Universität Koblenz-Landau

Campus Landau

Fachbereich 8: Psychologie

Einflüsse des zeitlichen Bezugsrahmens auf Angaben zur eigenen depressiven Befindlichkeit Teil 2

Diplomarbeit von

Tobias Fabian-Krause

Gutachter:

Prof. Dr. Manfred Schmitt

Dr. Christine Altstötter-Gleich

Juli 2011

Danksagung

An dieser Stelle möchte ich einigen Menschen, die mir bei der Erstellung dieser Arbeit mit fachlichem und/oder seelischem Beistand zur Seite standen meinen Dank aussprechen.

An erster Stelle zu nennen ist dabei der Betreuer dieser Arbeit, Herr Prof. Dr. Manfred Schmitt, der stets ein offenes Ohr für mich hatte und mir sowohl mit fachlichem Ratschlag als auch mit aufmunternden Worten zur Seite stand.

Danken möchte ich auch der Autorin der Vorgängerarbeit zu meiner Diplomarbeit, Frau Dipl.-Psych. Nina Heckmann, deren fachliche Ratschläge und Erfahrungsberichte ebenfalls in enormem Maße hilfreich waren.

Die seelische Unterstützung durch meine Freunde bedeutet mir viel. Es ist nicht möglich, alle zu nennen, die sich diesbezüglich verdient gemacht haben. Drei Personen möchte ich aber hervorheben und ihnen auf diesem Wege meinen herzlichen Dank aussprechen: Daniela Ferstl, Hady Nehm und Julia Nuber.

Schließlich möchte ich meinen Eltern und meinem Bruder Andreas danken. Was es mir bedeutet hat, jederzeit des Rückhalts und der Liebe meiner Familie sicher sein zu dürfen, ist mit Worten nicht auszudrücken. Ich belasse es daher bei einem schlichten *Dankeschön.*

Inhaltsverzeichnis

Zusammenfassung

Forschungsbefunde zum autobiographischen Gedächtnis (Sudman, Bradburn & Schwarz, 1996) lassen es zweifelhaft erscheinen, dass das von DSM-IV und ICD-10 vorgesehene Zweiwochenkriterium für die Diagnose einer Depression überhaupt im diagnostischen Prozess reliabel umsetzbar ist. Heckmann hatte 2008 festgestellt, dass es signifikante Unterschiede zwischen einer Untersuchungsgruppe gab, die das vereinfachte Beck-Depressions-Inventar (Schmitt & Maes, 2000) zweimal im Abstand von 14 Tagen mit der Instruktion, sich auf die letzten 14 Tage zu beziehen beantwortet hatte und einer zweiten Untersuchungsgruppe, die sich auf die letzten 3 Monate beziehen sollte. Mittels Latent-State-Trait-Analyse konnte dabei gezeigt werden, dass die Traitkonsistenz in der 3-Monats-Gruppe signifikant höher war als in der 14-Tage-Gruppe. Die Differenz hinsichtlich dieses Koeffizienten lag allerdings lediglich bei .05. Der durch stabile Eigenschaften erklärte Anteil lag mit .79 (14-Tage-Gruppe) und .84 in der 3-Monats-Gruppe sehr hoch. In der Folgestudie sollte einerseits geprüft werden, ob sich diese Auswirkungen der zeitlichen Instruktion replizieren lassen und andererseits, ob eine Erhöhung des Retest-Intervalls auf 3 Monate zu einer Absenkung des Anteils stabiler Dispositionen zugunsten des Situationseinflusses führt.

Mit Ausnahme der Erhöhung des Abstands zwischen den Messzeitpunkten wurde das Design von Heckmann (2008) übernommen. Die Untersuchung wurde an einer heterogenen Stichprobe durchgeführt. Die Stichprobengröße betrug N = 187 für die 3-Monatsgruppe und N = 240 für die 14-Tage-Gruppe. Da sich die untersuchten Gruppen hinsichtlich soziodemographischer Eigenschaften (Geschlecht, aktuelle berufliche Tätigkeit) unterschieden konnten Heckmanns Ergebnisse nicht repliziert werden, wobei keine klare Aussage darüber getroffen werden kann, ob die ermittelten Befunde insgesamt für oder gegen einen Einfluss der zeitlichen Instruktion sprechen. Die Erhöhung des Abstands zwischen den Messzeitpunkten führte zu einer Absenkung der Traitkonsistenz zugunsten der Zeitspezifität.

Schlüsselwörter: BDI-V, zeitliche Instruktion, Latent-State-Trait-Analysen

1 Einleitung

> The validity and reliability of measurements is of highest importance in clinical psychology because clinical judgments can have very important consequences for clients. Invalid measurements bear risks like over- or underestimation of treatment effects, they may lead to the wrong diagnosis, they may indicate a suboptimal treatment, or, in the worst case, they might even not detect a relevant symptom at all. (Courvoisier, Nussbeck, Eid, Geiser & Cole, 2008, S. 270).

Das Zitat verdeutlicht, wie wichtig eine valide und reliable Diagnostik auch und gerade für die klinische Psychologie ist. Der in der Praxis am meisten im Vordergrund stehende Aspekt ist dabei natürlich die Frage danach, ob bei einem vorstellig werdenden Klienten überhaupt eine behandlungsbedürftige Störung vorliegt und wenn ja, um welche es sich handelt. Dies ist Gegenstand der klassifikatorischen Diagnostik. Die aktuell für Praktiker verbindlichen Diagnosesysteme DSM-IV (zitiert nach der deutschen Übersetzung der textrevidierten Fassung (DSM-IV-TR) von Saß, Wittchen, Zaudig & Houben, 2003) und ICD-10 (zitiert nach der deutschen Ausgabe: Dilling, Mombour, Schmidt & Schulte-Markwort, 2006) geben strikte Abgrenzungskriterien zwischen den einzelnen Störungsbildern vor. Diese implizieren auch eine restriktive Trennung zwischen Personen, die eine psychische Störung haben und solchen, bei denen eine solche nicht vorliegt (Widiger & Clark, 2000). Diese Kriterien sind z.T. qualitativer Natur, d.h. sie bestehen aus definierten Symptomen, beide Diagnosesysteme enthalten aber auch quantitative Vorgaben wie z.B. Mindestanzahlen für die beschriebenen Symptome und Zeitkriterien, die vorgeben wie lange die einschlägigen Symptome mindestens vorliegen müssen bzw. wie lange sie höchstens vorliegen dürfen, damit die jeweilige Diagnose vergeben werden kann. So sehen DSM-IV und ICD-10 vor, das für die Major Depression (DSM) bzw. die mit F32 zu kodierenden Depressionsdiagnosen nur dann vergeben werden dürfen, wenn die einschlägigen Symptome mindestens zwei Wochen vorliegen (Saß et. al., 2003; Dilling et. al., 2006). Sofern klinisch bedeutsames Leiden bzw. signifikante

Funktionsbeeinträchtigung als nosologisch relevantes Kriterium für die Diagnose einer psychischen Störung angesehen wird, gibt es begründeten Anlass zum Zweifel bezüglich der Frage, ob dieses Kriterium tatsächlich ein brauchbares Abgrenzungskriterium zwischen „krank“ und „nicht krank“ darstellt (Howland et. al., 2008; Hautzinger & Meyer, 2002). Des Weiteren ist zu hinterfragen, ob die diagnostische Aufgabe der Feststellung, ob das Zweiwochenkriterium erfüllt ist im Rahmen des klinisch-diagnostischen Prozesses tatsächlich als reliabel und valide umsetzbar angesehen werden kann. Es gibt eine Reihe von Befunden, die Zweifel an der Reliabilität und Validität der Selbstauskünfte von Personen, die eine psychopathologisch relevante Symptomatik aufweisen rechtfertigen. So ist z.B. bekannt, dass Personen, die unter depressiven Erkrankungen leiden zu einer Übergeneralisierung neigen (z.B. Summner et. al., 2011). Obwohl direkte empirische Belege bislang fehlen erscheint es plausibel davon auszugehen, dass diese Tendenz dazu führt, dass der Zeitraum, während dessen die depressive Symptomatik vorliegt überschätzt wird. Dies hätte falsch-positive Diagnosen zur Folge. Neben diesen psychopathologisch begründeten Zweifeln stellt sich für Zeitkriterien die Frage danach, wie gut Menschen allgemein in der Lage sind, konkrete zeitliche Vorgaben bei der Befragung nach ihrem Befinden zu berücksichtigen. Der kognitive Aufwand, den eine solche Aufgabe mit sich bringt muss als erheblich angesehen werden. Nicht nur dass relevante Ereignisse für den erfragten Zeitraum erinnert werden müssen. Es ist darüber hinaus auch erforderlich, dass die erinnerten Einzelereignisse zu einem Gesamtbild integriert werden (Sudman, Bradburn & Schwarz, 1996). Es erscheint vor diesem Hintergrund zweifelhaft davon auszugehen, dass streng definierte Zeitkriterien wie das Zweiwochenkriterium für Depressionen dabei adäquat umgesetzt werden können. Alle genannten Aspekte lassen das Zweiwochenkritierum hinterfragenswert erscheinen und sprechen für eine dimensionale Konzeption depressiver Störungen.

Ein wichtiges Ziel therapeutischer Interventionen bei Depressionen besteht darin, eine Stabilisierung des affektiven Befindens des Klienten zu bewirken und vor allem negative Ausschläge seltener und weniger gravierend ausfallen zu lassen. Hieraus

ergibt sich ein weiterer wichtiger Aspekt der klinischen Diagnostik, nämlich die Frage danach, worauf sich aktuell feststellbare Veränderungen des affektiven Befindens eines Klienten zurückführen lassen: „Die verminderte Depressivität eines Psychotherapiepatienten kann eine völlig andere Bedeutung bekommen, wenn man situative Einflüsse wie etwa ein tagesspezifisches Lob durch den Arbeitgeber berücksichtigt, oder sie auf therapiebedingte, situationsübergreifende Traitveränderungen zurückführt.“ (Yousfi & Steyer, 2006, S. 351). Es ist also wünschenswert, für diagnostische Instrumente, wie beispielsweise Fragebögen, in Erfahrung zu bringen, ob diese in stärkerem Maße stabile Eigenschaften messen oder ob das Antwortverhalten in erster Linie von der Situation abhängt, in der sich die Person befindet, die den Fragebogen bearbeitet. Dazu ist es notwendig, Determinanten des Antwortverhaltens für den jeweiligen Fragebogen zu identifizieren. Zwei mögliche Determinanten für das Antwortverhalten in Depressions-Fragebögen stellen dabei die zeitliche Instruktion und der Retest-Abstand dar. Die zeitliche Instruktion betreffend wird in der psychologischen Forschung häufig angenommen, dass eine Instruktion wie „im Allgemeinen“ dazu führt, dass vornehmlich die Persönlichkeit des Befragten erfasst wird, wohingegen „im Moment“ eher zu Antworten führen sollte, die in dominierendem Maße durch die Situation beeinflusst wird (z.B. Krohne, Egloff, Kohlmann & Tausch, 1996). Außerdem ist davon auszugehen, dass es auch eine Rolle spielt, ob zwischen den Messzeitpunkten ein größerer oder kleinerer Abstand gewählt wird. Je größer der Retest-Abstand, desto höher ist die Wahrscheinlichkeit, dass sich die persönliche Situation der Person, die den Fragebogen bearbeitet verändert hat. Dementsprechend sollte die Situationsspezifität des gleichen Messinstruments bei gleicher zeitlicher Instruktion und einem größer gewählten Retest-Abstand höher sein.

Die vorliegende Arbeit verfolgt daher einerseits das Ziel zu prüfen, welchen Einfluss die zeitliche Instruktion auf das Antwortverhalten in einem Fragebogen zur Erfassung von Depressivität hat und andererseits zu prüfen, welchen Einfluss dabei der Abstand zwischen den Messzeitpunkten hat. Dies soll anhand des vereinfachten Beck-Depressions-Inventars (BDI-V) überprüft werden. Der Aspekt der Auswirkung der zeitlichen Instruktion wurde im Jahr 2008 für das gleiche Messinstrument von

Heckmann untersucht. Das Ziel der vorliegenden Studie ist es einerseits zu überprüfen, inwieweit die im Rahmen der Vorgängerarbeit gewonnenen Erkenntnisse hinsichtlich der zeitlichen Instruktion repliziert werden können. Andererseits soll geprüft werden, welche Auswirkung eine Vergrößerung des gewählten Retest-Abstands auf das Antwortverhalten hat. Sofern die oben ausgeführten Überlegungen zutreffend sind, sollte diese Variation zu einer Erhöhung des situativen Einflusses und korrespondierend zu einer Absenkung des Traiteinflusses führen.

Die Arbeit gliedert sich in die folgenden Abschnitte: Kapitel 2 dient der Darstellung des theoretischen Hintergrunds. Dabei werden depressive Störungen und Probleme der Klassifikation und Diagnostik dieser Erkrankungen dargestellt, bevor das Thema State- und Trait-Depressivität behandelt wird. Es folgen Betrachtungen zum autobiographischen Gedächtnis. Anschließend werden BDI-O und BDI-V vorgestellt. Danach erfolgt eine kurze Einführung in die Latent-State-Trait-Theorie. Das Kapitel endet mit der Darstellung für die Studie relevanter Forschungsergebnisse.

Kapitel 3 enthält die Herleitung der Fragestellung und die im Rahmen dieser Arbeit untersuchten Hypothesen.

Kapitel 4 beschreibt Durchführung und methodisches Vorgehen im Rahmen dieser Studie sowie die Beschreibung der Stichprobe. Die Ergebnisse der Studie werden in Kapitel 5 dargestellt und in Kapitel 6 diskutiert.

2 Theoretischer Hintergrund

Das zweite Kapitel dieser Arbeit widmet sich dem theoretischen Hintergrund der Untersuchung. Der erste Abschnitt bietet dabei eine Einführung in das Themengebiet „Depressive Störungen“ und untersucht ausgewählte Probleme der Klassifikation und Diagnostik dieser Erkrankungen (2.1.) Der zweite Abschnitt beschäftigt mit der Unterscheidung in State und Trait im Zusammenhang mit Depressivität (2.2). Im dritten Abschnitt werden Probleme des autobiographischen Gedächtnisses dargestellt, die hinsichtlich der Frage der Umsetzbarkeit zeitlicher Instruktionen von Interesse sind (2.3). Gegenstand des vierten Abschnitts sind BDI-O und BDI-V (2.4). Im vorletzten Abschnitt wird eine kurze Einführung in die Latent-State-Trait-Theorie gegeben (2.5), das Kapitel endet mit der Darstellung für die Studie relevanter Forschungsergebnisse.

2.1 Depressive Störungen und Depressionsdiagnostik

Dieser Abschnitt soll zunächst eine kurze Einführung in das Gebiet der depressiven Störungen bieten. Dazu soll zunächst die Relevanz des Problems verdeutlicht werden (Unterabschnitt 2.1.1). Danach werden die Diagnostischen Kriterien der Major Depression nach DSM-IV vorgestellt und überblicksartig mit denen der ICD-10 verglichen (2.1.2). Im Anschluss werden Differenzialdiagnostik und Komorbidität diskutiert (2.1.3 und 2.1.4). Es folgt eine kritische Auseinandersetzung mit der klassifikatorischen Konzeption depressiver Störungen, der die Alternative einer dimensionalen Konzeption gegenübergestellt wird (2.1.5). Der Abschnitt schließt mit einer Zusammenfassung des Dargestellten und stellt den Bezug zur Fragestellung der im Rahmen dieser Arbeit durchgeführten Studie her (2.1.6).

2.1.1 Die Relevanz des Problems

Wittchen & Jacobi (2005) schätzten auf Basis von 27 epidemiologischen Studien in 16 europäischen Ländern die Anzahl der Menschen, die im Gebiet der Europäischen Union im Untersuchungsjahr an Depressionen litten auf 18,4 Mio. Die Autoren

benennen depressive Störungen als die am häufigsten auftretenden psychischen Erkrankungen. Der Anteil der unbehandelten und unerkannten Fälle von Major Depression liegt nach ihrer Einschätzung bei rund zwei Dritteln. Zahlreiche Studien (z.B. Kessler et. al., 2003) konnten zeigen, dass Depressionen stark beeinträchtigende Erkrankungen darstellen, die nicht nur für die erkrankten Individuen, sondern auch für deren Angehörige als enorm belastend angesehen werden müssen. Des Weiteren hat die Erkrankung Auswirkungen auf die Mortalität der Betroffenen: Beesdo & Wittchen (2006) beziffern die Suizidrate unter depressiven Personen auf 15%. Die Unfall- und krankheitsbedingte Mortalität ist ebenfalls erhöht (de Jong-Meyer, 2005). All dies verdeutlicht die Relevanz des Problems.

