zugriff: 30.09.2015]

Deutsche Forschungsgemeinschaft (1999): Qualitätskriterien der Umfrageforschung. Berlin: Akademie Verlag. (Herausgegeben von Kaase, Max)

Diekmann, Andreas (2008): Soziologie und Ökonomie: Der Beitrag experimenteller Wirtschaftsforschung zur Sozialtheorie. IN: Kölner Zeitschrift für Soziologie und Sozialpsychologie 60, S. 528 – 550.

Diekmann, Andreas/ Jann, Ben (2001): Anreizformen und Ausschöpfungsquoten bei postalischen Befragungen. Eine Prüfung der Reziprozitätshypothese. IN: ZUMA-Nachrichten 48, S. 18 – 27.

Diekmann, Andreas/ Voss, Thomas (2003): Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme. München: Oldenbourg.

Dillman, Don A./ Smyth, Jolene D./ Christian, Leah Melani (2009): Internet, Mail and Mixed-Mode Surveys. The Tailored Design Method. Hoboken (New Jersey): Wiley & Sons.

Dohmen, Thomas/ Falk, Armin/ Huffman, David/ Sunde, Uwe (2009): Homo Reciprocans: Survey Evidence on Behavioural Outcomes. IN: The Economic Journal 119, S. 592 – 612.

Dykema, Jennifer/ Stevenson, John/ Day, Brendon/ Sellers, Sherrill L./ Bonham, Vence L. (2011): Effects of Incentives and Prenotification on Response Rates and Costs in a National Web Survey of Physicians. IN: Evaluation & the Health Professions 34, S. 434 – 447.

Ekeh, Peter P. (1974): Social Exchange Theory: The two Traditions. London: Heinemann.

Esser, Hartmut (1986): Können Befragte lügen? Zum Konzept des „wahren Wertes“ im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. IN: Kölner Zeitschrift für Soziologie und Sozialpsychologie 38, S. 314 – 336.

Esser, Hartmut (1991): Die Erklärung systematischer Fehler in Interviews: Befragtenverhalten als "rational choice". IN: Wittenberg, Reinhard (Hg.): Person-Situation-Institution-Kultur. Günther Büschges zum 65. Geburtstag. Berlin: Duncker & Humblot, S. 59 – 78.

Esser, Hartmut (1999): Soziologie. Spezielle Grundlagen. Band 1: Situationslogik und Handeln. Frankfurt/ New York: Campus Verlag.

Falk, Armin (2003): Homo Oeconomicus versus Homo Reciprocans: Ansätze für ein neues Wirtschaftspolitisches Leitbild? IN: Perspektiven der Wirtschaftspolitik 4, S. 141 – 172.

Falk, Armin/ Fischbacher, Urs (2000): A Theory of Reciprocity. IN: Working Paper 6. Verfügbar unter: www.unizh.ch/iew/wp

Faulbaum, Frank/ Prüfer, Peter/ Rexroth, Margit (2009): Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. Wiesbaden: VS Verlag.

Festinger, Leon (1957): A Theory of Cognitive Dissonance. Stanford (CA): Stanford University Press.

Fick, Patrick/ Diehl, Claudia (2013): Incentivierungsstrategien bei Minderheitenangehörigen. IN: methoden, daten, analysen. Zeitschrift für empirische Sozialforschung, Heft 1, S. 59 – 88.

Godwin, Kenneth R. (1979): The Consequences of Large Monetary Incentives in Mail Surveys of Elites. IN: Public Opinion Quarterly 43, S. 378 – 387.

Gouldner, Alvin W. (1960): The Norm of Reciprocity: A Preliminary Statement. IN: American Sociological Review 25, S. 161 – 178.

Göritz, Anja S. (2004): The impact of material incentives on response quantity, response quality, sample composition, survey outcome and cost in online access panels. IN: International Journal of Market Research 46, S. 327 – 345.

Grant, Ruth W. (2012): Strings Attached. Untangling the Ethics of Incentives. Princeton/ Oxford: Princeton University Press.

Groves, Robert M./ Fowler, Floyd J. Jr./ Couper, Mick P./ Lepkowski, James M./ Singer, Eleanor/ Tourangeau, Roger (2004): *Survey Methodology*. Hoboken (New Jersey): Wiley & Sons.

Groves, Robert M./ Lyberg, Lars (2010): Total Survey Error. Past, Present, and Future. IN: *Public Opinion Quarterly* 74, S. 849 – 879.

Groves, Robert M./ Singer, Eleanor/Corning, Amy (2000): Leverage-Saliency Theory of Survey Participation. Description and an Illustration. IN: *Public Opinion Quarterly* 64, S. 399 – 408.

Guay, Frédéric/ Vallerand, Robert J./ Blanchard, Céline (2000): On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale. IN: *Motion and Emotion* 24, S. 175 – 213.

Häder, Michael/ Kühne, Mike (2010): Mobiltelefonerfahrung und Antwortqualität bei Umfragen. IN: *methoden, daten, analysen. Zeitschrift für empirische Sozialforschung* 4 (2), S. 105 – 125.

Hansen, Robert A. (1980): A Self-Perception Interpretation of the Effect of Monetary and Nonmonetary Incentives on Mail Survey Respondent Behavior. IN: *Journal of Marketing Research* 17, S. 77 – 83.

Hartmann, Petra (1991): *Wunsch und Wirklichkeit: Theorie und Empirie Sozialer Erwünschtheit*. Wiesbaden: Deutscher Universitäts-Verlag.

Holbrook, Allyson (2008a): Acquiescence. IN: Lavrakas, Paul J. (Hg.): *Encyclopedia of Survey Research Methods*. Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE, S. 3 – 4.

Holbrook, Allyson (2008b): Recency Effect. IN: Lavrakas, Paul J. (Hg.): *Encyclopedia of Survey Research Methods*. Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE, S. 695 – 696.

Holbrook, Allyson L./ Anand, Sowmya/ Johnson, Timothy P./ Cho, Young Ik/ Shavitt, Sharon/ Chávez, Noel/ Weiner, Saul (2014): Response Heaping in Interviewer-Administered Surveys: Is It Really a Form of Satisficing? IN: Public Opinion Quarterly 78, S. 591 – 633.