2.1.2 Major Depression nach DSM-IV

Das aktuell gültige DSM-IV-TR (zitiert nach der deutschen Übersetzung von Saß, Wittchen, Zaudig & Houben, 2003) führt die depressiven Störungen im Kapitel „Affektive Störungen". Neben den depressiven Störungen umfasst dieses Kapitel die bipolaren Störungen sowie eine mit „Andere Affektive Störungen" bezeichnete Restkategorie, welche durch medizinische Krankheitsfaktoren ausgelöste, substanzinduzierte und nicht näher bezeichnete affektive Störungen enthält. Zu den depressiven Störungen nach DSM-IV-TR gehören neben der Major Depression die dysthyme Störung und die nicht näher bezeichnete depressive Störung. Das Kriterium A für die depressive Episode einer Major Depression benennt zunächst eine *Mindestsymptomzahl* sowie ein *Zeitkriterium* für die Vergabe der Diagnose: „Mindestens fünf der folgenden Symptome bestehen während derselben Zwei-Wochen-Periode…" (Saß et. al., 2003, S. 406). Im Anschluss daran werden die folgenden 9 Symptome benannt (Saß et. al., S. 406/407, leicht gekürzt wiedergegeben):

- Depressive Verstimmung an fast allen Tagen
- Deutlich vermindertes Interesse oder Freude an allen oder fast allen Aktivitäten
- Deutlicher Gewichtsverlust ohne Diät oder Gewichtszunahme
- Schlaflosigkeit oder vermehrter Schlaf an fast allen Tagen
- Psychomotorische Unruhe oder Verlangsamung an fast allen Tagen

- Müdigkeit oder Energieverlust an fast allen Tagen
- Gefühle von Wertlosigkeit oder übermäßige Schuldgefühle an fast allen Tagen
- Verminderte Fähigkeit zu denken oder sich zu konzentrieren
- Wiederkehrende Gedanken an den Tod

Das C-Kriterium besteht in der Verursachung von Leid in klinischem bedeutsamem Ausmaß bzw. in der Funktionsbeeinträchtigung für die betroffene Person. Die weiteren Kriterien sind differenzialdiagnostische Kriterien und werden, da das DSM-IV-TR in diesem Punkt etwas unübersichtlich ist, im folgenden Abschnitt zusammenfassend und ohne direkten Bezug zum DSM dargestellt. Erwähnenswert ist noch, dass das DSM-IV-TR eine Zusatzcodierung nach Schweregrad erlaubt, die aber voraussetzt, dass die genannten Mindestkriterien erfüllt sind und nur innerhalb dieser Kriterien differenziert wird. Beispielsweise liegt eine leichte Form der Major Depression dann vor, wenn die Mindestsymptomanzahl „gerade erreicht oder knapp überschritten wird." (Saß et. al., 2003, S. 462). Ein detaillierter Vergleich mit dem außerhalb der Vereinigten Staaten angewandten Klassifikationssystem der WHO (ICD) ist für die Fragestellung dieser Arbeit nicht zielführend und wird daher unterlassen. Interessant sind zwei Aspekte: erstens nimmt auch die aktuell gültige Version der ICD (ICD-10) für die mit F32 kodierten Diagnosen, welche das Analogon zur Major Depression im DSM-IV-TR darstellen, eine Unterteilung nach Schweregrad vor. Jedem Schweregrad wird dabei eine Mindestanzahl an Symptomen zugeordnet, die sich inhaltlich stark mit denen des DSM-IV-TR überlappen (Dilling, Mombour, Schmidt & Schulte-Markwort, 2006). Zweitens wird auch das Zweiwochenkriterium von der ICD-10 benannt und gilt für alle Schweregrade: „Die depressive Episode sollte mindestens zwei Wochen dauern." (Dilling et. al., 2006, S. 106)

2.1.3 Differenzialdiagnose

Hautzinger (2010) nennt die folgenden differenzialdiagnostisch relevanten Punkte, die bei der Verdachtsdiagnose Depression zu beachten sind:

- Depressive Stimmung wird nicht durch körperliche Erkrankungen wie z.B. eine Schilddrüsenunterfunktion oder durch Substanzabhängigkeit verursacht und ist nicht auf Medikamenteneinnahme zurückzuführen.
- Es liegt keine bipolare affektive Störung vor.
- Es liegt keine andere depressive affektive Störung wie eine Anpassungsstörung oder eine dysthyme Störung vor.
- Es handelt sich nicht um eine Trauerreaktion.

Zu ergänzen wäre an dieser Stelle das im DSM-IV-TR explizit benannte Abgrenzungskriterium: „es liegt keine schizoaffektive Störung vor und die depressive Störung überlagert keine Schizophrenie oder eine ähnliche Erkrankung“ (Saß et. al., 2003, S. 426). Unter den genannten soll exemplarisch die dysthyme Störung herausgegriffen und ihre diagnostischen Kriterien grob skizziert werden. Bei der Dysthymen Störung handelt es sich gemäß DSM-IV-TR um eine chronische depressive Verstimmung, die nach Kriterium A „für die meiste Zeit des Tages an mehr als der Hälfte der Tage besteht und über mindestens zwei Jahre hinweg andauert.“ (Saß et. al., 2003, S. 431). Die im B-Kriterium genannten Symptome entsprechen im Wesentlichen denen der Major Depression, wobei Freudlosigkeit, die psychomotorischen Symptome und Gedanken an den Tod nicht genannt werden und lediglich zwei Symptome als Mindestanzahl vorgegeben sind. Das C-Kriterium verlangt, dass es in der betreffenden Zweijahresperiode keinen Zeitraum von mehr als zwei Monaten gab, in denen der Patient symptomfrei war.

2.1.4 Komorbidität

Die Komorbiditätsrate der Depression ist hoch. Hautzinger (2010) beziffert diese auf 77% (S. 16). Als häufigste parallel auftretende Störungsgruppe benennt der Autor dabei die Angststörungen, die etwa die Hälfte der mit einer Depression

diagnostizierten Personen aufweisen, gefolgt von substanzinduzierten Abhängigkeiten und somatoformen Störungen (jeweils ungefähr ein Drittel).

2.1.5 Kategoriale versus dimensionale Konzeption

Die in den vorangegangenen Abschnitten dargestellte Konzeption depressiver Störungen des DSM-IV-TR entspricht einer klassifikatorischen oder kategorialen Sichtweise auf die Störungsbilder. Die Begriffe kategorial und klassifikatorisch werden heute häufig synonym verwendet (Michael & Margraf, 2003). Die zentrale Annahme der klassifikatorischen Diagnostik beschreiben die zitierten Autoren wie folgt: „Der kategorialen Klassifikation liegt die Annahme zugrunde, dass es sinnvolle Gruppierungen der beobachteten Phänomene gibt (z.B. überzufällig gemeinsames Auftreten bestimmter Symptome) und dass hinreichen qualitative Unterscheide zwischen den Gruppen bestehen, die die Einteilung in diskrete Klassen rechtfertigen." (Michael & Margraf, 2003, S. 238). Die Autoren diskutieren Vor- und Nachteile der kategorialen Klassifikation. Jeweils drei der Argumente sollen hier herausgegriffen werden (Michael & Margraf, 2003, S. 238):

Pro kategoriale Klassifikation:

- Erleichterung der Kommunikation durch eine klar definierte Nomenklatur
- Wirtschaftliche Informationsvermittlung, da von Diagnose auf Störungsmerkmale geschlossen werden kann
- Feststellen von überzufälligen Syndromen, d.h., bestimmte klinische Merkmale treten besonders häufig zusammen auf

Kontra kategoriale Klassifikation:

- Diagnostische Etiketten (Labels) fördern bzw. bewirken Stigmatisierung
- Künstliche Klassen erhalten einen unangemessenen Realitätsgehalt
- Klasse verdecken zugrunde liegende Dimensionen

Natürlich gibt es neben den genannten auch eine Reihe pragmatischer Argumente, die es für derzeit aktive Praktikerinnen und Praktiker unumgänglich machen, mindestens

auch die kategoriale Klassifikation zu berücksichtigen, wie z.B. das von Michael & Margraf in einem Extraabschnitt benannte Faktum, dass „[d]ie Krankenkassenabrechnung (…) das Vergeben einer klassifikatorischen Diagnose [erfordert].“ (S. 239). Gewichtige Nachteile ergeben sich aus den in Kapitel 2.1.2 exemplarisch für die Major Depression im DSM-IV-TR beschriebenen sehr rigiden Kriterien, welche die klassifikatorischen Diagnosesysteme benennen. Ein Problem erwächst dabei aus der Tatsache, dass DSM-IV-TR und ICD-10 eine strikte Trennung zwischen krank und nicht-krank vorsehen (Widiger & Clark, 2000). Angst & Merikangas (2001) kritisieren, dass die meisten strukturierten Interviews für die psychiatrische Diagnostik keine zusätzlichen Informationen über Patienten einholen, die das 14-Tage-Kriterium nicht erfüllen, selbst dann nicht, wenn diese alle Symptome einer Depression aufweisen. Hautzinger & Meyer (2002) bringen diesen Nachteil der kategorialen Klassifikation sehr anschaulich auf den Punkt:

> „Wie bei allen derartigen kategorialen Entscheidungen erhebt sich schnell die Frage, mit welchem Recht die Grenze z.B. bei fünf gleichzeitig vorhandenen Symptomen gezogen wird und nicht schon bei drei oder vier Symptomen. Oder ob verschiedene kurze heftige, doch niemals das 2-Wochen-Kriterium erfüllende Episoden affektiver Störungen möglicherweise nicht viel beeinträchtigender sind als eine einzelne längere Episode im Laufe eines Jahres.“ (S. 36).

Die Abstufung nach Schweregrad bei der Depressionsdiagnose nach ICD-10 stellt bis zu einem gewissen Grad einen Einstieg in eine dimensionale Konzeptionierung der Depression dar, wird doch das Skalenniveau von einer Nominalskala (Ja/Nein-Entscheidung bezüglich des Störungsbildes) zu einer Ordinalskala erhöht (Hautzinger, 2010). Doch wie dargestellt werden auch in der ICD-10 Mindestanzahlen für die Symptome genannt (siehe Kapitel 2.1.2) und das Zweiwochenkriterium gilt ebenfalls, so dass die beschriebenen Fragwürdigkeiten auch für das WHO-Diagnosesystem bestehen. Da es eine Vielzahl an Menschen gibt, die eine Reihe depressiver Symptome aufweisen, aber nicht die Diagnosekriterien für die Major Depression im Sinne des DSM-IV erfüllen, haben sich in der Forschungsliteratur Begriffe wie „Subsyndromal

Symptomatic Depression“ oder „Minor Depressive Disorder“ (Judd, Rapaport, Paulus & Brown, 1994; Rapaport & Judd, 1998) etabliert. In der textrevidierten Fassung des DSM-IV wird die Minor Depressive Disorder im Anhang erwähnt (Saß et. al., 2003, S. 848). Der Hauptunterschied zur Major Depression ergibt sich hinsichtlich der Symptomanzahl: von einer Minor Depressive Disorder kann gesprochen werden, wenn zwei bis vier der für die Major Depression genannten Symptome vorhanden sind. Das Zweiwochenkriterium wird aufrecht erhalten. Kessler, Zhao, Blazer & Swartz (1997) ermittelten eine Lebenszeitprävalenz für die Minor Depressive Disorder von 7.5%. Howland et. al. (2008) verglichen das psychosoziale Funktionsniveau von Personen mit Minor Depressive Disorder mit dem von Personen mit einer diagnostizierten Major Depression und dem von Personen, die gar keine Symptome aufwiesen. Die Resultate zeigten, dass die funktionelle Beeinträchtigung der Personen, die an einer Minor Depression litten signifikant höher ausfiel als jene von symptomfreien Personen und in einigen Subskalen nicht signifikant von der Beeinträchtigung von Personen mit vollausgeprägter Major Depression abwich. Ähnliche Ergebnisse erzielten Nierenberg et. al. (2010).

Alle genannten Aspekte werfen die Frage nach Verbesserungen und Alternativen auf. Eine beträchtliche Anzahl von Studien beschäftigte sich mit der Frage, ob affektive Störungen nicht valider durch eine *dimensionale* Konzeptionierung erfasst werden können. Hankin, Fraley, Lahey & Waldman (2005) stellten in einer taxonometrischen Analyse von strukturierten klinischen Interviews bei Kindern fest, dass das Verständnis von Depression als Kontinuum den erhobenen Daten deutlich besser gerecht würde als eine kategoriales Verständnis. Angst & Merikangas (2001) plädierten auf Basis einer auf 15 Jahre angelegten Längsschnittanalyse bei Erwachsenen ebenfalls für eine dimensionale Konzeptionierung von Depressionen und empfahlen dabei neben der Symptomanzahl, die einen linear positiven Zusammenhang zu den verwendeten Validitätskriterien wie beispielsweise dem beruflichen Funktionsniveaus aufwies, die Dauer und Häufigkeit depressiver Episoden ebenfalls zu berücksichtigen und ebenfalls mit kontinuierlichen Maßen zu messen.

2.1.6 Zusammenfassung und Bezug zur Studie

Depressionen stellen die am häufigsten vorkommende psychische Krankheit dar. Ihre Auswirkung auf die betroffenen Individuen und deren Angehörige sind oftmals ausgesprochen gravierend. Eine Vielzahl der Betroffenen bleibt unerkannt und unbehandelt (Wittchen & Jacobi, 2005). Dies macht die Bedeutsamkeit, einer zuverlässigen Diagnostik deutlich. Die aktuelle gültigen Diagnosesysteme DSM-IV und ICD-10 basieren auf einer klassifikatorischen bzw. kategorialen Konzeption der depressiven Störungen. Die Validität des verwendeten Ansatzes ist aus einer Mehrzahl von Gründen fragwürdig. So sehen beide Diagnosesysteme Mindestanzahlen für die einschlägigen Symptome vor, deren Festlegung empirisch nicht zu begründen ist und die daher willkürlich erscheint. Des Weiteren nennen sowohl ICD-10 als auch DSM-IV eine Mindestdauer von zwei Wochen für das Bestehen der Symptome. Auch hinsichtlich dieses Kriteriums müssen Zweifel hinsichtlich der Validität angemeldet werden. Empirische Untersuchungen sprechen dafür, dass auch kürzer andauernde depressive Episoden die betroffenen Personen in starkem Maße beeinträchtigen können. Ein weiteres Problem ist die Frage danach, ob die Erfüllung des Zeitkriteriums im Rahmen des diagnostischen Prozesses überhaupt in realistischer Form festgestellt werden kann. Dies zu überprüfen ist eines der Ziele der Vorgängerarbeit von Heckmann und ist eines der Ziele dieser Arbeit.

2.2 State- und Trait-Depressivität

In der Persönlichkeitspsychologie hat sich hinsichtlich der zeitlichen Stabilität von Konstrukten die Unterscheidung in States (zeitlich instabil und stark durch Situationsfaktoren beeinflusst) und Traits (zeitlich stabil und transsituativ konsistent) etabliert (z.B. Kelava & Schermelleh-Engel, 2007). Vor diesem Hintergrund stellt sich die Frage, ob es sich beispielsweise bei der Major Depression im Sinne des DSM-IV um einen State oder einen Trait handelt. So wie die Frage im vorangegangenen Satz formuliert wurde, impliziert sie eine kategoriale Trennung zwischen State und Trait. Diese Sichtweise geriet in der Persönlichkeitspsychologie zunehmend in die Kritik. Cooper & McConville (1990) ließen Probanden über einen Monat täglich ihre

Stimmung einschätzen. Dabei erklärten interindividuelle Unterschiede 25% der Varianz der dabei festgestellten Stimmungsschwankungen. Die Autoren schlossen daraus, dass der Zusammenhang zwischen Stimmung und Traitmaßen in der Literatur unterschätzt wird (zitiert nach Amelang & Bartussek, 1997). Allen & Potkay (1981) charakterisierten Traits als Summation von State-Einheiten und bezeichneten daher die Übergänge zwischen States und Traits als fließend und Grenzziehungen zwischen den beiden Konstruktarten als willkürlich (zitiert nach Amelang & Bartussek, 1997). Um diesem Umstand Rechnung zu tragen müsste man daher eher „von Traits als den *relativ* stabilen und überdauernden, von States hingegen als den *relativ* temporären Charakteristika“ sprechen (Amelang & Bartussek, 1997, S. 59, Hervorhebungen im Text). Janke & Hüppe (1991) sehen Stimmungen als „zeitlich ausgedehnte Gefühle“ und setzen zwischen diese und die langfristigen Merkmale des emotionalen Erlebens „mittelfristige Zustände“ wie z.B. depressive Verstimmungen (zitiert nach Amelang & Bartussek, 1997). In diesem Zusammenhang darf auch die sogenannte Konsistenzkontroverse in der Persönlichkeitspsychologie und die daraus resultierende Interaktionismusdebatte nicht unerwähnt bleiben. Im Rahmen der Konsistenzkontroverse wurde in Zweifel gezogen, ob die Annahme von Traits als zeitlich stabilen und transsituativ konsistenten Eigenschaften überhaupt aufrecht erhalten werden kann (z.B. Schmitt, 1992, 2005). Im Zuge dieser Diskussion gewann die Einsicht, „dass die Leistungsfähigkeit des Eigenschaftsmodells gesteigert werden kann, wenn es mit interaktionistischem Gedankengut angereichert wird und die Verhaltenswirksamkeit von Situationsmerkmalen einbezogen wird.“ (Schmitt, 2005, S. 107). Aus diesem Grund geht man heute davon aus, dass Messungen im Rahmen der psychologischen Diagnostik sowohl von transsituativ konsistenten, stabilen Merkmalen als auch von situationsspezifischen, zeitlich instabilen Faktoren beeinflusst wird (Kelava & Schermelleh-Engel, 2007). Die Latent-State-Trait-Theorie (Steyer, Ferring & Schmitt, 1992) bietet die testtheoretische Grundlage, um dem Rechnung zu tragen, ohne eine systematische Erfassung der situativen Faktoren vorzunehmen (siehe Kapitel 2.5).