Hubbard, Raymond/ Little, Eldon L. (1988): Promised contributions to charity and mail survey responses. Replication with extension. IN: Public Opinion Quarterly 52, S. 223 – 230.

Jabine, Thomas B./ Straf, Miron L./ Tanur, Judith M./ Tourangeau, Roger (1984): Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. Washington D.C.: National Academy Press.

James, Jeannine M./ Bolstein, Richard (1990): The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys. IN: Public Opinion Quarterly 54, S. 346 – 361.

James, Jeannine M./ Bolstein, Richard (1992): Large Monetary Incentives and their Effects on Mail Survey Response Rates. IN: Public Opinion Quarterly 56, S. 442 – 453.

Költringer, Richard (1993): Gültigkeit von Umfragedaten. Wien/ Köln/ Weimar: Böhlau Verlag.

Krenzke, Thomas/ Mohadjer, Leyla/ Ritter, Grant/ Gadzuk, Anita (2005): Incentive effects on self-report of drug use and other measures of response quality in the alcohol and drug services study. IN: Journal of Economic and Social Measurement 30, S. 191 – 217.

Kreuter, Frauke/ McCulloch, Susan/ Presser, Stanley/ Tourangeau, Roger (2011): The Effects of Asking Filter Questions in Interleaved Versus Grouped Formats. IN: Sociological Methods and Research 40, S. 88 – 104.

Krosnick, Jon A. (1991): Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. IN: Applied Cognitive Psychology 5, S. 213 – 236.

Krosnick, Jon A. (2000): The Threat of Satisficing in Surveys: The Shortcuts Respondents Take in Answering Questions. IN: Survey Methods Newsletter 20, S. 4 – 8.

Krosnick, Jon A./Alwin, Duane F. (1987): An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. IN: Public Opinion Quarterly 51, S. 201 – 219.

Krosnick, Jon A./ Narayan, Sowmya S./ Smith, Wendy R. (1996): Satisficing in surveys: Initial evidence. IN: Braverman, Marc T./ Slater, Jana Kay (Hg.): Advances in survey research. San Francisco: Jossey-Bass, S. 29 – 44.

Kühnel, Steffen M./ Dingelstedt, André (2014): Kausalität. IN: Baur, Nina/ Blasius, Jörg (Hg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, S. 1017 – 1028.

Laurie, Heather/ Lynn, Peter (2008): The Use of Respondent Incentives on Longitudinal Surveys. IN: Working Paper Series 42, Institute For Social & Economic Research.

Lenski, Gerhard E./ Legett, John C. (1960): Caste, Class, and Deference in the Research Interview. IN: American Journal of Sociology 65, S. 463 – 467.

Lischewski, Julia (2015): Soziale Erwünschtheit im Licht des Rational-Choice Ansatzes. E-Dissertation. Göttingen: <http://hdl.handle.net/11858/00-1735-0000-0022-5DC6-A>

Mauss, Marcel (1923): Die Gabe. Die Form und Funktion des Austauschs in archaischen Gesellschaften. Reprint: 1990. Frankfurt (Main): Suhrkamp

Mayerl, Jochen/ Urban, Dieter (2008): Antwortreaktionszeiten in Survey Analysen. Messung, Auswertung und Anwendungen. Wiesbaden VS-Verlag.

McDaniel, Stephen W./ Rao C. P. (1980): The Effect of Monetary Inducement on Mailed Questionnaire Response Quality. IN: Journal of Marketing Research 17, S. 265 – 268.

McPhee, Cameron/ Hastedt, Sarah (2012): More Money? The Impact of Larger Incentives on Response Rates in a Two-Phase Mail Survey. Link:https://fcsm.sites.usa.gov/files/2014/05/Hastedt_2012FCSM_I-A.pdf [letzter Zugriff: 14.11.2014]

Medway, Rebecca L. (2012): Beyond Response Rates: The Effect of Prepaid Incentives on Measurement Error. Link: http://drum.lib.umd.edu/bitstream/1903/13646/1/Medway_umd_0117E_13833.pdf [letzter Zugriff am 14.11.2014]

Medway, Rebecca L./ Tourangeau, Roger (2015): Response Quality in Telephone Surveys. Do Prepaid Cash Incentives Make a Difference? IN: *Public Opinion Quarterly* 79, S. 524 – 543.

Mikrozensusgesetz (2005): Gesetz zur Durchführung einer Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt sowie die Wohnsituation der Haushalte (Mikrozensusgesetz 2005 – MZG 2005). Erreichbar unter: https://www.destatis.de/DE/Methoden/Rechtsgrundlagen/Statistikbereiche/Inhalte/054a_MZG_2005.pdf?__blob=publicationFile [letzter Zugriff: 30.09.2015]

Moosbrugger, Helfried/ Kelava, Augustin (2012): *Testtheorie und Fragebogenkonstruktion*. Berlin/ Heidelberg: Springer-Verlag.

Pallak, Suzanne R./ Costomiris, Steven/ Sroka, Susan/ Pittman, Thane S. (1982): School Experience, Reward Characteristics, and Intrinsic Motivation. IN: *Child Development* 53, S. 1382 – 1391.

Paulhus, Delroy L. (1984): Two-component models of socially desirable responding. *Journal of Personality and Social Psychology* 46, S. 598 – 609.

Perugini, Marco/ Gallucci, Marcello/ Presaghi, Fabio/ Ercolani, Anna Paola (2003): The Personal Norm of Reciprocity. IN: *European Journal of Personality* 17, S. 251 – 283.

Pffor, Klaus (2015): Incentives. *SDM Survey Guidelines (Gesis)*.

Porter, Stephen R./ Whitcomb, Michael E. (2003): The Impact of Lottery Incentives on Student Survey Response Rates. IN: *Research in Higher Education* 44, S. 389 – 407.

Pretty, Grace H./ Seligman, Clive (1984): Affect and the Overjustification Effect. IN: *Journal of Personality and Social Psychology* 46, S. 1241 – 1253.

Rammstedt, Beatrice/ Kemper, Christoph J./ Klein, Mira Céline/ Beierlein, Constanze/ Kovalova, Anastassiya (2013): Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit. IN: *methoden, daten, analysen. Zeitschrift für empirische Sozialforschung. Heft 2*, S. 233 – 249.

Rässler, Susanne/ Riphahn, Regina T. (2006): Survey item nonresponse and its treatment. IN: *Allgemeines Statistisches Archiv* 90, S. 217 – 232.

Regan, Dennis T. (1971): Effects of an Favor and Liking on Compliance. IN: *Journal of Experimental Social Psychology* 7, S. 627 – 639.

Robertson, Dan H./Bellenger, Danny N. (1978): A New Method of Increasing Mail Survey Responses: Contributions to Charity. IN: *Journal of Marketing Research* 15, S. 632 – 633.

Rosenthal, Gabriele (1999): *Der Holocaust im Leben von drei Generationen. Familien von Überlebenden der Shoah und von Nazi-Tätern.* Gießen: Psychosozial-Verlag.

Rubin, Donald B. (1976): Inference and Missing Data. IN: *Biometrika* 63, S. 581 – 592.

Rummel, Amy/ Feinberg, Richard (1988): Cognitive Evaluation Theory: A meta-analytic review of the literature. IN: *Social Behavior and Personality* 16, S. 147 – 164.

Ryan, Richard M./ Deci, Edward L. (2000a): Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. IN: *Contemporary Educational Psychology* 25, S. 54 – 67.

Ryan, Richard M./ Deci, Edward L. (2000b): Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. IN: American Psychologist 55, S. 68 – 78.

Ryu, Erica/ Couper, Mick P./ Marans, Robert W. (2006): Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error. IN: International Journal of Public Opinion Research 18, S. 89 – 106.

Sahlins, Marshall D. (2005): Zur Soziologie des primitiven Tauschs. IN: Adloff, Frank / Mau, Steffen (Hg.): Vom Geben und Nehmen. Zur Soziologie der Reziprozität. Campus Verlag: Frankfurt am Main, S. 73 – 94.

Saris, Willem E./ Gallhofer, Irmtraud N. (2007): Design, Evaluation, and Analysis of Questionnaires for Survey Research. Erste Auflage. Hoboken (New Jersey): Wiley & Sons.

Saris, Willem E./ Gallhofer, Irmtraud N. (2014): Design, Evaluation, and Analysis of Questionnaires for Survey Research. Zweite Auflage. Hoboken (New Jersey): Wiley & Sons.

Schimpl-Neimanns, Bernhard (2006): Zur Datenqualität der Bildungsangaben im Mikrozensus. IN: ZUMA-Arbeitsbericht 2006/03.

Schnell, Rainer (2012): Survey-Interviews. Methoden standardisierter Befragungen. Wiesbaden: VS-Verlag.

Shadish, William R./ Cook, Thomas D./ Campbell, Donald T. (2002): Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston/ New York: Houghton Mifflin Company.

Shettle, Caroly/ Mooney, Geraldine (1999): Monetary Incentives in U.S. Government Surveys. IN: Journal of Official Statistics 15, S. 231 – 250.

Silber, Henning (2015): Kontexteffekte in der Umfrageforschung: Interaktion zwischen Erhebungsdesign und Antwortverhalten. Selbstverlag.

Silber, Henning/ Lischewski, Julia/ Leibold, Jürgen (2013): Comparing Different Types of Web Surveys: Examining Drop-Outs, Non-Response and Social Desirability. IN: metodoloskizveski. Advances in Methodology and Statistics 10, S. 121 – 143.

Simon, Herbert A. (1947): Administrative Behavior. New York: The Macmillan Company.

Simon, Herbert A. (1957): Models of Man, Social and Rational. New York: Wiley.

Simmons, Eleanor/ Wilmot, Amanda (2004): Incentive Payments on Social Surveys: a Literature Review. IN: Survey Methodology Bulletin 53, S. 1 – 11.

Singer, Eleanor (1998): Incentives for Survey Participation: Research on Intended and Unintended Consequences. IN: ZUMA-Nachrichten 42, S. 7 – 29.

Singer, Eleanor/ Couper, Mick P. (2008): Do Incentives Exert Undue Influence on Survey Participation? Experimental Evidence. IN: Journal of Empirical Research on Human Research Ethics 3, S. 49 – 56.

Singer, Eleanor/ Gebler, Nancy/ Van Hoewyk, John (1997): Does \$10 equals \$10? The Effect of Framing on the Impact of Incentives. IN: Proceedings of the Survey Research Methods Section, American Statistical Association, S. 946 – 951.

Singer, Eleanor/ Hoewyk van, John/ Gebler, Nancy/ Raghunathan, Trivellore/ McGonagle, Katherine (1999): The Effect of Incentives on Response Rates in Interviewer-Mediated-Surveys. IN: Journal of Official Statistics 15, S. 217 – 230.

Singer, Eleanor/ Hoewyk van, John/ Maher, Mary P. (2000): Experiments with Incentives in Telephone Surveys. IN: Public Opinion Quarterly 64, S. 171 – 188.

Skarbak-Kozietulska, Anna/ Preisendörfer, Peter/ Wolter, Felix (2012): Leugnen oder Gestehen? Bestimmungsfaktoren wahrer Antworten in Befragungen. IN: Zeitschrift für Soziologie 41, S. 5 – 23.

Stadtmüller, Sven (2009): Rücklauf gut, alles gut? Zu erwünschten und unerwünschten Effekten monetärer Anreize bei postalischen Befragungen. IN: methoden, daten, analysen. Zeitschrift für empirische Sozialforschung 3 (2), S. 167 – 185.

Stadtmüller, Sven/ Porst, Rolf (2005): Zum Einsatz von Incentives bei postalischen Befragungen. ZUMA How-to-Reihe 14.

Staszynska, Katarzyna M. (2011): Cognitive Determinants of Data Quality in Public Opinion Polls: Respondents Definition of the Survey. IN: Polish Sociological Review 176, S. 493 – 514.

Stegbauer, Christian (2011): Reziprozität. Einführung in soziale Formen der Gegenseitigkeit. Wiesbaden: VS Verlag.