Die Frage nach dem State- oder Traitcharakter der Major Depression ist also falsch gestellt oder mindestens unpräzise formuliert. Es ist davon auszugehen, dass das Depressivitätsniveau einer Person sowohl von stabilen Eigenschaften der Person als auch von situativen Einflüssen beeinflusst wird. Aus der Persönlichkeitsforschung ist dabei bekannt, dass ein hoher Zusammenhang zwischen Neurotizismus und sowohl dem Schweregrad als auch der Dauer von depressiver Symptomatik besteht (z.B. Brown, 2007; Gershuny & Sher, 1998). Was situative Einflüsse betrifft, so konnte beispielsweise (Kessler, 1997) zeigen, dass gravierende Lebensereignisse wie etwa der Verlust des Arbeitsplatzes einen hohen Einfluss auf das Depressivitätsniveau haben können. Doch auch alltägliche Stressoren alltäglichen Stressoren ("daily hassles") wie z.B. Streitigkeiten mit Freunden können zu verstärkter Depressivität führen (Lovejoy & Steuerwald, 1997). Für die Interaktion zwischen Personenfaktoren und Stressoren sprechen beispielsweise die Befunde von Wetter & Hankin (2009), die zeigen, dass der Zusammenhang zwischen Neurotizismus und depressiver Symptomatik durch das Stressniveau mediiert wird.

2.3 Das autobiographische Gedächtnis

Ein wichtiger Aspekt der Frage danach, ob zeitliche Instruktionen in Fragebögen umgesetzt werden können ist die Frage danach, ob die Versuchspersonen ausreichend Informationen über die für die Beantwortung relevanten Ereignisse aus ihrem Gedächtnis abrufen können. Fragen dieser Art behandelt die Forschung zum autobiographischen Gedächtnis (Schwarz & Sudman, 1994). Die Vorgängerarbeit von Heckmann hat sich diesem Gebiet sehr ausführlich gewidmet. Daher wird es im Rahmen dieser Arbeit nur überblicksartig dargestellt. Dabei sollen zunächst typische Fehler beim Abruf aus dem biographischen Gedächtnis betrachtet werden (Kapitel 2.3.1) und im Anschluss die Unterscheidung zwischen semantischem und episodischem Emotionswissen (Kapitel 2.3.2).

2.3.1 Fehler im autobiographischen Gedächtnis

Sollen zeitliche Instruktionen bei der Bearbeitung psychologischer Instrumente adäquat berücksichtigt werden, ist es dazu notwendig, dass die erinnerten Ereignisse kategorisiert werden. Damit ist gemeint: die Versuchsperson muss sich darüber im Klaren sein, ob ein erinnertes Ereignis in den durch die Instruktion vorgegebenen Zeitraum fällt oder nicht. Dazu muss zunächst der Beginn des erfragten Zeitraums erinnert werden, dann das relevante Ereignis und im Anschluss erfolgt die Integration zu einem Gesamturteil (Heckmann, 2008, S. 8). Kommt es dabei dazu, dass entgegen den Vorgaben der Instruktion Ereignisse berücksichtigt werden, die *nicht* in den erfragten Zeitraum fallen, so spricht man von *Telescoping* (Sudman, Bradburn & Schwarz, 1996, zitiert nach Heckmann, 2008, S. 9). Werden die erfragten Informationen nicht direkt erinnert, so wenden die befragten Personen häufig *Schätzstrategien* an. Ein Beispiel hierfür wäre, dass eine Basisrate des relevanten Verhaltens verwendet wird und diese dann für den erfragten Zeitraum „hochgerechnet" wird (Menon, 1994, zitiert nach Heckmann, 2008, S. 9). Des Weiteren kann das Gesamturteil dadurch verzerrt werden, dass bestimmte besonders saliente Erinnerungsmomente das Urteil in höherem Maße beeinflussen als weniger saliente (Hank, Schwenkmezger & Schumann, 2001, zitiert nach Heckmann, 2008, S. 9).

2.3.2 Semantisches und episodisches Emotionswissen

Robinson und Clore (2002 a, b, zitiert nach Heckmann, 2008, S. 10) unterscheiden zwischen episodischem und semantischem Emotionswissen. Das episodische Emotionswissen ist an Zeit und Ort gebunden und wird vor allem bei Berichten über aktuelle Emotionen genutzt, wohingegen das semantische Emotionswissen aus allgemeinen Annahmen über die eigenen Emotionen besteht (Heckmann, 2008, S. 11). Sofern episodisches Wissen verfügbar ist, wird es bevorzugt für die Beantwortung von Fragen benutzt. Weil die Fähigkeit, sich an bestimmte Details zu erinnern mit der Zeit relativ schnell abnimmt, wodurch das episodische Emotionswissen relativ schnell fehleranfällig wird, wird bei längeren Erinnerungszeiträumen tendenziell eher semantisches Emotionswissen verwendet. Der Bezug auf das semantische

Emotionswissen beginnt bereits, wenn die Instruktion für die Versuchspersonen lautet, sie sollen sich auf die letzten paar Wochen ("last few weeks") beziehen. Die Autoren gehen davon aus, dass beide Arten des Emotionswissens eigene Fehlerquellen besitzen. Das Emotionswissen ist also fehleranfällig, was eine mögliche Erklärung für eventuelle Probleme bei der Umsetzung zeitlicher Instruktionen in Fragebögen wäre. Des Weiteren könnte der verstärkte Einsatz des semantischen Emotionswissens bei abgefragten Bezugszeiträumen von mehreren Wochen dazu führen, dass Fragebögen wie der BDI-V unabhängig von der zeitlichen Instruktion einen hohen Traitanteil aufweisen (Heckmann, 2008, S. 63).

2.4 BDI-O und BDI-V

Der folgende Abschnitt dient der Vorstellung des im Rahmen dieser Studie eingesetzten Messinstruments. Dazu wird zunächst das Beck-Depressions-Inventar (im Folgenden BDI-O genannt) vorgestellt (Kapitel 2.4.1), welches die Grundlage lieferte, bevor die vereinfachte Version von Schmitt & Maes (2000) (im Folgenden BDI-V genannt) präsentiert wird (Kapitel 2.4.2).

2.4.1 Das BDI-O

Das Beck-Depressions-Inventar (im Folgenden zur Abgrenzung vom BDI-V als BDI-O bezeichnet) ist das international am meisten gebrauchte Selbstbeurteilungsinstrument in der Diagnostik affektiver Störungen (Hautzinger & Meyer, 2002). Das BDI-O umfasst 84 Items bzw. 21 Symptomkategorien, zu denen jeweils 4 Aussagen mit steigender Intensität vorgegeben sind. Diesen Aussagen sind Zahlen von null (niedrigste Intensität) bis drei (höchste Intensität) zugeordnet (Richter, 1991). Die durch die 21 Items gebildete Skala umfasst die folgenden Aspekte: traurige Stimmung, Pessimismus, Versagen, Unzufriedenheit, Schuldgefühle, Strafbedürfnis, Selbsthass, Selbstanklagen, Selbstmordimpulse, Weinen, Reizbarkeit, sozialer Rückzug und Isolierung, Entschlussunfähigkeit, negatives Körperbild, Arbeitsunfähigkeit, Schlafstörungen, Ermüdbarkeit, Appetitverlust, Gewichtsverlust, Hypochondrie und Libidoverlust (Hautzinger & Meyer, 2002).

Das BDI-O hat eine Reihe von Modifikationen erfahren, die unter anderem den zeitlichen Bezugsraum des Erfragten betrafen. Als 1961 die ursprüngliche Version von Beck, Ward, Mendelson, Mock & Erbaugh publiziert wurde, lautete die zeitliche Instruktion: "right now". Im Jahr 1978 änderte Beck diese Anweisung. Ab diesem Zeitpunkt lautete sie: "past week, including today". Die aktuelle Version (Beck, Steer & Brown, 1996, deutsche Übersetzung: Hautzinger, Keller & Kühner, 2006) berücksichtigt das beschriebene Zweiwochenkriterium des DSM-IV ersucht die Befragten daher darum, sich bei der Beantwortung auf die letzten 14 Tage zu beziehen (Hautzinger, Keller & Kühner, 2006). Beck, Steer & Garbin (1988) gehen davon aus, dass die zeitliche Instruktion einen Einfluss darauf hat, ob mit dem Instrument eher der gegenwärtige Zustand oder länger anhaltende Gefühlszustände bzw. Einstellungen erfasst werden, nennen allerdings keine empirische Belege für die Annahme (Heckmann, 2008, S. 12).

Hautzinger & Meyer (2002) berichten, dass die interne Konsistenz des BDI-O bei Patientenstichproben meist über .90 liege. Für die Stabilität des Summenwertes über eine Woche geben sie den Wert von .75 an. Die Korrelationen mit anderen Selbstbeurteilungsmaßen beziffern sie auf .76 bis über .80. Reliabilität und Validität wurden in einer Vielzahl von Studien für gut befunden (Lukesch, 1974; Beck, Steer & Garbin, 1988; Richter, 1991; Kühner, Bürger, Keller & Hautzinger, 2007). Als Erfolgsmaß bei Interventionsstudien erwies sich das BDI-O als veränderungssensitiv (Hautzinger & De Jong-Meyer, 1996).

2.4.2 Das BDI-V

Schmitt & Maes (2000) sahen ein Manko des BDI-O in seiner unökonomischen Schwierigkeitsskalierung. Diese beeinträchtige die vor allem für epidemiologische Untersuchungen und multivariate Fragenbogenstudien bedeutsame Nützlichkeit des Messinstruments. Des Weiteren gehen Schmitt et. al. (1993) davon aus, dass die Bearbeitung der 84 Items des BDI-O einen für depressive Patienten unnötigen hohen Aufwand darstellen. Außerdem verweisen die Autoren auf die erhöhten Kosten, die ein derart umfangreiches Messinstrument bei multivariaten Fragebogenstudien an großen Stichproben verursacht.

Das Item zum Gewichtsverlust hatte in Voruntersuchungen regelmäßig die geringste Trennschärfe besessen (Hautzinger, Bailer, Worall & Keller, 1994; Kammer, 1984, zitiert nach Schmitt & Maes, 2000). Aus diesem Grund wurde es nicht in das BDI-V übernommen. Die verbleibenden 20 Symptomkategorien wurden mit jeweils einem Item abgefragt. Die vierstufige Schwierigkeitsskalierung wurde durch einer Häufigkeitsskala ersetzt. Diese umfasst sechs Stufen von 0 („nie“) bis 5 („fast immer“). Als zeitliche Instruktion wählten die Autoren: „Wie ist Ihr gegenwärtiges Lebensgefühl?“ (Schmitt & Maes, 2000, S. 39).

Auf Basis der bisherigen Studien kann dem BDI-V eine gute Reliabilität attestiert werden. Erste Belege, die auf eine gute Validität hindeuten konnten ebenfalls erbracht werden. Schmitt & Maes (2000) schätzten auf Basis der Untersuchung einer demographisch heterogenen Stichprobe von ca. 2500 Personen die Reliabilität des BDI-V anhand einer Latent-State-Trait-Analyse auf .95. Für eine gute Validität des Messinstruments sprechen die im Rahmen der zitierten Studie ermittelten hohen Korrelationen mit anderen Indikatoren der seelischen Gesundheit. Die Studie von Schmitt et. al. (2003) konnte den Befund der sehr guten Messeigenschaften bestätigen. Zudem konnten sie zeigen, dass die Messeigenschaften von BDI-O und BDI-V sehr ähnlich sind. Die Korrelationen von BDI-V und BDI-O liegt höher als die Korrelationen der beiden Instrumente mit anderen Depressionsmaßen, beide korrelieren in ähnlicher Höhe mit einem Expertenurteil. Auf der Ebene der Einzelsymptome konnte Konvergenz zwischen den beiden Fragebögen festgestellt werden. Zudem konnten beide Version ähnlich gut zwischen Personen mit depressiver Symptomatik und symptomfreien Personen diskriminieren. Im Jahr 2006 normierten Schmitt, Altstötter-Gleich, Hinz, Maes & Brähler das BDI-V an einer großen, demographisch heterogenen Stichprobe. Hierbei zeigte sich ein signifikanter Geschlechtsunterschied: Frauen wiesen im Durchschnitt höhere Depressivitätswerte auf als Männer. Aus diesem Grund wurden für die Geschlechter getrennte Normwerte ermittelt. Die Autoren geben 35 als Grenze für den BDI-V-Summenwert an. Wird dieser Wert überschritten liegt mit 90%iger Wahrscheinlichkeit eine klinisch relevant depressive Erkrankung vor. Wird er

unterschritten, so kann mit der gleichen Wahrscheinlichkeit davon ausgegangen werden, dass eine solche nicht vorliegt.

2.5 Latent-State-Trait-Theorie

Das folgende Kapitel widmet sich der Latent-State-Trait-Theorie (LST-Theorie) (Steyer, Ferring & Schmitt, 1992). Diese bietet die testtheoretische Grundlage für die im Rahmen dieser Studie durchgeführten Analysen.

Als Ausgangspunkt für die Entwicklung der Latent-State-Trait-Theorie sehen Deinzer et. al. (1995) die Diskussion über die Person-Situation-Interaktion sowie die durch diese angeregte Forschung (für Zusammenfassungen dieser Diskussion siehe z.B. Steyer, Schmitt & Eid, 1999; Lucas & Donnellan, 2009). Lange Zeit dominierte in diesem Bereich der experimentelle Forschungsansatz kombiniert mit varianzanalytische Verfahren als Auswertungsmethode der Wahl, wohingegen die im wesentlichen mit Fragebögen operierende Korrelationsforschung ins Hintertreffen geraten war, da sie notorisch unter dem Problem litt, dass sie die situativen Einflüsse und die der Person-Situation-Interaktion nicht systematisch erfassen konnte (Deinzer et. al., 1995). Van Heck (1984, 1989) beispielsweise unternahm den Versuch, eine allgemeine Taxonomie der Situationsbedingungen aufzustellen. Doch der Einfluss dieser Bemühungen auf die Forschung blieb gering. Weiterhin muss man konstatieren, dass die Identifizierung und Erfassung relevanter Situationen ein enorm aufwändiges Unterfangen darstellt (Deinzer et. al., 1995). In Hogans (2009) Worten: "…the conceptual status of 'situations' is a mess."

Die LST-Theorie hat eine Möglichkeit geschaffen, diesem Problem aus dem Weg zu gehen:

> LST theory differs from the experimental paradigm in that the situations in which measurement takes place do not have to be known and do not have to be observed (…) in order to determine the proportion of variance accounted for by situations and/or interactions. However, if variables describing specific

> characteristics of situations are present, it will be possible to estimate their effects as well. (Steyer, Schmitt & Eid, 1999, S. 392)

Um dieses Ziel zu erreichen dekomponiert die LST-Theorie jede beobachtete Variable, wie z.B. den Summenwert, den eine Person beim Ausfüllen eines bestimmten Fragebogens erzielt, in eine latente State-Variable und einen Messfehleranteil. Der latente State repräsentiert also den aus der klassischen Testtheorie bekannten *wahren Wert* der Ausprägung des untersuchten Merkmals. Die klassische Testtheorie wird aber im Rahmen der LST-Theorie erweitert: die latente State-Variable ihrerseits wird wiederum in den Anteil einer latenten Trait-Variable einerseits und den Anteil des Latent-State-Residuums andererseits aufgeteilt. Der Trait-Anteil reflektiert den Einfluss der Persönlichkeit, das Latent-State-Residuum beinhaltet den Einfluss der Situation und der Interaktion von Person und Situation (Steyer et. al., 1999).