Stocké, Volker (2002): Framing und Rationalität. Die Bedeutung der Informationsdarstellung für das Entscheidungsverhalten. München: Oldenbourg Wissenschaftsverlag.

Stocké, Volker (2004): Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. IN: Zeitschrift für Soziologie 33, S. 303 – 320.

Tang, Shu-Hua/ Hall, Vernon C. (1995): The overjustification effect: A meta-analysis. IN: Applied Cognitive Psychology 9, S. 365 – 404.

The Council of Professional Associations on Federal Statistics (1993): Providing Incentives to Survey Respondents. Erreichbar unter: www.copafs.org/reports/providing_incentives_to_surveys_respondents.aspx#defining [letzter Zugriff: 20.07.2015]

Thierry, Henk, (1998): Motivation and Satisfaction. IN: Drenth, Pieter Johan Diederik/ Thierry, Henk/ de Wolff, Charles Johannes (Hg.): A Handbook of Work and Organizational Psychology: Volume 4: Organizational Psychology; Hove East Sussex: Psychology Press, S. 253 – 320.

Thomas, William I./ Thomas, Dorothy S. (1928): The Child in America. Behavior Problems and Programs. New York: Alfred A. Knopf.

Tourangeau, Roger (1984): Cognitive science and survey methods. IN: Jabine, Thomas B./ Straf, Miron L./ Tanur, Judith M./ Tourangeau, Roger (1984): Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. Washington D.C.: National Academy Press, S. 73 – 100.

Tourangeau, Roger (2007): Incentives, Falling Response Rates, and the Respondent-Researcher Relationship. Tagungsbeitrag: Ninth Conference on Health Survey Research Methods, S. 244 – 253.

Tourangeau, Roger/ Groves, Robert M./ Redline, Cleo D. (2010): Sensitive Topics and Reluctant Respondents. Demonstrating a Link between Nonresponse Bias and Measurement Error. IN: Public Opinion Quarterly 74, S. 413 – 432.

Tourangeau, Roger/ Rasinski, Kenneth (1988): Cognitive Processes Underlying Context Effects in Attitude Measurement. IN: Psychological Bulletin 103, S. 299 – 314.

Tourangeau, Roger/ Rips, Lance J./ Rasinski, Kenneth (2000): The Psychology of Survey Response. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo, Delhi: Cambridge University Press.

Tourangeau, Roger/ Ye, Cong (2009): The Framing of the Survey Request and Panel Attrition. IN: Public Opinion Quarterly 73, S. 338 – 348.

Yu, Julie/ Cooper, Harris (1983): A Quantitative Review of Research Design Effects on Response Rates to Questionnaires. IN: Journal of Marketing Research 20, S. 36 – 44.

Vogt, Katrin (2004): Interessenerzeugung durch individuelle Belohnung oder Übung zur Verhinderung von social loafing in Kooperationssituationen. E-Dissertation:
<http://hdl.handle.net/10900/48774>

Vosswinkel, Stephan (2005): Reziprozität und Anerkennung in Arbeitsbeziehungen. IN: Adloff, Frank / Mau, Steffen (Hg.): Vom Geben und Nehmen. Zur Soziologie der Reziprozität. Campus Verlag: Frankfurt am Main, S. 237 – 256.

Warriner, Keith/ Goyder, John/ Gjertsen, Heidi/ Hohner, Paula/ McSpurren, Kathleen (1996): Charities, no; Lotteries, no; Cash, yes. IN: Public Opinion Quarterly 60, S. 542 – 562.

Weisberg, Herbert F. (2005): The Total Survey Error Approach. A Guide To The New Science Of Survey Research. Chicago/ London: The University of Chicago Press.

Wiersma, (1992): The effects of extrinsic rewards in intrinsic motivation: A meta-analysis. IN: Journal of Occupational and Organizational Psychology 65, S. 101 – 114.

Williams, Dmitri/ Kennedy, Tracy L. M./ Moore, Robert J. (2011): Behind the Avatar: The Patterns, Practices and Functions of Role-Playing in MMOs. IN: Games and Culture 6, S. 171 – 200.

Willimack, Diane K./ Schuman, Howard/ Pennell, Beth-Ellen/ Lepkowksi, James M. (1995): Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey. IN: Public Opinion Quarterly 59, S. 78 – 92.

Wolbring, Tobias/ Groß, Jochen (2009): Reciprocity at the Football World Cup 2006. An Empirical Investigation of Passing Behavior. Arbeitspapier des Instituts für Soziologie: LMU München Nr.3. Erreichbar unter:

http://www.ls4.sozioologie.uni-muenchen.de/forschung/arbeitspapiere_lsbr/index.html

[letzter Zugriff: 30.09.2015]

Wolbring, Tobias/ Hellmann, Anja (2010): Attraktivität, Reziprozität und Lehrveranstaltungsevaluation. Eine experimentelle Untersuchung. IN: Kölner Zeitschrift für Soziologie und Sozialpsychologie 62, S. 707 – 730.

Wotruba, Thomas R. (1966): Monetary Inducements and Mail Questionnaire Research. IN: Journal of Marketing Research 3, S. 393 – 400.

Zaller, John/ Feldman, Stanley (1992): A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences. IN: American Journal of Political Science 36, S. 579 – 616.

Anhang

Anschreiben per eMail

Betreff: Forschungsprojekt: Persönlichkeit und Identität

Liebe Kommilitonin, lieber Kommilitone,

mein Name ist André Dingelstedt und im Rahmen meiner Promotionsarbeit führe ich eine Befragung zum Thema „Persönlichkeit und Identität“ unter den Studierenden der Universität Göttingen durch.

Das Ziel dieser Arbeit ist die Erhebung und Analyse des Studierverhaltens und der Persönlichkeitsmerkmale von Studierenden. Damit diese Arbeit gelingen kann, benötige ich Ihre Unterstützung, d.h. Ihre Teilnahme an der Befragung.

Ich möchte Sie daher bitten, an meiner Befragung teilzunehmen!

Die Befragung findet am Zentralcampus, im ersten Stock des Oeconomicums (Oec 1.123) statt. Die Teilnahme ist bis Mitte des nächsten Monats möglich, immer von Montag bis Freitag (15:00 bis 20:00 Uhr). Ein Termin oder eine Anmeldung ist nicht erforderlich – Sie brauchen einfach nur spontan vorbeizukommen! Die Befragung dauert nur etwa 27 Minuten und Sie würden mir damit wirklich sehr helfen.

Die Befragung ist natürlich vollkommen anonym und die erhobenen Daten werden selbstverständlich nur im Kontext der wissenschaftlichen Analysen genutzt.

Die zentralen Ergebnisse der Abschlussarbeit werden später, in Absprache mit Herrn Prof. Dr. Steffen M. Kühnel, auf der Seite des [Methodenzentrums](#) veröffentlicht. Die Ergebnisse werden dann mit dem Studiendekan besprochen und diskutiert.

Vielen Dank für das Lesen dieser eMail und ich hoffe bis demnächst ;)

Mit besten Grüßen

André Dingelstedt

PS: Eine Karte mit der genauen Lage / Adresse des Oeconomicums finden Sie unter diesem [Link!](#)

Anschreiben per eMail (Erinnerung)

Betreff: [Erinnerung] Forschungsprojekt: Persönlichkeit und Identität

Liebe Kommilitonin, lieber Kommilitone,

mein Name ist André Dingelstedt und ich möchte Sie hiermit noch einmal daran erinnern, dass ich im Rahmen meiner Abschlussarbeit noch freiwillige Studierende für meine Befragung zum Thema "Persönlichkeit und Identität" suche. Das Ziel dieser Arbeit ist die Erhebung und Analyse des Studierverhaltens und der Persönlichkeitsmerkmale von Studierenden.

Die Befragung findet noch an 9 Tagen statt und es ist für die Aussagekraft der Studie - und damit für meine Abschlussarbeit - sehr wichtig, dass eine ausreichende Stichprobengröße erreicht wird. Nur dadurch sind spätere statistische Verallgemeinerungen der Ergebnisse zulässig.

Ich möchte Sie daher bitten, an meiner Befragung teilzunehmen!

Der Ort der Befragung ist am Zentralcampus, im ersten Stock des Oeconomicums (Oec 1.123) und eine Teilnahme ist immer von Montag bis Freitag (15:00 bis 20:00 Uhr) möglich. Ein Termin oder eine Anmeldung ist nicht erforderlich – Sie brauchen einfach nur spontan vorbeizukommen! Die Befragung dauert nur etwa 27 Minuten und Sie würden mir damit wirklich sehr helfen.

Die Befragung ist hierbei natürlich vollkommen anonym und die erhobenen Daten werden selbstverständlich auch nur im Kontext der wissenschaftlichen Analysen genutzt.

Die zentralen Ergebnisse der Abschlussarbeit werden später auf der Homepage des [Methodenzentrums](#) veröffentlicht und darüber hinaus mit dem Studiendekan besprochen und diskutiert.

Bei den Studierenden, die bereits an dieser Umfrage teilgenommen haben, möchte ich mich auf diesem Wege auch noch einmal für die Unterstützung bedanken!

Mit freundlichen Grüßen

André Dingelstedt

PS: Eine Karte mit der genauen Lage / Adresse des Oeconomicums finden Sie unter diesem [Link!](#)

Flyer für die Rekrutierung

Forschungsprojekt: Persönlichkeit und Identität

Liebe Kommilitonin, lieber Kommilitone,

im Rahmen meiner Abschlussarbeit führe ich am Methodenzentrum Sozialwissenschaften eine Befragung zum Thema „Persönlichkeit und Identität“ unter den Studierenden der Universität Göttingen durch.

Das Ziel dieser Arbeit ist die Erhebung und Analyse des Studierverhaltens und der Persönlichkeitsmerkmale von Studierenden. Damit diese Arbeit gelingen kann, benötige ich Ihre Unterstützung, d.h. Ihre Teilnahme an der Befragung.

Ich möchte Sie daher bitten, an meiner Befragung teilzunehmen!

Die Befragung findet am Zentralcampus, im ersten Stock des **Oeconomicums (Oec 1.123)** statt. Die Teilnahme ist diese und nächste Woche immer von **Montag bis Freitag (15:00 bis 20:00 Uhr)**. Ab dem Haupteingang des Oeconomicums gibt es Wegweiser, um den Weg sicher zu finden!

Ein Termin oder eine Anmeldung ist nicht erforderlich – Sie brauchen einfach nur spontan vorbeizukommen! Die Befragung dauert nur etwa 27 Minuten und Sie würden mir damit wirklich sehr helfen.

Die zentralen Ergebnisse der Abschlussarbeit werden später auf der Seite des Methodenzentrums veröffentlicht. Die Ergebnisse werden dann mit dem Studiendekan besprochen und diskutiert.

André Dingelstedt

Die 13 Faktoren von Deming (1944)

1. Variability in response
2. Differences between different kinds and degrees of canvass;
 - a) Mail, telephone, telegraph, direct interview;
 - b) Intensive vs. Extensive interviews;
 - c) Long vs. short schedules;
 - d) Check block plan vs. Response;
 - e) Correspondence panel and key reporters;
3. Bias and variation arising from the interviewer;
4. Bias of the auspices;
5. Imperfections in the design of the questionnaire and tabulation plans;
 - a) Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
 - b) Omitting questions that would be illuminating to the interpretation of other questions;
 - c) Emotionally toned words; leading questions; limiting response to a pattern;
 - d) Failing to perceive what tabulations would be most significant;
 - e) Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available;
7. Bias arising from nonresponse (including omissions);
8. Bias arising from late reports;
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered;
10. Bias arising from an unrepresentative selection of respondents;
11. sampling errors and biases;
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating);
13. Errors in interpretation
 - a) Bias arising from bad curve fitting; wrong weighting; incorrect adjusting;
 - b) Misunderstanding the questionnaire;
failure to take account of the respondents' difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data;
 - c) Personal bias in interpretation.