2.6 Forschungsstand

Der folgende Abschnitt gibt den aktuellen Forschungsstand für die im Rahmen dieser Studie relevanten Fragen wieder. Zunächst wird dabei die Forschung zur Auswirkung zeitlicher Instruktionen betrachtet (2.6.1), bevor Untersuchungen über die Retest-Stabilität bzw. den Trait- und State-Anteil von BDI-O und BDI-V berichtet werden (2.6.2). Schließlich werden die Ergebnisse der Vorgängerstudie von Heckmann (2008) dargestellt.

2.6.1 Zeitliche Instruktionen

Es gibt nur wenige Untersuchungen zur Auswirkung zeitlicher Instruktionen bei Fragebögen. Systematisch untersucht wurde dies für die „Positive and Negative Affect Schedule“ (PANAS, Watson, 1988; Watson, Clark & Tellegen, 1988). Für die wurden sieben unterschiedliche zeitliche Instruktionen getestet (“right now, that is, at the present moment“, “today“, “during the past few days“, ”during the past week“, “during the past few weeks“, “during the past few years”, “in general, that is, on the average”).

Watson, Clark & Telegen hatten 1988 die Teilnehmerinnen und Teilnehmer ihrer Studie gebeten, die sieben unterschiedlichen Instruktionen für die PANAS jeweils mit einer Woche Abstand zu beantworten. Dabei waren zwei Durchgänge zu absolvieren, für jede Instruktion wurde die PANAS also von jeder Versuchsperson zweimal ausgefüllt. Die Retest-Korrelationen für die Skala Negativer Affekt wies dabei den erwartenden Anstieg mit größer werdendem vorgegebenem zeitlichen Bezugsrahmen auf. Lediglich die auf den aktuellen Moment bezogene Instruktion fiel aus dem Rahmen, indem sie eine höhere Stabilität aufwies als die Instruktionen heute, die letzten Tage und die letzte Woche. Gleiches galt für die Momentinstruktion bei der Skala Positiver Affekt. Bei dieser Skala kam hinzu, dass die Unterschiede zwischen den Instruktionen nur sehr gering ausfielen, der erwartete Anstieg also letztlich nicht konstatiert werden konnte. Die Autoren interpretieren ihre Ergebnisse dahingehend, dass die PANAS mit kürzeren Instruktionen eine hohe Sensibilität für Stimmungsänderungen zeige, wohingegen sie bei der Vorgabe längerer Zeiträume vornehmlich stabile Eigenschaften messe.

Krohne, Egloff, Kohlmann & Tausch hatten 1996 ähnliche Untersuchungen für die deutsche Übersetzung der PANAS durchgeführt. Dabei wurden sechs unterschiedliche zeitliche Instruktionen getestet („im Moment“, „heute“, „in den letzten Tagen“, „in den letzten Wochen“, „im letzten Jahr“, „im Allgemeinen“). Die Autoren gingen dabei davon aus, dass je größer der zeitliche Bezugsrahmen ist, desto größer ist der Zusammenhang mit der Instruktion „im Allgemeinen“, bei der sie davon ausgingen, dass sie vor allem stabile Eigenschafen misst und dass der Zusammenhang zur State-Version („im Moment“) umso geringer wird. Diese Hypothesen konnten weitestgehend bestätigt werden. Das von den Autoren angenommene Korrelationsmuster zeigte sich allerdings nur für den aktuellen Affekt, beim Trait-Affekt konnte dieses Muster nur auf der Ebene der Regressionskoeffizienten nachgewiesen werden.

2.6.2 BDI-V und BDI-O

Für das BDI-O gibt es bisher keine Studien, die sich mit der Auswirkung zeitlicher Instruktionen auf das Antwortverhalten befasst haben. Untersuchungsergebnisse gibt es allerdings hinsichtlich der zeitlichen Stabilität und der Änderungssensitivität. Zimmerman (1986) berichtete bei einer studentischen Stichprobe für ein Restest-Intervall von einer Woche für das BDI-I eine Test-Retest-Korrelation von lediglich .64. Hatzenbuehler, Parpal & Matthews (1983) berichteten, dass der wiederholte Einsatz des BDI-I zu einem signifikanten Rückgang der Summenwerte führe und rieten daher bei Therapieevaluationsstudien, die das BDI verwenden für Testwiederholungseffekte zu kontrollieren. Ähnliche Effekte beschreiben Ahava, Iannone, Grebstein & Sherling (1998). Barkham, Mullin, Leach, Stiles & Lucock (2007) hatten das BDI-I Patientinnen und Patienten in britischen Psychotherapiepraxen mehrfach vorgelegt und dabei verschiedene Messintervalle untersucht: die Test-Retest-Korrelation lag dabei zwischen .91 für den Retest-Abstand von einem Monat. Für 3 Monate lag sie bei .81, für 9-12 Monate bei .64. Ein Problem der Interpretation von Retest-Korrelationen besteht allerdings in der Konfundierung von Reliabilität und differentiellen Veränderungen (Mohiyeddini, Hautzinger & Bauer, 2002). Es kann nicht klar gesagt werden, ob eine vergleichsweise niedrige Retestkorrelation auf einer niedrigen Reliabilität beruht oder auf einer geringen Stabilität des untersuchten Merkmals. Die in Kapitel 2.5 vorgestellte Latent-State-Trait-Theorie bietet eine Möglichkeit, diese beiden Komponenten zu trennen. Es liegen bisher nur wenige solcher Analysen für das BDI vor. Mohiyeddini, Hautzinger & Bauer (2002) schätzten für das BDI-I mittels Latent-State-Trait-Analyse den durch die Traitkonsistenz erklärten Anteil der Varianz auf 49%, den Anteil der Zeitspezifität auf 22%. Der vorgegebene zeitliche Bezugsrahmen war dabei die vergangene Woche inklusive des Tages der Befragung. Der Abstand zwischen den beiden Messzeitpunkten hatte 4 Monate betragen. Schmitt & Maes (2000) hatten für das BDI-V mit der Instruktion „Wie ist Ihr gegenwärtiges Lebensgefühl?“ bei einem Retest-Intervall von 2 Jahren einen Wert für die Traitkonsistenz von .59 und für die Zeitspezifität von .26 auf der Ebene der Testhälften ermittelt.

2.6.3 Ergebnisse der Vorgängerstudie

Der folgende Abschnitt soll die Ergebnisse der Vorgängerstudie von Heckmann (2008) kurz zusammenfassen. In dieser Untersuchung wurde den Teilnehmerinnen und Teilnehmern das BDI-V zweimal im Abstand von 14 Tagen als Online-Fragebogen zur Bearbeitung dargeboten. Die Versuchspersonen wurden dabei in zwei Instruktionsgruppen aufgeteilt. Die eine erhielt die Instruktion, sich bei ihren Antworten auf die vorangegangenen 14 Tage inklusive des Tags der Bearbeitung zu beziehen (im Folgenden 14-Tage-Gruppe genannt), die andere sollte sich auf die vorangegangenen 3 Monate beziehen (im Folgenden 3-Monats-Gruppe genannt) (Heckmann, 2008, S. 43). Die Autorin hatte dabei die Hypothese untersucht, dass in der 14-Tage-Gruppe eine signifikant höhere Zeitspezifität und korrespondierend dazu, eine niedrigere Traitkonsistenz als in der 3-Monats-Gruppe zu ermitteln ist, wobei der Einfluss der Methodenfaktoren und die Reliabilität jeweils gleich ist (S. 25). Diese Hypothesen wurden mittels Latent-State-Trait-Analysen geprüft. Für die 3-Monats-Gruppe ermittelte Heckmann (2008) dabei eine Traitkonsistenz von .84, die Zeitspezifität betrug .08 (die angegebenen Werte beziehen sich auf den Gesamttest) (S. 57). Für die 14-Tage-Gruppe lag die Traitkonsistenz bei .79, die Zeitspezifität bei .12. Das von Heckmann aufgestellte Modell, welches Unterschiede hinsichtlich der eine Auswirkung der zeitlichen Instruktionen repräsentierenden Parameter zuließ passte signifikant besser auf die von Heckmann (2008) erhobenen Daten als das Modell, welches keine Unterschiede zwischen den Instruktionsgruppen zuließ (S. 57). Heckmann (2008) konnte ihre Hypothesen insgesamt als bestätigt ansehen konnte, wenngleich die Gruppenunterschiede minimal ausfielen und damit insgesamt zu konstatieren war, dass das BDI-V zu einem sehr hohen Anteil stabile Eigenschaften misst (S. 61).

3 Fragestellung

Im folgenden Kapitel soll auf Basis der dargestellten Untersuchungen die Fragestellung der vorliegenden Untersuchung hergeleitet werden (3.1), bevor Fragestellung und Hypothesen der Arbeit dargestellt werden (3.2).

3.1 Herleitung der Fragestellung

Für die Depressionsdiagnostik ist es in hohem Maße von Bedeutung, reliable und valide Messinstrumente zur Verfügung zu haben. Von großem Interesse ist dabei die Frage, ob ein verwendetes Messinstrument in stärkerem Maße stabile Eigenschaften misst oder ob das Antwortverhalten vornehmlich durch situative Umstände beeinflusst wird (Mohiyeddini, Hautzinger & Bauer, 2002). Eine wichtige Einflussgröße ist dabei die zeitliche Instruktion, auch vor dem Hintergrund, dass DSM-IV und ICD-10 für psychische Erkrankungen wie beispielsweise die Major Depression oder die Dysthyme Störung exakt definierte zeitliche Kriterien vorgeben (Saß et. al. 2003; Dilling et. al., 2006). Der kognitive Aufwand bei der Repräsentation zeitlicher Instruktionen muss als hoch betrachtet werden (Sudman, Bradburn & Schwarz, 1996). Dennoch konnte beispielsweise mit der Positive Negative Affect Schedule (Watson, 1988; Watson, Clark & Tellegen, 1988; für die Deutsche Version: Krohne, Egloff, Kohlmann & Tausch, 1996) nachgewiesen werden, dass zeitlichen Instruktionen einen Einfluss auf das Antwortverhalten bei Fragebögen zur Affektivität haben. Dabei konnte der Zusammenhang, je größer der zeitliche Bezugsrahmen, desto höher der Trait-Anteil und desto niedriger der State-Anteil, annähernd nachgewiesen werden. In der Vorgängerstudie von Heckmann (2008) hatte sich ein signifikanter Unterschied zwischen der Instruktionsgruppe, die sich bei der Beantwortung des BDI-V von Schmitt & Maes (2000) auf die letzten 14 Tage beziehen sollte gegenüber der Instruktionsgruppe, die sich auf die letzten 3 Monate beziehen sollte ergeben. Der Trait-Anteil der 14-Tage-Gruppe hatte unterhalb des für die 3-Monats-Gruppe ermittelten Werts gelegen. Korrespondierend dazu lag der Anteil der situativen Einflüsse und der Person-Situations-Interaktion (Zeitspezifität) in der 14-Tage-Gruppe höher. Für diese Untersuchung werden die zeitlichen Instruktionen der

Vorgängerarbeit übernommen und dabei angenommen, dass sich der gleiche Effekt zeigen lässt.

Werden Mehrfachmessungen mit einem Instrument angestrebt, z.B. um den Verlauf einer Störung oder den Einfluss einer therapeutischen Intervention zu untersuchen, ergibt sich daraus ein weiterer interessierender Aspekt, nämlich die Auswirkungen des gewählten Messintervalls. Es ist dabei davon auszugehen, dass je größer der Abstand zwischen den Messzeitpunkten ist, umso wahrscheinlicher ist es, dass sich eine Veränderung der Lebenssituation der untersuchten Personen ergibt. Barkham et. al. (2007) hatten für das BDI-II bei einer Untersuchung an Personen, die sich in Psychotherapie befanden einen kontinuierlichen Abfall der Retestkorrelation ermittelt und fanden für den Messabstand 1 Monat eine Retest-Korrelation von .91, für 3 Monate .81 bei 9-12 Monaten lediglich .64. Mohiyeddini, Hautzinger & Bauer hatten 2002 hatten für das BDI-I bei einem Messabstand von 4 Monaten 49% der Varianz durch die Traitkonsistenz und 22% durch die Zeitspezifität erklären können. Der zeitliche Bezugsrahmen war dabei die vergangene Woche. Heckmann hatte für das BDI-V bei einem Retest-Intervall von 14 Tagen Werte von .79 für die Traitkonsistenz in der 14-Tage-Gruppe und von .84 in der 3-Monatsgruppe errechnet. Die Zeitspezifität lag bei .12 (14-Tage-Gruppe) und .08 (3-Monats-Gruppe). Auch wenn die letztgenannten Ergebnisse nur eingeschränkt interpretierbar sind, da zeitliche Instruktion und Messintervall miteinander konfundiert sind, scheint sich insgesamt der Trend abzuzeichnen, dass bei einem Abstand von mehreren Monaten die Stabilität des BDI-V abnimmt und der Anteil der Zeitspezifität an der Varianz des Antwortverhaltens zunimmt. Dieser Zusammenhang soll im Rahmen dieser Studie mit einem Retest-Intervall von 3 Monaten geprüft werden. Dieser Abstand wurde gewählt, um korrespondierend zu Heckmanns 14-Tage-Intervall beide zeitlichen Instruktionen auch im Retest-Intervall repräsentiert zu sehen. War bei Heckmann der zeitliche Bezugsrahmen für die 3-Monats-Gruppe für die beiden Messzeitpunkte nahezu identisch (2 ½ Monate wurden zweimal erfragt), so überschneiden sich die Bezugszeiträume der beiden Messzeitpunkte für diese Instruktionsgruppe in dieser Studie lediglich bezüglich eines Tages, nämlich des Untersuchungstages zu

Messzeitpunkt eins. Für die 14-Tage-Gruppe hatte sich bei Heckmann eine Überschneidung von einem Tag ergeben, in dieser Untersuchung werden sich die Bezugszeiträume der 14-Tage-Gruppe nicht überschneiden. Vielmehr liegen 2 ½ Monate zwischen dem ersten Messzeitpunkt und dem ersten beim zweiten Messzeitpunkt erfragten Tag. Vor diesem Hintergrund und aufgrund der präsentierten Forschungsergebnisse bezüglich BDI-O und BDI-V wird angenommen, dass in dieser Studie im Vergleich der Instruktionsgruppen eine höhere Zeitspezifität und eine niedrigere Traitkonsistenz erzielt wird als bei Heckmann (2008).

3.2 Fragestellung und Hypothesen

Eine der Grundannahmen dieser Arbeit besteht darin, dass die zeitliche Instruktion einen Einfluss auf das Antwortverhalten im BDI-V hat. Es wird davon ausgegangen, dass die Vorgabe eines *längeren* Zeitraums zu einem größeren Einfluss der stabilen Dispositionen führt und korrespondierend dazu der Einfluss der Situation sowie der Interaktion aus Person und Situation geringer ausfällt. Die Grundannahme sowie die Hypothesen 1a und 1b entsprechen denen in der Vorgängerstudie (Heckmann, 2008, S. 24/25).

Fragestellung 1: Führen unterschiedliche zeitliche Instruktionen zu Unterschieden in der Situationssensitivität und Stabilität des BDI-V?

Hypothese 1a: Die 3-Monatsgruppe weist eine signifikant höhere Trait-Konsistenz und eine signifikant geringere Zeitspezifität auf als die 14-Tage-Gruppe. Ein Mehrgruppenmodell, das für beide Instruktionsgruppen die gleichen Werte für Traitvarianz und Situationseinflüsse annimmt, zeigt keine akzeptable Modellgüte. Ein Modell, das für Traitvarianz und Situationseinflüsse unterschiedliche Werte in beiden Gruppen zulässt, führt zu signifikanter Verbesserung des Modellfits und zeigt einen akzeptablen Modellfit.

Hypothese 1b: Zwischen den beiden Instruktionsgruppen zeigen sich keine signifikanten Unterschiede hinsichtlich der Methodenspezifität und der Reliabilität.

Eine weitere Grundannahme dieser Arbeit besteht darin, dass eine Vergrößerung des Abstands zwischen den Messzeitpunkten dazu führt, dass depressivitätsrelevante Veränderungen der Lebensbedingungen mit höherer Wahrscheinlichkeit und in größerem Ausmaß eintreten. Daher wird davon ausgegangen, dass ein größer gewähltes Retest-Intervall dazu führt, dass für beide Instruktionsgruppen der Vergleich mit der jeweils gleichen Instruktionsgruppe in der Vorgängeruntersuchung so ausfällt, dass sich ein niedrigerer Anteil stabiler Eigenschaften und ein höherer Anteil der Situationseinflüsse bzw. der Interaktion aus Person und Situation ermitteln lässt.