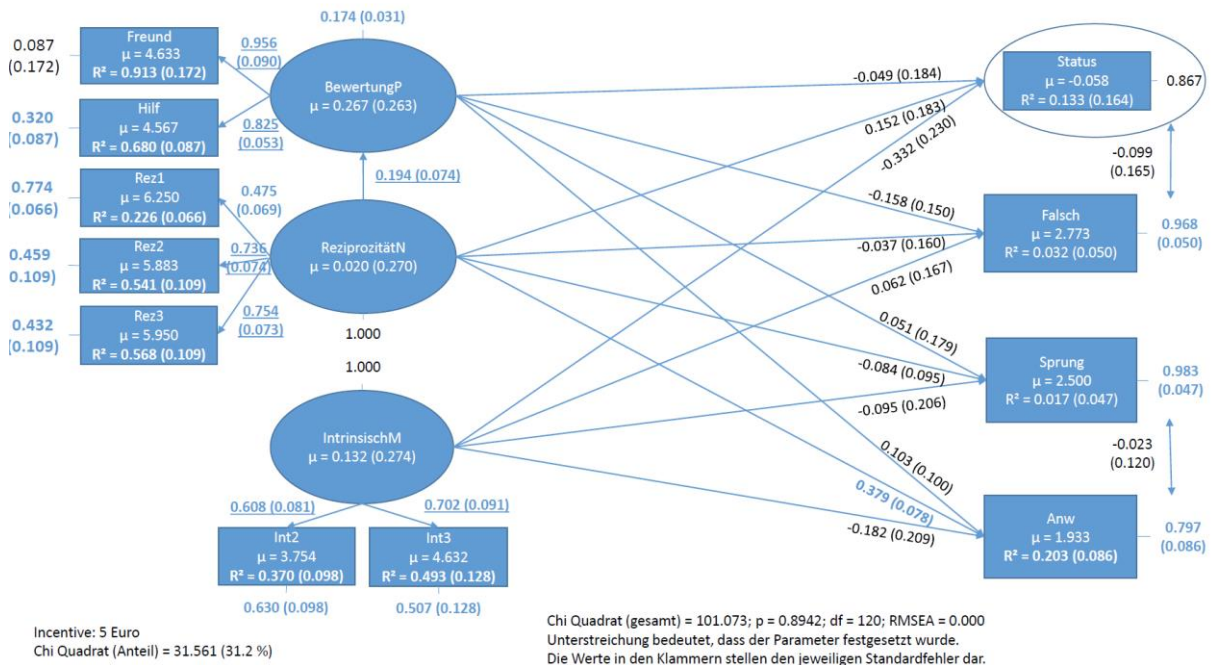
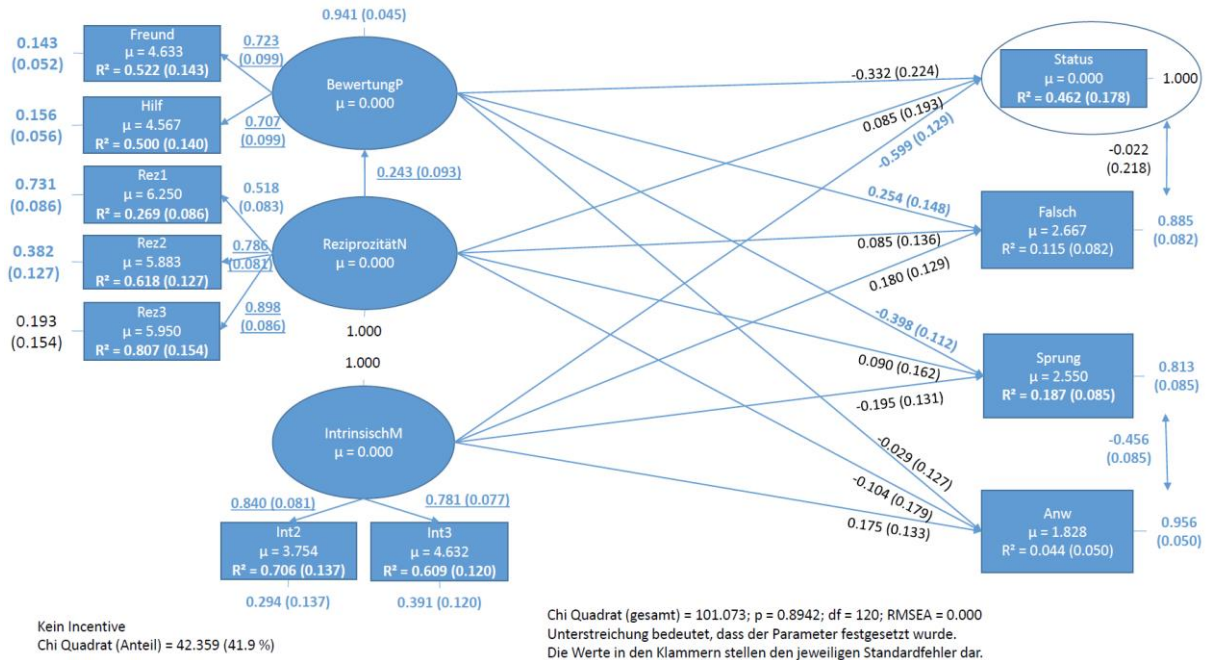
Quelle: Deming, W. Edwards (1944): On Errors in Surveys. IN: American Sociological Review 9, S. 359 – 369.