Fragestellung 2: Führt bei zweimaliger Messung der Depressivität unter Einsatz des BDI-V eine Veränderung des Abstands zwischen den beiden Messzeitpunkten zu Unterschieden hinsichtlich der Situationsspezifität und der Stabilität des BDI-V?

Hypothese 2: Ein zeitlicher Abstand von 3 Monaten zwischen den beiden Messzeitpunkten führt bei konstanter zeitlicher Instruktion zu einer *höheren* Zeitspezifität und einer *niedrigeren* Traitkonsistenz als ein Retest-Abstand von 14 Tagen.

4 Methode

Das vierte Kapitel dieser Arbeit widmet sich zunächst dem Untersuchungsdesign und der Beschreibung der Stichprobe (4.1). Im Anschluss daran werden die im Rahmen der Studie erhobenen Konstrukte näher erläutert (4.2), bevor die Durchführung der Untersuchung beschrieben wird (4.3). Danach wird das hier angewendete statistische Modell der Latent-State-Trait-Analyse mit zwei Methodenfaktoren vorgestellt (4.4.), um im darauf folgenden Unterkapitel (4.5.) die Bildung der im Rahmen der Analysen verwendeten Testhälften zu erklären. Das Kapitel schließt mit einer Erläuterung der Methode, die für die Parameterschätzung verwendet wurde und einer Beschreibung der Kennwerte, die zur Beurteilung des Modellfits herangezogen wurden (4.6).

4.1 Untersuchungsdesign und Stichprobe

4.1.1 Untersuchungsdesign

Das Hauptziel dieser Studie ist es zu zeigen, wie sich eine Veränderung des Abstands zwischen den beiden Messzeitpunkten auf das Antwortverhalten der Versuchspersonen hinsichtlich des Vereinfachten Beck-Depressions-Inventars (BDI-V) (Schmitt & Maes, 2000) auswirkt. Vergleichsmaßstab ist dabei die Arbeit von Heckmann. Die folgenden Angaben zum Design entsprechen daher im Wesentlichen denen der Arbeit von Heckmann (2008, Kapitel 4.1). In beiden Studien waren die Versuchspersonen dazu aufgerufen, zu zwei Messzeitpunkten einen kurzen Fragebogen auszufüllen. In Heckmanns Arbeit betrug der Abstand zwischen den beiden Messzeitpunkten zwei Wochen. Im Rahmen dieser Arbeit wurde ein Intervall von *drei Monaten* zwischen den Erhebungszeitpunkten gewählt. Die Instruktion der Untersuchung enthielt als Information über das Ziel der Studie die Aussage, es sollten die Messeigenschaften eines psychologischen Instruments untersucht werden. Die Probanden wurden in zwei Gruppen aufgeteilt: Die eine wurde dazu ersucht, sich bei der Beantwortung der Frage auf die letzten drei Monate zu beziehen (im folgenden „3-Monats-Gruppe“ genannt), die andere sollte sich auf die letzten 14 Tage beziehen (14-Tage-Gruppe).

4.1.2 Auswahl der Teilnehmerinnen und Teilnehmer

Die Rekrutierung der Teilnehmerinnen und Teilnehmer erfolgte in erster Linie durch das Anschreiben der Moderatorinnen und Moderatoren von Internetforen der Anbieter *yahoo!* und *Google*. Die Auswahl der Gruppen erfolgte mit dem Ziel, eine möglichst große thematische Bandbreite herzustellen, damit die erzielte Stichprobe möglichst heterogen ausfällt. Dies geschah vor dem Hintergrund, dass die Varianz bei der Fähigkeit zur Retrospektion als groß angesehen werden muss und nicht bekannt ist, welchen prädiktiven Wert hierbei soziodemographische Variablen haben (Schwarz & Sudman, 1994, zitiert nach: Heckmann, 2008, S. 26) und zum anderen vor dem Hintergrund, dass sich soziodemographische Variablen als valide Prädiktoren für depressive Erkrankungen erwiesen haben (Beesdo & Wittchen, 2006). *Yahoo!*- bzw. G*oogle*-Gruppen, die ausweislich ihres Namens oder/und ihrer Beschreibung vornehmlich Jugendliche als Zielgruppe hatten wurden nicht kontaktiert. Gruppen, die in erster Linie Senioren ansprechen wurden bevorzugt in die Untersuchung aufgenommen. Obwohl sich beispielsweise ein Aufwärtstrend beim Anteil der Menschen in der Altersgruppe der 70-75jährigen zeigt, muss weiterhin davon ausgegangen werden, dass Menschen jenseits des 70. Lebensjahres im Internet weit unterdurchschnittlich aktiv sind (Yoon, Yoon & George, 2011). Daher besteht die Gefahr, dass diese Altersgruppe beim gewählten Ansatz der Onlinestudie unterrepräsentiert ist. Dem sollte durch die Privilegierung von „Senioren-Foren“ entgegengewirkt werden. Es wurden nur deutschsprachige Foren ausgewählt, um möglichst gut ausschließen zu können, dass eventuell auftretende Probleme beim Verständnis des Fragebogens auf mangelnden Deutschkenntnissen beruhen. Neben Internetforen wurden aus dem Bekanntenkreis des Autors Personen über die sozialen Netzwerke *facebook* und *StudiVZ* angeschrieben. Außerdem wurden die Fachschaften der Universität Koblenz-Landau (beide Campi) mit Ausnahme der Fachschaft für Psychologie angeschrieben und darum gebeten, den Link zur Studie an ihre E-Mailverteiler weiterzuleiten. Psychologiestudierende der im Sommersemester 2010 jüngsten und zweitjüngsten Kohorte des Campus‘ Landau wurden direkt über ihre E-Mail-Verteiler angeschrieben. Ältere Kohorten dieses Studiengangs wurden ebenfalls

per E-Mail angeschrieben. In diesen Anschreiben wurde allerdings lediglich darum gebeten, den Link zur Studie im Bekanntenkreis weiterzuleiten. Zu einer eigenen Teilnahme wurde diese Gruppe nicht aufgefordert, da bei höheren Semestern im Fach Psychologie das Risiko überdurchschnittlich hoch erscheint, dass das BDI bekannt ist, was u.U. dazu führen könnte, dass Effekte wie z.B. soziale Erwünschtheit die Varianz des Antwortverhaltens künstlich verkleinern. Mit dem Ziel, die Motivation zur Teilnahme an der Studie zu erhöhen wurde unter denjenigen, die zu beiden Messzeitpunkten den Fragebogen abschickten fünf Gutscheine à 10 Euro des Internetversandhauses *Amazon* verlost.

4.1.3 Stichprobenbeschreibung

Zum ersten Messzeitpunkt nahmen insgesamt 682 Personen an der Untersuchung teil, zum zweiten Zeitpunkt waren es 433 Personen. 255 Personen mussten aufgrund der Tatsache, dass nur die Daten für den ersten Erhebungszeitpunkt vorhanden waren ausgeschlossen werden. Probanden, die den für sie avisierten zweiten Messzeitpunkt um mehr als vier Tage verfehlten wurden ebenfalls ausgeschlossen. Dies betraf 6 Personen. Die Größe der Analysestichprobe betrug daher 427 Personen.

Zur Prüfung, ob signifikante Unterschiede zwischen der Gruppe, die beide Erhebungszeitpunkte absolvierte (im folgenden Analyse-Gruppe genannt) und der Gruppe, die nur am ersten Messzeitpunkt teilnahm, besteht wurde eine Dropout-Analyse (Chi^2-Test und ANOVA) durchgeführt. Diese ergab keine signifikanten Unterschiede zwischen den beiden Gruppen hinsichtlich der Merkmale Geschlecht, Alter und BDI-V-Summenwert. Die Gruppen unterschieden sich aber signifikant hinsichtlich des höchsten Bildungsabschlusses und der Tätigkeit, der die Personen zum ersten Zeitpunkt der Erhebung laut ihrer Angaben nachgingen. Die höheren Bildungsabschlüsse, also Abitur und Hochschulabschluss, dominierten beide Gruppen, wobei diese Überrepräsentation in der Analysegruppe mit etwa 89% signifikant höher ausfiel als in der Dropout-Gruppe mit etwa 80%. Innerhalb der Gruppen der Personen mit höheren Bildungsabschlüssen zeigte sich, dass in der Analysegruppe der Anteil der Personen mit Hochschulabschluss mit annähernd 50% gegenüber 39%

Abiturienten deutlich höher lag, während in der Dropout-Gruppe beide Gruppen jeweils etwa 40% ausmachten. Hinsichtlich der Tätigkeit ergab sich als größter Unterschied zwischen den beiden Gruppen ein um 8 Prozentpunkte höherer Anteil (37% vs. 29%) von Studierenden in der Analysegruppe. Die einzelnen Ergebnisse der Dropout-Analyse können den Tabellen B.1 (Chi^2-Tests) und B.2 (ANOVA) entnommen werden.

In der Gesamtstichprobe befanden sich 58% Frauen. Bei einer Spannbreite von 16 bis 75 Jahren betrug der Altersdurchschnitt 34,7 Jahre. 4% der Befragten gaben an, einen Hauptschlussabschluss als höchsten Bildungsabschluss zu haben, 10% einen Realschulabschluss, 40% das Abitur und 46% einen Hochschulabschluss. 52% der Probanden waren berufstätig, 2% Schüler, 1% in Ausbildung außerhalb von Schule und Hochschule, 34% Studierende und 11% waren entweder arbeitslos oder im Ruhestand. Unter den Teilnehmerinnen und Teilnehmern befanden sich 31% verheiratete, 36% lebten in einer Beziehung ohne verheiratet zu sein, 29% gaben an, in keiner Beziehung zu leben, 1% war verwitwet und 4% geschieden.

Die 3-Monats-Gruppe umfasste zum ersten Erhebungszeitpunkt 305 Personen, von denen 187 in die Analysestichprobe eingingen. Zur 14-Tage-Gruppe gehörten zum ersten Messzeitpunkt 377 Personen, von denen 240 am zweiten Messzeitpunkt fristgerecht teilnahmen. Eine zufällige Zuteilung der Personen zu diesen beiden Gruppen war nur im Fall der Versuchspersonenanwerbung über *facebook* und *StudiVZ* möglich. In den anderen Fällen erfolgte die zufällige Zuteilung auf der Ebene vollständiger Mailinglisten. Dies war aufgrund der Tatsache, dass ein Zugang zu den Einzeladressen der Teilnehmer technisch nur schwer und vermutlich auch nicht ohne einen Verstoß gegen das Datenschutzrecht zu realisieren gewesen wäre. Problematisch ist dieses Vorgehen insofern als es möglich wäre, dass auf die beschriebene Weise bestimmte Personengruppen in einer der beiden Gruppen der Analysestichprobe stark überrepräsentiert sind. Diese könnten gemeinsame psychologische und/oder soziodemographische Merkmale aufweisen, die im Zusammenhang mit den im Rahmen dieser Studie untersuchten Fragestellungen stehen. So könnten sich 3-Monats- und 14-Tage-Gruppe z.B. hinsichtlich der Erinnerungsfähigkeit hinsichtlich ihres

Befindens unterscheiden oder sie könnten unterschiedliche Depressivitätswerte aufweisen. Zur Überprüfung, ob sich die beiden Instruktionsgruppen hinsichtlich der erfassten soziodemographischen Variablen oder der Höhe der Depressivität unterscheiden wurden die beiden Gruppen mittels Chi^2-Test (genaue Ergebnisse in Tabelle B.3) und ANOVA (Tabelle B.4) miteinander verglichen. Es zeigten sich hierbei keine signifikanten Unterschiede hinsichtlich des Bildungsabschlusses, des Alters und der Depressivität. Signifikante Unterschiede ergaben sich hinsichtlich des Geschlechts und der Tätigkeit. Während die 3-Monats-Gruppe ein annähernd gleichverteiltes Geschlechterverhältnis aufwies (53% Frauen), war die 14-Tage-Gruppe zu annähernd zwei Dritteln mit Frauen besetzt. In Bezug auf die Tätigkeit zeigte sich in der 14-Tage-Gruppe vor allem ein weitaus höherer Anteil Studierender (43% gegenüber 28% in der 3-Monats-Gruppe) zulasten der Berufstätigen (48% in der 14-Tage-Gruppe, 56% in der 3-Monats-Gruppe). Die nach Instruktionsgruppen getrennte Stichprobenbeschreibung für die Analysegruppen ist in den Tabellen B.5 (3-Monats-Gruppe) und B.6 (14-Tage-Gruppe) dargestellt.

4.2 Operationalisierung des erhobenen Konstrukts

Depressivität wurde im Rahmen dieser Studie mittels des Vereinfachten Beck Depressionsinventars (BDI-V, Schmitt & Maes, 2000) gemessen. Dieses Messinstrument umfasst 20 Items wie z.B. „Ich bin traurig." oder „Mir fehlt das Interesse an Menschen" (die vollständigen Items sind Tabelle 5.1 zu entnehmen). Das Antwortformat des BDI-V ist eine sechsstufige Häufigkeitsskala, deren Extrempunkte sprachlich mit „nie" (0) und „fast immer" (5) gekennzeichnet sind. Untersuchungen zur Messgüte des Instruments ergaben gute bis sehr gute psychometrische Kennwerte, die bei höherer Messökonomie gleich gut sind wie die des Original-BDI (Schmitt et. al., 2003). Durch ein Missgeschick seitens des Autors wurde dabei eine ältere Version des BDI-V verwendet (Schmitt, Maes & Schmal, 1999). Diese ist mit der von Heckmann (2008) verwendeten identisch, enthält aber noch zusätzlich am Ende die beiden Items: „Ich bin des Lebens überdrüssig." sowie „Ich sehne mich nach dem Tod."

4.3 Untersuchungsdurchführung

4.3.1 Die Durchführung der Onlinebefragung

Der erste Erhebungszeitraum erstreckte sich vom 1. Juli bis zum 31. August des Jahres 2010. Die Teilnehmer erhielten das Anschreiben zum 2. Teil der Studie exakt drei Monate nach dem Tag, an dem sie den Fragebogen zum ersten Zeitpunkt abgeschickt hatten. Der zweite Untersuchungszeitraum erstreckte sich daher vom 1. Oktober bis zum 1. Dezember des Jahres 2010 (am 1. Dezember wurden die Teilnehmer angeschrieben, die am 31. August ihren ersten Fragebogen abgeschickt hatten). Zur Erstellung des Fragebogens sowie zur Durchführung der Onlinebefragung wurde die Open Source Software LimeSurvey Version 1.86 (2010) verwendet. Die Befragung wurde als reine Onlineuntersuchung durchgeführt, da die Erfahrungen der Voruntersuchung von Heckmann gezeigt hatten, dass auf diese Art und Weise eine genügend große Anzahl an Versuchsteilnehmern gewonnen werden kann. Die Ausgabe von ausgedruckten Fragebögen wurde daher zur Vermeidung des zusätzlichen organisatorischen und finanziellen Aufwands unterlassen.

Die Teilnehmerinnen und Teilnehmer erhielten entweder durch die Moderatoren ihrer Internetforen (*yahoo!*- und *Google*-Gruppen) oder durch den Autor der Studie (andere Teilnehmer) den durch LimSurvey generierten Hyperlink zur Studie. Klickten sie diesen an, so wurden sie zum Fragebogen weitergeleitet, der auf dem Server des Methodenzentrums der Universität in Landau lag. Zwei Tage vor ihrem individuellen zweiten Messzeitpunkt wurden die Versuchspersonen per E-Mail darüber informiert, dass sie in zwei Tagen den Link zum zweiten Teil der Studie bekommen würden. In diesem Anschreiben wurden die Teilnehmerinnen und Teilnehmer darum ersucht, den Fragebogen nach Möglichkeit noch am gleichen Tag, an dem sie den Link geschickt bekommen auszufüllen. Sofern zwei Tage nach der Einladung zur Teilnahme am zweiten Messzeitpunkt kein Eingang des Fragebogens vermerkt werden konnte, wurde den betreffenden Versuchspersonen eine Erinnerungsnachricht per E-Mail geschickt, die auch nochmals den Hyperlink zur Studie enthielt.

Die erste Seite des Onlinefragebogens enthielt eine kurze Erklärung zum Ziel der Studie, eine Erläuterung des Ablaufs und die Instruktion zur Beantwortung der Fragen. Des Weiteren enthielt die Startseite Hinweise zum Datenschutz sowie die Auslobung der *Amazon*-Gutscheine. Durch Klick auf den Button „Weiter" gelangten die Teilnehmenden jeweils zur nächsten Seite. Seite 2 enthielt die ersten 11 Items des Vereinfachten Beck Depression Inventars (BDI-V, Schmitt & Maes, 2000), Seite 3 die Items 12-20 sowie die in Kapitel 4.2 erwähnten Zusatzitems. Die technischen Einstellungen in LimSurvey wurden so gewählt, dass zum Messzeitpunkt 1 ab Seite 2 jede folgende Seite erst dann für die Versuchsperson erreichbar war, wenn die vorangegangenen Seiten vollständig ausgefüllt waren. Die nach Seite 3 folgenden Seiten enthielten jeweils pro Seite eine Abfrage einer soziodemographischen Variablen (Alter, Geschlecht, höchster Bildungsabschluss, Tätigkeit und Familienstand). Im Anschluss wurde ein Code[1] abgefragt mittels dessen das Datenset einer Person zu einem Messzeitpunkt dem des anderen Messzeitpunkts zugeordnet werden konnte. Im Anschluss wurden die Teilnehmerinnen und Teilnehmer gebeten, ihre E-Mailadresse anzugeben. Dies war notwendig, damit die Erinnerungsnachricht sowie die eigentliche Einladung, die den Link zur zweiten Untersuchung enthielt verschickt werden konnte, auch wenn dadurch die Anonymität der Versuchsteilnehmer nicht in vollem Umfang gewährleistet werden konnte. Der Onlinefragebogen des ersten Messzeitpunktes endete an diesem Punkt. Anders als bei Heckmann (2008, S. 30) wurde nur zum zweiten Messzeitpunkt die Kontrollfrage gestellt. Diese lautete: „Auf welchen Zeitraum haben Sie sich bei den Angaben zu Ihrem Lebensgefühl bezogen?". Des Weiteren wurden die Versuchspersonen gefragt, was ihrer Meinung nach Hauptgegenstand und Ziel der Untersuchung sei. Schließlich wurde den Teilnehmenden die Möglichkeit gegeben, einen Kommentar zur Studie abzugeben. Die Beantwortung der Fragebögen dürfte in etwa fünf bis maximal zehn Minuten in Anspruch genommen haben.

[1] Dieser fünfstellige Code setzte sich aus dem Vornamen der Mutter, dem Vornamen des Vaters, dem Geburtsort der Versuchsperson sowie dem Monatstag des Geburtstages der Personen zusammen (der letztgenannte Wert sollte zweistellig angegeben werden, also z.B. „01" für den 1. des Monats)

4.3.2 Umgang mit hochdepressiven Personen

Der Autor sah eine ethische Verpflichtung darin, Personen, die einen hohen BDI-V-Summenwert erzielten darauf hinzuweisen, dass bei ihnen unter Umständen eine klinisch relevante Erkrankung vorliegt. Schmitt, Altstötter-Gleich, Hinz, Maes & Brähler (2006) ermittelten 35 als kritischen Wert, ab dem von einer 90%igen Wahrscheinlichkeit für eine klinisch relevante Erkrankung ausgegangen werden muss. Darüber hinaus stellten sie Geschlechtsunterschiede für das BDI-V fest. Für diese Studie wurden leicht höhere, nach Geschlecht differenzierte kritische Werte verwendet. Überschritten weibliche Versuchsteilnehmer den Wert von 40 und männliche den Wert von 38, erhielten die betroffenen Versuchspersonen per E-Mail eine Rückmeldung über ihr Testresultat, in der sie darüber unterrichtet wurden, dass das verwendete Messinstrument Depressivität misst und dass der Wert, den erzielt hatten ein möglicher Hinweis auf eine klinisch relevante Erkrankung sei. Den Personen wurde nahegelegt für den Fall, dass die Symptome länger andauern therapeutische Hilfe in Anspruch zu nehmen. Das Item Nr. 9 des BDI-V („Ich denke daran, mir etwas anzutun.“) sowie die beiden zusätzlich erhobenen Items („Ich bin des Lebens überdrüssig.“ und „Ich sehne mich nach dem Tod.“), die nicht in die Berechnung des BDI-Summenwertes einflossen wurden extra berücksichtigt. Versuchspersonen, die bei mindestens zwei dieser drei Items die Zahlen 4 oder 5 ankreuzten wurden ebenfalls kontaktiert. Da in diesem Fall akute Suizidalität befürchtet wurde, wurden die Versuchspersonen gebeten, den Autor telefonisch zu kontaktieren. Dies betraf 10 Versuchspersonen, von denen 4 das Gesprächsangebot annahmen. Diese konnten sich glaubhaft von akuter Suizidalität distanzieren, so dass keine weiteren Schritte unternommen werden mussten. Beide beschriebenen Interventionsarten wurden zu beiden Messzeitpunkte durchgeführt. Zum ersten Erhebungszeitpunkt wurden so insgesamt 30 Personen pro Instruktionsgruppe, also insgesamt 60 Personen kontaktiert. Von diesen nahmen 22 aus der 14-Tage-Gruppe und 18 aus der 3-Monats-Gruppe am zweiten Erhebungszeitpunkt teil. Im zweiten Erhebungszeitraum wurden insgesamt 48 Personen kontaktiert.

4.4 Auswertungsmethode

In diesem Kapitel wird die statistische Umsetzung der Hypothesen in Latent-State-Trait-Modelle näher beschrieben. Im ersten Unterkapitel (4.4.1) werden dabei die Modelle erläutert, die für die 3-Monats- sowie die 14-Tage-Gruppe jeweils einzeln gerechnet wurden. Gegenstand des zweiten Unterkapitels sind dann die Modelle, die zum Vergleich zwischen den Gruppen herangezogen wurden. Die LISREL-Syntax für die unterschiedlichen Modelle kann Anhang C entnommen werden.

4.4.1 Einzelgruppenmodell

In diesem Unterkapitel soll das Modell der Latent-State-Trait-Theorie mit zwei Methodenfaktoren vorgestellt werden, welches die Grundlage für die jene Analysen bot, die für die 3-Monats- und 14-Tage-Gruppe getrennt gerechnet wurden. Dieses Modell hatte sich in den Vorstudien zum BDI bewährt (Mohiyeddini, Hautzinger & Baer, 2002; Schmitt & Maes, 2000) und war ebenfalls Grundlage der Vorgängerstudie von Heckmann (2008, Kapitel 4.4.1). Es ist in Abbildung 4.1 graphisch dargestellt.

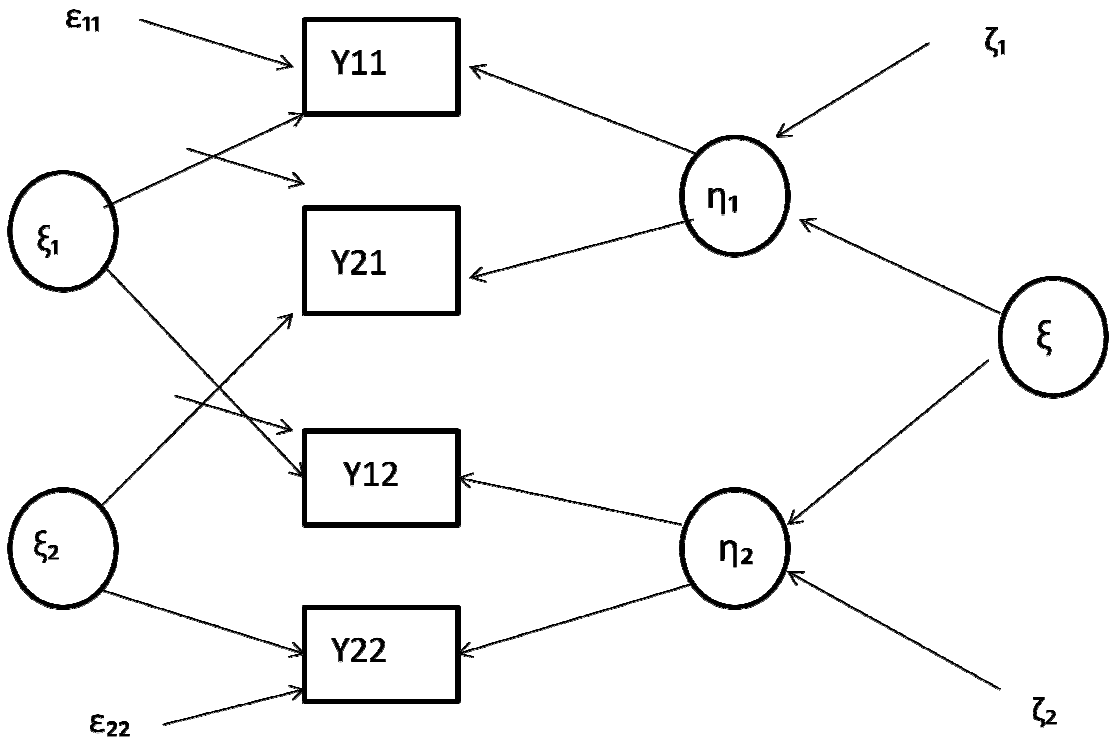


Abbildung 4.1: Latent-State-Trait-Modell mit 2 Methodenfaktoren

Zur Erläuterung des Modells: Y_{ik} bezeichnet den Wert der manifesten Variable, die mit der Testhälfte i zum Zeitpunkt k erhoben wurde. Mit η_k sind die latenten State-Variablen, mit ζ_k die Latent-State-Residuen zum Zeitpunkt k bezeichnet. ξ ist in diesem Modell das Symbol für den Traitfaktor. Der Messfehler der Testhälfte i zum Messzeitpunkt k ist jeweils mit ε_{ik} gekennzeichnet. Schließlich symbolisiert ξ_{ik} in diesem Modell den jeweiligen Methodenfaktor der Testhälfte i. Alle Faktorladungen dieses Modells sind auf den Wert eins gesetzt. Wie bereits in Kapitel 2.3 erläutert stellt in diesem Modell die latente State-Variable den wahren Wert im Sinne der klassischen Testtheorie dar. Die latente State-Variable wird im Rahmen der LST-Theorie zusätzlich in die latente Traitvariable und das Latent-State-Residuum dekomponiert. Die latente Trait-Variable steht dabei für den Einfluss der Persönlichkeit auf die State-Variable und wird als stabil angesehen. Das Latent-State-Residuum steht für den Einfluss der Situation, in der sich die Person zum jeweiligen Erhebungszeitpunkt befindet und gleichzeitig für die Interaktion aus Persönlichkeit und Situation. Die Effekte dieser Variable sind systematisch, aber nicht stabil. Die Methodenfaktoren schließlich geben den Einfluss der unterschiedlichen Testhälften auf die manifeste Variablen wieder. Der Einfluss der Testhälften wird als über die Messzeitpunkte hinweg stabil angesehen. Schließlich beschreiben die latenten Messfehlervariablen den Messfehler der Testhälften sowie alle weiteren unsystematischen Varianzquellen.

In Anlehnung an Heckmann (2008, S. 32) wurde im Rahmen dieser Arbeit das Modell in seiner restriktivsten Variante getestet. Dies bedeutet, dass die Varianzen der zwei Latent-State-Residuen, die Varianzen der Methodenfaktoren sowie die Varianzen der Messfehler gleichgesetzt wurden. Die Latent-State-Trait-Koeffizienten für beide Testhälften wurden ebenfalls gleichgesetzt. Folglich beträgt für die Einzelgruppenmodelle die Zahl der zu schätzenden Parameter, anders als es das Modell in der präsentierten Grundform implizieren würde, nicht neun, sondern lediglich vier. Die Anzahl der Freiheitsgrad liegt dementsprechend bei sechs, gegenüber einem in der Grundform. Die folgenden Parameter sind zu schätzen: die Varianzen der latenten Traitvariable, der Latent-State-Residuen, der Methodenfaktoren sowie der Messfehler. Sofern das spezifizierte Modell auf die erhobenen Daten passt,

lassen sich für die Messwertvariablen (hier: die Testhälften) die folgenden Koeffizienten errechnen (Steyer et. al. 1999; Steyer, Ferring & Schmitt, 1992). Die *Traitkonsistenz* (TKon) gibt den Varianzanteil der Messwertvariable (Y_{ik}) an, der durch den Trait bestimmt wird [Var(ξ)/Var(Y_{ik})]. Als *Messgelegenheitsspezifität* oder *Zeitspezifität* (ZSpe) bezeichnet man den Varianzanteil, der durch die Latent-State-Residuen erklärt wird. [Var (ζ_k)/Var (Y_{ik})]. Der Varianzanteil der Methodenfaktoren [Var(ξ_i)/Var(Y_{ik})] wird *Methodenspezifität* (MSpe) genannt. Schließlich ergibt sich die *Reliabilität* (Rel) als der systematische Varianzanteil von Y_{ik} durch Addition von TKon, ZSpe und MSpe. Die soeben beschriebenen Koeffizienten beziehen sich lediglich auf die *Testhälften.* Es besteht allerdings die Möglichkeit, diese mit Hilfe der folgenden Gleichungen, die an die Spearman-Brown-Formel für die Reliabilität angelehnt sind, auf den Gesamttest hochzurechnen (Steyer & Eid, 1993). Gleichung 1 stellt die Formel für die Traitkonsistenz dar, Gleichung 2 jene für die Zeitspezifität (Steyer & Schmitt, 1990, S. 85). Die dritte Gleichung bietet die Möglichkeit, die Reliabilität für den Gesamttest zu berechnen (Steyer & Eid, 1993, S. 146), Gleichung 4 ermöglicht dies für die Methodenspezifität (Heckmann, 2008, S. 32).

(1) TKon (Test) = $\frac{2TKon(Hälfte)}{1+TKon(Hälfte)+ ZSpe(Hälfte)}$

(2) ZSpe (Test) = $\frac{2ZSpeHälfte)}{1+TKon(Hälfte)+ ZSpe(Hälfte)}$

(3) Rel (Test) = $\frac{2Rel(Hälfte)}{1+Rel(Hälfte)}$

(4) MSpe(Test) = *Rel(Test) – TKon(Test) – ZSpe(Test)*

4.4.2 Mehrgruppenvergleiche

An die Analyse der Einzelgruppenmodelle, welche für beide Instruktionsgruppen getrennt durchzuführen ist, schließen sich Mehrgruppenvergleiche (Multi-sample-Analysen) an. Mittels dieser Analyseart ist es möglich, ein Strukturgleichungsmodell an den Daten mehrerer Stichproben gleichzeitig zu testen. In diesem Fall wurden beide Instruktionsgruppen simultan getestet. Diese Methode ermöglicht es, zu prüfen, ob Unterschiede zwischen den Instruktionsgruppen hinsichtlich der zu schätzenden Parameter bestehen. Ein weiterer Vorteil der Multi-sample-Analysen liegt in der Tatsache, dass sie es ermöglichen, eventuell vorgefundene Unterschiede hinsichtlich der Parameter durch Signifikanztests abzusichern, während die Einzelgruppenmodelle nur deskriptiv durch Vergleich der einzelnen Kennwerte des Modellfits verglichen werden können.

Die im Folgenden vorzustellenden Modelle der Mehrgruppenvergleiche entsprechen jenen, die in der Vorgängerarbeit verwendet wurden (Heckmann, 2008, S. 33/34). Das erste Modell zeichnet sich dadurch aus, dass alle Parameter zwischen den Instruktionsgruppen gleichgesetzt sind (im Folgenden als „Invarianzmodell“ bezeichnet). Gruppenunterschiede hinsichtlich der zu schätzenden Parameter sind also im Rahmen dieses Modells nicht zugelassen.

Zum Vergleich wird in Schritt 2 ein Modell geprüft, bei dem die Restriktionen für die Traitvarianz (ξ) und die Latent-State-Residuen (ζ) gelockert werden. Die beiden gewählten Parameter sind genau jene, in denen sich der angenommene Einfluss der zeitlichen Instruktion widerspiegeln würde. Das Modell impliziert für sie gleiche Ladungsmuster und Startwerte, die endgültigen Werte der Parameterschätzung können aber zwischen den Gruppen unterschiedlich ausfallen. Die in Kapitel 4.4.1 beschriebenen Formeln für die Latent-State-Trait-Koeffizienten implizieren, dass Traitkonsistenz und Latent-State-Residuum einander wechselseitig bedingen. Das bedeutet: sofern die Reliabilität und der Einfluss des Methodenfaktors konstant sind, geht eine Veränderung der Traitkonsistenz mit einer Veränderung des Latent-State-Residuums in entgegengesetzter Richtung einher und umgekehrt. Aus diesem Grund

werden beide Parameter in einem Schritt gelockert. Sollte das in Schritt 2 gerechnete Modell signifikant besser auf die Daten der beiden Instruktionsgruppen passen als das Invarianzmodell, so würde dies für einen Einfluss der zeitlichen Instruktion auf Traitkonsistenz und Messgelegenheitsspezifität sprechen.