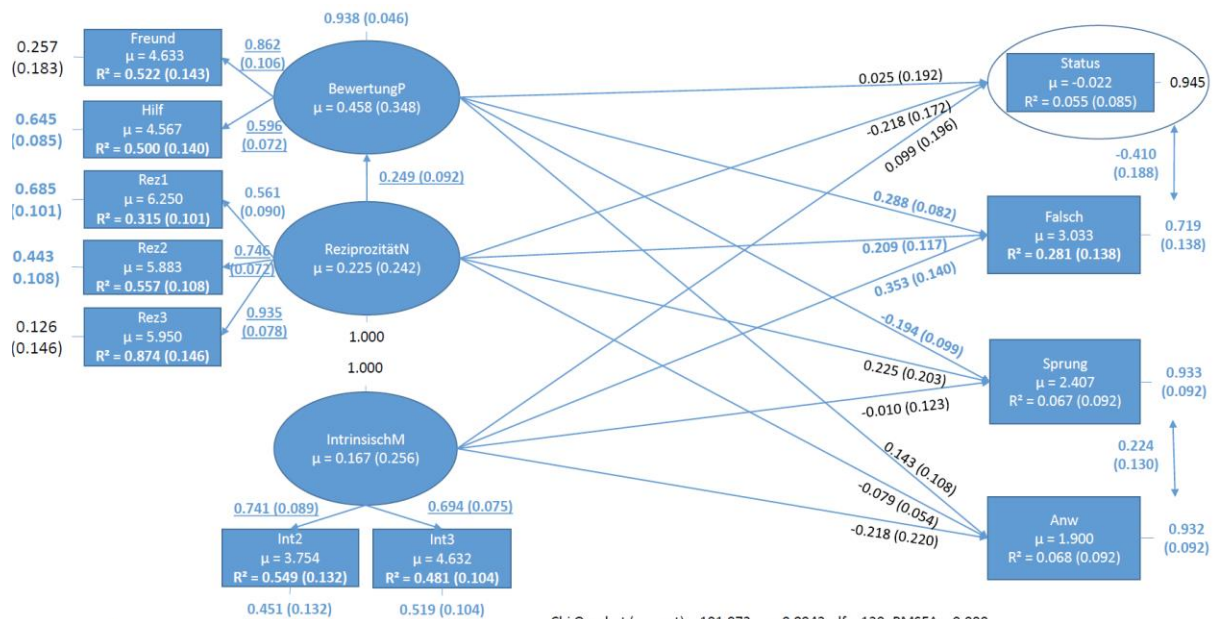
**Korrelationen zwischen den einzelnen Indikatoren für ein durchdachtes Bearbeiten
des Fragebogens.**

Status Quo-Effekt	1						
Akquieszenz	-0.114 (p = 0.129) n = 179	1					
Erinnerungs- strategie	0.010 (p = 0.892) n = 179	-0.048 (p = 0.520) n = 180	1				
Konsistenz: 24 h Tag	0.054 (p = 0.479) n = 177	-0.046 (p = 0.541) n = 178	-0.053 (p = 0.483) n = 178	1			
Konsistenz: 100% Studium	-0.017 (p = 0.821) n = 171	0.000 (p = 1.000) n = 171	0.083 (p = 0.281) n = 171	-0.075 (p = 0.335) n = 169	1		
Konsistenz: Politikerbewertung	0.140 (p = 0.175) n = 95	0.048 (p = 0.644) n = 96	-0.097 (p = 0.350) n = 96	-0.093 (p = 0.368) n = 96	-0.027 (p = 0.799) n = 92	1	
Anzahl an Worten	0.115 (p = 0.127) n = 178	-0.083 (p = 0.272) n = 179	0.089 (p = 0.238) n = 179	-0.007 (p = 0.931) n = 178	0.167 (p = 0.030) n = 170	0.138 (p = 0.179) n = 96	1
	Status Quo-Effekt	Akquieszenz	Erinnerungs- strategie	Konsistenz: 24 h Tag	Konsistenz: 100% Studium	Konsistenz: Politikerbewertung	Anzahl an Worten

Quelle: eigene Daten.

Die standardisierten Koeffizienten der simultanen Gruppenvergleiche für das Strukturmodell I