In Schritt 3 wird schließlich ein Modell getestet, bei dem zusätzlich zu den in Schritt 2 gelockerten Parametern auch die Varianz der Methodenfaktoren und die Messfehlervarianz bei gleichem Ladungsmuster und gleichen Startwerten zwischen den Gruppen variieren dürfen. Die Parameter werden also unabhängig voneinander geschätzt (im Folgenden wird das Modell daher „Unabhängigkeitsmodell" genannt). Angesichts der Tatsache, dass im Rahmen dieser Studie davon ausgegangen wird, dass sich die beiden Gruppen lediglich hinsichtlich der zeitlichen Instruktionen für die Beantwortung des BDI-V unterscheiden, sollte die zusätzliche Aufhebung der Parameterrestriktionen in Schritt 3 keine signifikante Verbesserung des Modellfits erbringen. Sollte dies dennoch der Fall, so ist dies ein Hinweis darauf, dass sich die beiden Instruktionsgruppen in Bezug auf Merkmale unterscheiden, welche nicht auf die unterschiedlichen zeitlichen Instruktionen zurückzuführen sind. In diesem Fall müsste auch das Ergebnis des zuerst angestellten Modellvergleichs in Frage gestellt werden, da dann nicht ausgeschlossen werden könnte, dass es auf Unterschieden zwischen den Gruppen beruht, die nicht durch die Hypothesen dieser Arbeit abgedeckt sind und z.B. auf der Stichprobenziehung beruhen.

Die drei soeben präsentierten Mehrgruppenmodelle sind hierarchisch ineinander geschachtelt, so dass es möglich ist, Unterschiede hinsichtlich des Modellfits durch einen Chi^2-Modelldifferenzentest auf Signifikanz zu testen (siehe auch Heckmann, 2008, S. 34).

4.5. Bildung der Testhälften

Die in den vorangegangenen Abschnitten präsentierten Modelle werden mittels einer Latent-State-Trait-Analyse ausgewertet (Steyer et. al., 1992) ausgewertet. Für diese ist es erforderlich, dass für die zu analysierenden latenten States mindestens zwei manifeste Indikatoren pro Messzeitpunkt vorliegen. In Vorgängerstudien zum BDI-V (Schmitt & Maes, 2000; Heckmann, 2008) wurde das Messinstrument aus diesem Grund in zwei Testhälften aufgeteilt. Das Zusammenfassen von Items zu Testhälften oder ähnlichen Päckchen ist bei der Anwendung von Strukturgleichungsmodellen weit verbreitet (Hau & Marsh, 2004). Wenngleich dieses Vorgehen keineswegs unumstritten ist (Little, Cunningham, Shahar & Widaman 2002), erscheint es im Rahmen dieser Arbeit vor allem aus zwei Gründen sinnvoll, wenn nicht gar unumgänglich. Als erstes zu nennen wäre die Stichprobengröße: Eid (1995) beziffert den Bedarf an Versuchspersonen bereits für eine Latent-State-Trait-Analyse mit zwei Messzeitpunkten und sechs Items auf 500. Werden Items zu Päckchen zusammengefasst sind weniger Parameter zu schätzen und die Schätzungen für kleinere Stichproben sind stabiler (West, Finch & Curran, 1995). Der zweite Grund ist die in den erwähnten Vorgängerstudien (Schmitt & Maes, 2000; Heckmann, 2008) festgestellte Rechtsschiefe der Items des BDI-V. Da annähernde Normalverteilung eine Voraussetzung für das gewählte Parameterschätzverfahren ist spricht auch dies gegen eine Verwendung der Einzelitems als Indikatoren der Depressivität. Itempäckchen hingegen weisen in der Regel weitaus geringere Abweichungen von der Normalverteilung als die ihnen zugrunde liegenden Einzelitems auf (West et. al., 1995). Eine Möglichkeit der Zusammenfassung von Items ist dabei die Summenbildung (Marsh, Antill & Cunningham, 1989). Aus diesem Grund wurde die in den zitierten Vorgängerarbeiten gewählte Methode, die Items des BDI-V in zwei Testhälften aufzuteilen, die Itemresultate aufzusummieren und die Summenwerte der beiden Testhälften als Indikatoren der Depressivität für die beiden Zeitpunkte zu wählen, übernommen.

Um eine möglichst gute Vergleichbarkeit der Ergebnisse dieser Studie mit denen der Arbeit von Heckmann zu gewährleisten, wurden zunächst die Testhälften von

Heckmann (2008, S. 35/36 u. 43) übernommen. Unter Verwendung dieser Testhälften konnten für die beiden Einzelgruppenmodelle (siehe Kapitel 4.4.1) akzeptable bis gute Modellkennwerte erzielt werden. Bei den Mehrgruppenmodellen (siehe Kapitel 4.4.2) hingegen waren die errechneten Fitstatistiken weit jenseits des Bereichs, der als akzeptabel angesehen werden. Vor dem Hintergrund, dass die aufgestellten Modelle eine ausgezeichnete Möglichkeit bieten, die aufgestellten Hypothesen zu testen und sich in der Arbeit von Heckmann bestens bewährt hatten wurden anstelle einer Modifikation der Modelle zunächst intensiv Versuche unternommen, möglichst gut parallele Testhälften für die erhobenen BDI-V-Daten der beiden Instruktionsgruppen zu finden. Schmitt und Maes (2000) verwendeten in ihrer Arbeit nach der Odd-even-Methode gebildete Testhälften. Diese wurden als erstes geprüft. Doch auch mit diesen Testhälften konnten keine akzeptablen Modellfits erzielt werden: sowohl für die beiden Einzelgruppenmodelle als auch für die Mehrgruppenmodelle erwiesen sie sich im Zusammenspiel mit den erhobenen Daten dieser Studie als ungeeignet. Unter Verwendung der split-half-Methode (erste 10 Items bilden Testhälfte 1, die zweiten 10 Testhälfte 2) wurden gute Modellkennwerte für die Einzelgruppenmodelle erzielt, jedoch nicht für die Mehrgruppenmodelle. Little et. al. (2002) nennen die Zuweisung der Items zu den Testhälften per Zufall als eine empfehlenswerte Methode zur Gewinnung von Itempäckchen. Aus diesem Grund wurden auch Versuche in diese Richtung unternommen. So wurden unter Verwendung der Onlinesoftware Random Sequence Generator (1998) per Zufall zwei Spalten erzeugt, auf die sich die Zahlen 1-20 aufteilten. Die Zahlen wurden als gleichbedeutend mit denen der Nummerierung des BDI-V interpretiert, so dass auf diese Art zwei Testhälften gewonnen wurden. Diese Prozedur wurde mehrfach wiederholt. Leider ergaben die Latent-State-Trait-Analysen, die mit diesen Testhälften gerechnet wurden ebenfalls keine akzeptable Passung auf die erhobenen Daten. Dies galt hierbei in allen gerechneten Fällen schon auf der Ebene der Einzelgruppenmodelle. Schließlich wurde in leichter Abwandlung die von Heckmann (2008, S. 35/36) angewandte Methode der Testhälftengenerierung nachvollzogen. Heckmann hatte die Items anhand der Daten der 14-Tage-Gruppe nach Mittelwert sortiert in eine Rangreihe gebracht und im Anschluss die Items den Testhälften so zugeordnet, dass die durchschnittlichen Mittelwerte der Items in den

beiden Testhälften möglichst nahe beieinander lagen. Im Anschluss hatte sie geprüft, ob die durchschnittliche Itemvarianz der beiden Testhälften ebenfalls hinreichend ähnlich ist. Zur Generierung der Testhälften für diese Arbeit wurden die Items des BDI-V ebenfalls unter Rückgriff auf die Daten der 3-Monats-Gruppe in eine Rangreihe gebracht. Anders als bei Heckmann wurde dabei allerdings nicht der Mittelwert der Items als Maßstab verwendet, sondern die Itemvarianz. Im Anschluss wurde eine Kombination ermittelt, welche die Items zu zwei Testhälften bündelt, die eine möglichst gleich große durchschnittliche Itemvarianz aufwies. Danach wurde geprüft, ob der durchschnittliche Mittelwert ebenfalls in etwa gleich groß ist. Die auf diese Art erzielte Kombination lautete für Testhälfte 1: Items 5, 6, 8, 9, 10, 13, 14, 15, 17, 19; für Testhälfte 2: Items 1, 2, 3, 4, 7, 11, 12, 16, 18, 20 (die Nummerierung entspricht der im BDI-V). Schließlich wurde untersucht, ob für diese Testhälften auch in der 14-Tage-Gruppe die durchschnittliche Itemvarianz und der durchschnittliche Itemmittelwert in etwa gleich groß sind. Schlussendlich wurden unter Verwendung dieser Testhälften die Latent-State-Trait-Analysen in einem Probelauf gerechnet. Die im Rahmen dieses Probelaufs erzielten Modellkennwerte erschienen in der Gesamtschau den zuvor getesteten Varianten überlegen. Daher wurden die oben beschriebenen Testhälften als Grundlage für die Analysen im Rahmen dieser Arbeit verwendet.

4.6 Parameterschätzung und Modelltests

Dieser Abschnitt widmet sich zunächst der Beschreibung der in dieser Arbeit verwendeten Methode der Parameterschätzung. Daran schließen sich Erläuterungen zu den Kennwerten an, die herangezogen wurden, um die Modellgüte zu beurteilen.

4.6.1 Software und Parameterschätzung

Für die Parameterschätzungen und zur Beurteilung der Modellgüte wurde der LISREL 8.80 Student Edition (Jöreskog & Sörbom, 2006a) verwendet. Als Schätzverfahren wurde in Anlehnung an Heckmann (2008, S. 36) die *Maximum-Likelihood-Methode* (ML) gewählt. Diese ist die am häufigsten angewandte Methode (Moosbrugger & Schermelleh-Engel, 2007, S. 319) und sie bietet den Vorteil, dass sie Inferenzstatistiken

zur Verfügung stellt (Backhaus, Erichson, Plinke & Weiber, 2006). Es ergeben sich zwei Voraussetzungen für die Anwendung von ML: der Stichprobenumfang sollte mindestens 100 betragen (Lei & Lomax, 2005; Backhaus et. al., 2006) und die manifesten Variablen sollten multinormalverteilt sein. (Moosbrugger & Schermelleh-Engel, 2007, S. 319). In dieser Studie wurde die genannte Mindestgröße der Stichprobe für beide Instruktionsgruppen erreicht. Die Voraussetzung der Multinormalverteilung der Testhälften wurde mit der PRELIS Version 2.80 Student Edition (Jöreskog & Sörbom, 2006b) geprüft. Hinsichtlich der univariaten Verteilungseigenschaften der Testhälften ergaben sich für Schiefe Werte zwischen 0.40 und 0.92 (*p*-Werte: .000 bis .026) und für Exzess zwischen -0.51 und +0.73 (*p*-Werte: .050 bis .854). Der multivariate Test ergab in der 14-Tage-Gruppe Werte von 2.96 für Schiefe und 29.17 (beide $p < .001$) für Exzess, in der 3-Monatsgruppe ergab sich für Schiefe der Wert 3.37 ($p < .001$), für Exzess 28.13 ($p = .001$). Die Abweichungen von der Normalverteilung erreichen also statistische Signifikanz. Die Frage, ab wann Abweichungen von der Normalverteilung als kritisch anzusehen sind und daher korrektive Maßnahmen ergriffen werden müssen, ist in der Literatur nicht vollständig geklärt. Es existieren keinerlei Richtlinien. Einige Autoren geben zwei als kritischen Wert für univariate Schiefe und sieben für univariaten Exzess an (Curran, West & Finch, 1996; West, Finch & Curran, 1995). Diese Werte wurden in dieser Studie deutlich unterschritten. Die ML-Methode gilt zudem als relativ robust gegenüber Verletzungen der Normalverteilung (z.B. Chou & Bentler, 1995). Lei & Lomax (2005) kommen bei ihren Simulationsstudien zu dem Ergebnis, dass selbst bei einer heftigen Abweichung von der Normalverteilung die maximale Abweichung der Parameterschätzung unter Einsatz von ML deutlich unter 10% gelegen habe und schließen daraus: "Therefore, the usual interpretation of SEM parameter estimates can be accepted, even under the severe nonnormality conditions." (S. 16). Vor diesem Hintergrund erscheint die Anwendung von ML im Rahmen dieser Studie als gerechtfertigt.

4.6.2 Beurteilung der Modellgüte

Für die Bewertung der Modellgüte von Strukturgleichungsmodellen war es lange Zeit üblich, diese ausschließlich anhand des Chi^2-Tests zu beurteilen. Da der Chi^2-Test

allerdings als sehr empfindlich gegenüber der Stichprobengröße gilt, geriet diese Praxis in die Kritik (Saris, Satorra & van der Veld, 2009). So wurden eine Reihe weiterer Fitindizes entwickelt und Programme wie LISREL (Jöreskog & Sörbom, 2006a) bieten mittlerweile eine Vielzahl solcher Modellkennwerte an (Saris, Satorra & van der Veld, 2009). Zur Beurteilung der Modellgüte dieser Arbeit werden die Empfehlungen von Schermelleh-Engel, Moosbrugger & Müller (2003) übernommen, denen auch Heckmann in ihrer Arbeit folgte (Heckmann, 2008, S. 37). Diese umfassen die folgenden Kennwerte: *Chi*² (χ^2), dessen *p*-Wert, χ^2/df, den Root Mean Square Error of Approximation (*RMSEA)*, dessen Konfidenzintervall, das standardisierte Root Mean Square Residual (*SRMR*), den Nonnormed Fit Index (*NNFI*) und den Comparative Fit Index (*CFI*). Aufgrund seiner Eignung für Modellvergleiche wird außerdem das Akaike Information Criterion (*AIC*) berichtet. Die Erweiterungen der Empfehlung von Schermelleh-Engel et. al. (2003), die Heckmann (2008, S. 38) in ihrer Arbeit vornahm werden ebenfalls übernommen, so dass auch der Goodness-of-Fit-Index (*GFI*) berichtet wird, da dieser Kennwert bei Berechnung der Multi-Sample-Analysen für beide untersuchten Gruppen zur Verfügung gestellt wird.

Der χ^2-Test bietet die Möglichkeit einer inferenzstatistischen Überprüfung der Nullhypothese, alle übrigen Indizes sind deskriptive Gütekennwerte. *RMSEA* und *SRMR* sind Kennwerte für das Gesamtmodell. *NFFI*, *CFI* und *GFI* sind Indizes, welche auf Modellvergleichen zwischen dem zu untersuchenden Modell und einem Vergleichsmodell fußen. Mittels des *AIC* kann die Sparsamkeit des Modells einer Beurteilung unterzogen werden.

Der χ^2-Test überprüft, sofern alle Verteilungsannahmen erfüllt sind, ob die empirische Kovarianzmatrix mit der modellimplizierten Kovarianzmatrix übereinstimmt (Saris, Satorra & van der Veld, 2009). Sofern der *p*-Wert des χ^2-Tests größer als .05 ist, wird die Nullhypothese, die eine Übereinstimmung von empirischer Kovarianzmatrix und modellimplizierter annimmt, beibehalten. Der χ^2-Test ist allerdings stark von der Stichprobengröße abhängig: "the decision for accepting or rejecting a particular model may vary as a function of sample size, which is certainly not desirable" (Hu & Bentler, 1998, S. 429). Aus diesem Grund und da der χ^2-Test auch sensibel auf eine Verletzung

der Voraussetzung der Multinormalverteilung der Variablen reagiert, empfehlen Jöreskog & Sörbom (1993), ihn eher als rein deskriptiven Gütekennwert anzusehen. Des Weiteren schlagen die Autoren vor, den Quotienten χ^2/df als Gütekennwert zu verwenden. Werte unterhalb von zwei betrachten sie dabei als gut, Werte unter drei seien als akzeptabel anzusehen.

Der *RMSEA* (Steiger, 1990) prüft eine Nullhypothese *näherungsweisen* Fits des Modells in der Population (Brown & Cudeck, 1993; Steiger, 2000). *RMSEA*-Werte zwischen 0 und .05 weisen auf eine guten Fit, Werte zwischen .05 und .08 auf einen mittelmäßigen, akzeptablen Fit und Werte größer als .10 auf einen schlechten Modellfit hin (Kelley & Lai, 2011; MacCallum, Browne & Sugawara, 1996). Da in der angewandten Forschung davon auszugehen ist, dass die Parameterschätzungen für die untersuchte Stichprobe von den korrespondierenden Parametern in der Population abweichen ist es sinnvoll, die Konfidenzintervalle für den *RMSEA* ebenfalls zur Beurteilung der Modellgüte ins Kalkül zu ziehen (Kelley & Lai, 2011). Schermelleh-Engel et. al. (2003) bezeichnen 0 als den Wert für die untere Grenze des Konfidenintervalls, der von einem guten Modell sprechen lässt und Werte kleiner .05 als untere Grenze für einen als akzeptabel zu bezeichnenden Modellfit.