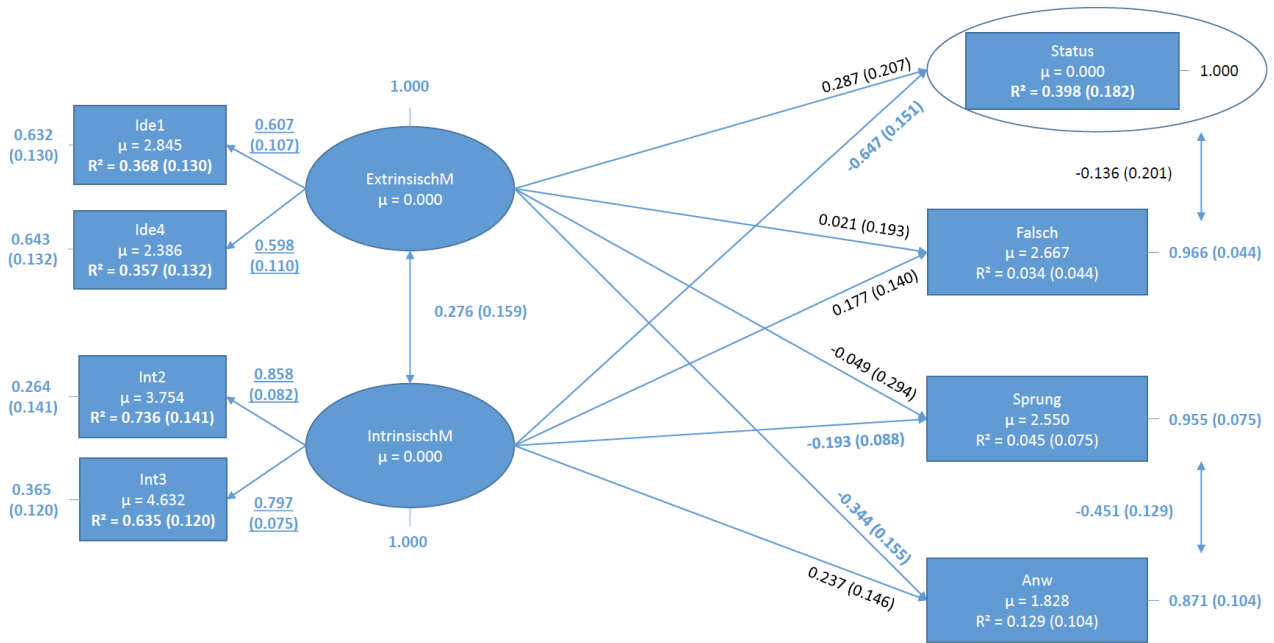




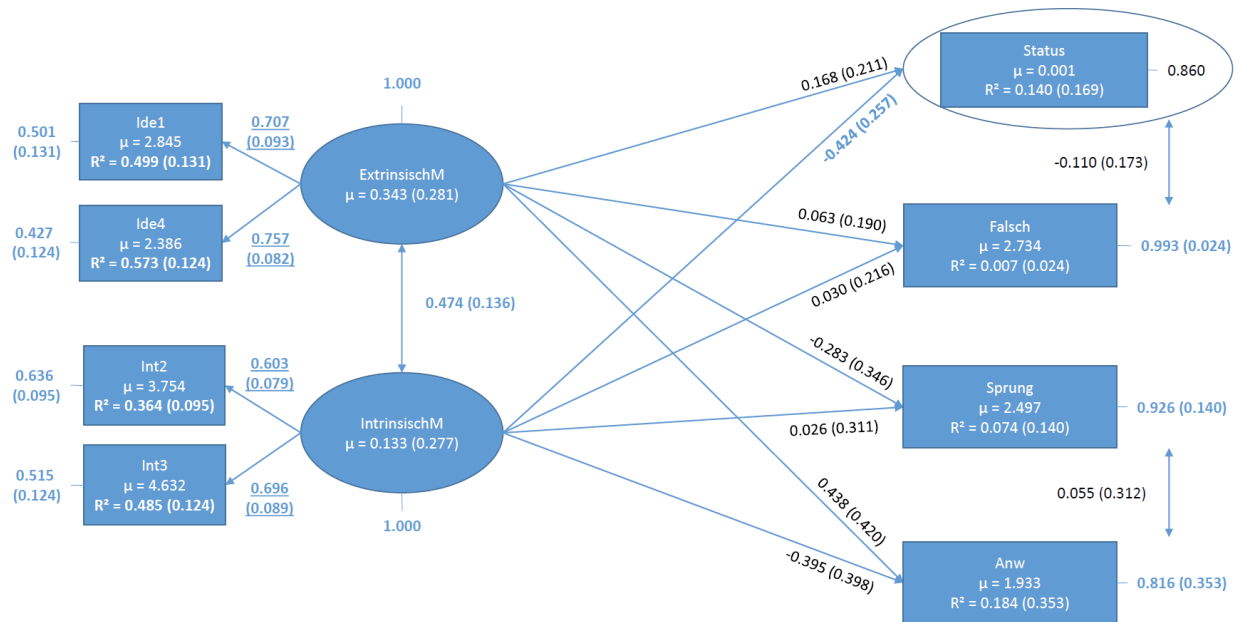
Incentive: 20 Euro
 Chi Quadrat (Anteil) = 27.153 (26.9 %)

Chi Quadrat (gesamt) = 101.073; p = 0.8942; df = 120; RMSEA = 0.000
 Unterstreichung bedeutet, dass der Parameter festgesetzt wurde.
 Die Werte in den Klammern stellen den jeweiligen Standardfehler dar.

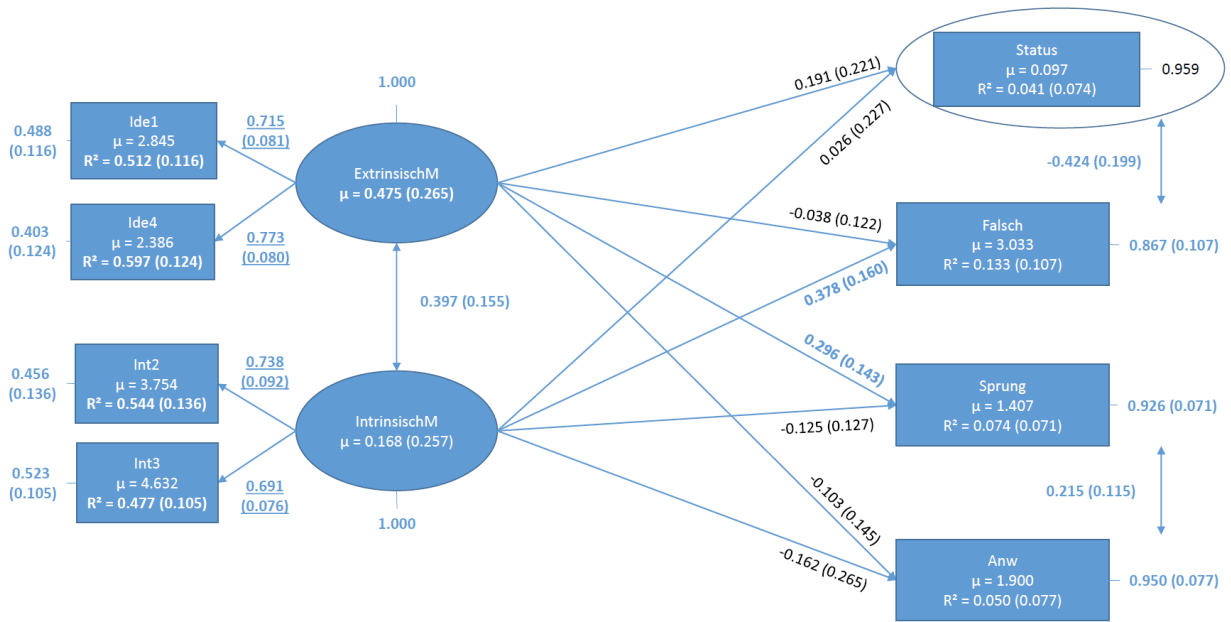
Die standardisierten Koeffizienten der simultanen Gruppenvergleiche für das Strukturmodell II



Kein Incentive
 Chi Quadrat (Anteil) = 17.143 (41.8 %)



Incentive: 5 Euro
 Chi Quadrat (Anteil) = 10.687 (26.1 %)



Incentive: 20 Euro
 Chi Quadrat (Anteil) = 13.165 (32.1 %)

Chi Quadrat (gesamt) = 40.995; p = 0.7850; df = 49; RMSEA = 0.000
 Unterstrichung bedeutet, dass der Parameter festgesetzt wurde.
 Die Werte in den Klammern stellen die jeweiligen Standardfehler dar.

Versicherung

„Ich versichere, dass ich die eingereichte Dissertation „Die Wirkung von Incentives auf die Antwortqualität in Umfragen“ selbstständig und ohne unerlaubte Hilfsmittel verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß den Schriften anderer Autoren entnommenen Stellen habe ich kenntlich gemacht.“

André Dingelstedt