Heckmann (2008) hatte sich in der Vorgängerarbeit gegen das *Root Mean Square Residual* entschieden, da dieses zur Bewertung des Modellfits nicht ohne die Berücksichtigung der Skalierung der Variablen herangezogen werden könne (S. 39) und sich stattdessen für das von Bentler (1995, zitiert nach Heckmann, 2008, S. 39) vorgeschlagene *SRMR* entschieden. Dieses Vorgehen wird übernommen. Für das *SRMR* gilt: Werte zwischen 0 und .05 sprechen für einen guten Modellfit, Werte zwischen .05 und .10 sind als akzeptabel anzusehen (Schermelleh-Engel et. al. 2003, zitiert nach Heckmann, 2008, S. 39).

Der *NNFI* (Bentler & Bonett, 1980; Tucker & Lewis, 1973) vergleicht das aufgestellte Modell mit dem so genannten Unabhängigkeitsmodell. Dieses basiert auf der Grundannahme einer messfehlerfreien Messung aller manifesten Variablen, was zur Folge hat, dass alle Fehlervarianzen auf null und alle Faktorladung auf eins fixiert sind.

Zudem wird davon ausgegangen, dass alle Variablen nicht untereinander korrelieren. Im Allgemeinen nimmt der *NNFI* Werte zwischen null und eins an, wobei er diesen Korridor unter Umständen verlassen kann, da er nicht normiert ist. Höhere Werte zeigen eine bessere Modellpassung an. Ein akzeptabler Modellfit kann hierbei für Werte größer als 0.95 angenommen werden, ein guter für Werte größer 0.97 (Schermelleh-Engel et. al., 2003).

Bei der Berechnung des *CFI* werden die χ^2-Werte sowie die Freiheitsgrade des Zielmodells zu den korrespondierenden Werten des Vergleichsmodells in Beziehung gesetzt, wobei die χ^2-Werte durch einen Nonzentralitätsparameter an eine nichtzentrale χ^2-Verteilung angenähert werden. Vergleichsmodell ist hierbei das für den *NNFI* bereits beschriebene Unabhängigkeitsmodell (Schermelleh-Engel & Moosbrugger, 2002). Der *CFI* kann Werte zwischen null und eins annehmen, wobei Werte über .95 auf einen akzeptablen und Werte über .97 auf einen guten Modellfit hinweisen.

Der *GFI* wurde 1984 von Jöreskog und Sörbom vorgeschlagen. Die Idee des *GFI*-Index basiert auf dem gleichen Prinzip wie der Determinationskoeffizient in der Regressionsanalyse (MacCallum & Hong, 1997; Mulaik et. al., 1989): "The GFI can be thought of as 1 minus the ratio of residuals (weighted) sum of squares to total (weighted) sum of squares." (MacCallum & Hong, 1997, S. 200). Der bestmögliche Wert für den *GFI* ist damit eins. MacCallum & Hong berichten, dass lange Zeit .90 als gängiger Cut-off-Wert für den *GFI* gesehen wurde, dass Hu & Bentler (1995, zitiert nach MacCallum & Hong, 1997, S. 209) bei Vorliegen bestimmter Bedingungen dieses Kriterium als zu restriktiv, unter anderen Bedingungen als zu permissiv bezeichnet hätten. Schumacker & Lomax schlugen 1996 vor, Werte größer als .95 als Hinweis auf einen guten Modellfit anzusehen.

Mit Hilfe des *AIC* (Akaike, 1974, 1987) ist es möglich, Modelle, welche auf der gleichen Kovarianzmatrix beruhen, aber nicht hierarchisch geschachtelt sind, deskriptiv miteinander zu vergleichen. Das *AIC* bevorzugt sparsame Modelle, da bei der Berechnung des Werts für das Zielmodell der χ^2-Wert um die Zahl der schätzenden

Parameter angepasst wird, wodurch komplexere Modell bestraft werden. Als das beste Modell wird dabei das Modell beurteilt, welches den kleinsten *AIC*-Wert aufweist (Schermelleh-Engel et. al., 2003). LISREL gibt neben dem beschriebenen Zielmodell auch Werte für das Unabhängigkeitsmodell und das saturierte Modell (null Freiheitsgrade) aus. Tabelle 4.1 bietet eine Übersicht über die soeben beschriebenen Modellgütekennwerte und die Wertebereiche innerhalb derer der betreffende Index für einen akzeptablen bzw. guten Modellfit spricht. Tabelle 4.1 wurde von Heckmann (2008, S. 41) übernommen.

Tabelle 4.1: Überblick über die Fitstatistiken

Fit-Index	Guter Modellfit	Akzeptabler Modellfit
χ^2	0-2 *df*	2*df*-3-*df*
p-Wert χ^2	.05-1.00	.01-.05
χ^2/df	0-2	2-3
RMSEA	0-.05	.05-.08
CI (RMSEA)	Linke Grenze = .00	Nahe an *RMSEA*
SRMR	0-.05	.05-.10
NNFI	0.97-1.00	0.95-.97
CFI	.97-1.00	.95-97
GFI	.95-1.00	.90-.95
AIC	Kleiner als *AIC* für Vergleichsmodell	

Anmerkungen: RMSEA = Root Mean Square Error of Approximation, *SRMR* = Standardizes Root Mean Square Residual, *NNFI* = Nonnormed Fit Index (da er nicht standardisiert ist, kann er auch Werte außerhalb 0-1 annehmen), *CFI* = Comparative Fit Index, *GFI* = Goodness-of-Fit Index, *AIC* = Akaike Information Criterion

5 Ergebnisse

Im fünften Kapitel werden die Resultate der Studie beschrieben. Als Erstes wird dabei die deskriptive Analyse des BDI-V dargestellt (Kapitel 5.1). Daran schließt sich die Präsentation der Latent-State-Trait-Analysen an. Dabei werden zunächst die in Kapitel 4.4.1 beschriebenen Einzelgruppenmodelle (Kapitel 5.2) dargestellt, danach die in Kapitel 4.4.2 erläuterten Multi-Sample-Analysen (Kapitel 5.3).

5.1 Deskriptive Analyse des BDI-V

Die folgenden beiden Abschnitte dieser Arbeit widmen sich den im Rahmen dieser Studie erzielten Deskriptivstatistiken. Dabei werden zunächst Itemkennwerte und interne Konsistenz des BDI-V berichtet (Kapitel 5.1.1). Es folgen Darstellungen der Mittelwerte, Kovarianzen und Korrelationen für die im Rahmen der Latent-State-Trait-Analyse verwendeten Testhälften (Kapitel 5.1.2).

5.1.1 Itemkennwerte und interne Konsistenz

In Tabelle 5.1 sind die Mittelwerte, Standardabweichungen und part-whole-korrigierten Trennschärfen (r_{it}) der Items des BDI-V abgebildet. Wie in den Vorgängerstudien zum BDI-V (Schmitt & Maes, 2000; Schmitt et. al., 2006; Heckmann, 2008) weisen die Items dieses Messinstruments eine stark rechtsschiefe Verteilung auf. Mit Ausnahme des Suizidalitätsitems (Nr. 9) weisen alle Items eine Standardabweichung größer eins auf. Die Größe der Streuung kann als ein Indikator für die Diskriminationsfähigkeit eines Items interpretiert werden (Amelang & Schmidt-Atzert, 2006, S. 119). Dem folgend kann den Items des BDI-V eine gute Diskriminationsfähigkeit bescheinigt werden.

Tabelle 5.1: Mittelwerte, Standardabweichungen und Trennschärfen der Items des BDI-V

	Item	*M*	*s*	r_{it}
1.	Ich bin traurig. (2)	1.65	1.09	.63
2.	Ich sehe mutlos in die Zukunft. (2)	1.07	1.16	.67
3.	Ich fühle mich als Versager(in). (2)	.91	1.13	.69
4.	Es fällt mir schwer, etwas zu genießen. (2)	1.33	1.36	.59
5.	Ich habe Schuldgefühle. (1)	1.19	1.61	.56
6.	Ich fühle mich bestraft. (1)	.62	1.04	.53
7.	Ich bin von mir enttäuscht. (2)	1.21	1.13	.69
8.	Ich werfe mir Fehler und Schwächen vor. (1)	1.66	1.22	.68
9.	Ich denke daran, mir etwas anzutun. (1)	.30	.77	.51
10.	Ich weine. (1)	.99	1.08	.45
11.	Ich fühle mich gereizt und verärgert. (2)	1.82	1.14	.53
12.	Mir fehlt das Interesse an Menschen. (2)	.94	1.11	.52
13.	Ich schiebe Entscheidungen vor mir her. (1)	2.03	1.34	.52
14.	Ich bin besorgt um mein Aussehen. (1)	1.70	1.30	.37
15.	Ich muss mich zu jeder Tätigkeit zwingen. (1)	1.36	1.18	.68
16.	Ich habe Schlafstörungen. (2)	1.28	1.37	.45
17.	Ich bin müde und lustlos. (1)	1.56	1.20	.72
18.	Ich habe keinen Appetit. (2)	.66	1.01	.41
19.	Ich mache mir Sorgen um meine Gesundheit. (1)	1.30	1.18	.40
20.	Sex ist mir gleichgültig. (2)	1.24	1.36	.44

Anmerkungen: $N = 427$, die eingeklammerte Zahl hinter den Items gibt an, zu welcher Testhälfte das Item gehört.

Die Trennschärfen der Items entsprechen in der Gesamtschau in etwa jenen, die in den zitierten Vorgängerstudien erzielt wurden. Während bei Schmitt & Maes (2000) und Schmitt et. al. (2006) insbesondere das Sexualitätsitem (Nr. 20) die mit Abstand schlechteste Trennschärfe aufwies, erzielte dieses Item in dieser Studie einen vergleichsweise ordentlichen Wert für die Trennschärfe, der in der Nähe des von Heckmann (2008, S. 43) ermittelten Wertes liegt. Die schlechteste Trennschärfe wurde hier mit .37 für das Item „Ich bin besorgt um mein Aussehen." (Nr. 14) festgestellt. Die beste Trennschärfe wies mit .72, wie in den zitierten Vorgängerstudien, das Item „Ich bin müde und lustlos." (Nr. 17) auf. Für die interne Konsistenz wurde ein *Cronbach's Alpha* (Cronbach, 1951) von .91 ermittelt, der gleiche Wert wie bei Heckmann (2008, S. 43). Dieser Wert liegt nahe bei jenen Werten, die in den anderen zitierten Vorgängerstudien zum BDI-V gefunden wurden und in Studien zur deutschen Originalversion des BDI II erzielt wurden (Hautzinger et. al., 2006; Schmitt & Maes, 2000; Schmitt et. al., 2003; Schmitt et. al., 2006).

5.1.2 Mittelwerte, Kovarianzen und Korrelationen der Testhälften

In den folgenden Tabellen sind Mittelwerte, Kovarianzen und Korrelationen der beiden Testhälften zu den beiden Erhebungszeitpunkten, getrennt nach Instruktionsgruppen (Tabelle 5.2.: 3-Monats-Gruppe, 5.3.: 14-Tage-Gruppe), abgebildet. Es zeigt sich, dass die Korrelation zwischen den beiden Testhälften *innerhalb* der Messzeitpunkte (kursiv und fett) hoch ausfallen. Hierin spiegelt sich die im vorangegangenen Abschnitt berichtete hohe interne Konsistenz des BDI-V wider. Die Varianzen der Testhälften *innerhalb* der Zeitpunkte sind bei der 3-Monatsgruppe in etwa gleich groß, was darauf hinweist, dass die verwendeten Testhälften als zumindest annähernd parallel angesehen werden können.

Tabelle 5.2: Mittelwerte, Standardabweichungen, Varianzen (Diagonale), Kovarianzen (unterhalb der Diagonale) und Korrelationen (oberhalb der Diagonale) der beiden Testhälften des vereinfachten Beck-Depressions-Inventars (BDI-V) für die 3-Monatsgruppe für beide Erhebungszeitpunkte in der Analysestichprobe.

	BDI-V_{11}	BDI-V_{21}	BDI-V_{12}	BDI-V_{22}
BDI-V_{11}	51.67	***.85***	.75	*.69*
BDI-V_{21}	43.71	51.61	*.62*	.74
BDI-V_{12}	36.80	30.71	47.09	***.83***
BDI-V_{22}	34.14	36.76	39.54	47.40
M	12.91	11.81	12.34	11.45
s	7.19	7.18	6.86	6.88

Anmerkungen: N = 187. Die erste Zahl des Index' der Variablen steht für die Testhälfte, die zweite für den Erhebungszeitpunkt. Alle Korrelationen sind grau schattiert, Korrelationen zwischen den Testhälften sind kursiv, Korrelationen innerhalb eines Messzeitpunktes sind fettgedruckt, die Retestkorrelationen sind unterstrichen.

Tabelle 5.3: Mittelwerte, Standardabweichungen, Varianzen (Diagonale), Kovarianzen (unterhalb der Diagonale) und Korrelationen (oberhalb der Diagonale) der beiden Testhälften des vereinfachten Beck-Depressions-Inventars (BDI-V) für die 14-Tagegruppe für beide Erhebungszeitpunkte.

	BDI-V_{11}	BDI-V_{21}	BDI-V_{12}	BDI-V_{22}
BDI-V_{11}	53.88	***.82***	.68	*.61*
BDI-V_{21}	47.85	62.99	*.56*	.71
BDI-V_{12}	33.74	30.04	45.88	***.80***
BDI-V_{22}	32.05	39.83	38.79	50.74
M	12.52	12.35	11.63	11.55
s	7.34	7.94	6.77	7.12

Anmerkungen: N = 240. Die erste Zahl des Index' der Variablen steht für die Testhälfte, die zweite für den Erhebungszeitpunkt. Alle Korrelationen sind grau schattiert, Korrelationen zwischen den Testhälften sind kursiv, Korrelationen innerhalb eines Messzeitpunktes sind fettgedruckt, die Retestkorrelationen sind unterstrichen.

Diese positive Beurteilung der verwendeten Testhälften gilt nur deutlich eingeschränkt für die 14-Tage-Gruppe, bei welcher auffällt, dass zu beiden Messzeitpunkten die zweite Testhälfte eine deutlich stärkere Streuung erzeugte. Vollständig parallele Testhälften würden allerdings neben gleichen Varianzen innerhalb der Messzeitpunkte auch implizieren, dass z.B. die Korrelation von Testhälfte eins zum Messzeitpunkt eins mit Testhälfte zwei zum Messzeitpunkt zwei gleich groß ist wie z.B. die Retestkorrelation von Testhälfte eins. Dies ist nicht der Fall. Die Retestkorrelationen (unterstrichen) sind größer als die Korrelationen unterschiedlicher Testhälften zu unterschiedlichen Zeitpunkten (kursiv, weder fett noch unterstrichen). Dies gilt für beide Gruppen. Um den Einfluss der Testhälften auf die manifesten Variablen zu modellieren wurden im Rahmen der LST-Theorie Methodenfaktoren integriert (Schmitt & Steyer, 1993). Die oben beschriebenen Befunde spiegeln den Einfluss der Testhälften wider und zeigen, dass es sinnvoll war, für die Analysen im Rahmen dieser Arbeit ein Modell zu wählen, welches Methodenfaktoren berücksichtigt (siehe Kapitel 4.4.1). Die Retestkorrelationen können als Maßstab für die Stabilität des Messinstruments interpretiert werden. Diese fällt insgesamt deutlich niedriger aus als Heckmann (2008, S. 44). Dies ist zunächst durchaus erwartungsgemäß und kann mit dem größeren Zeitraum, der zwischen den beiden Erhebungszeitpunkten liegt begründet werden. Allerdings mag ein weiterer Grund für diesen Zusammenhang darin zu suchen sein, dass die Varianz zu Messzeitpunkt zwei für beide Instruktionsgruppen insgesamt niedriger ausfällt. Dieses Phänomen stellt einen Unterschied zu den Ergebnissen von Heckmann dar. Die Ursachen hierfür sind beispielsweise in der Stichprobenziehung, oder in der in Kapitel 4.3.2 beschriebenen Intervention zu suchen. Des Weiteren kann konstatiert werden, dass die Korrelationen zwischen den Messzeitpunkten niedriger ausfallen als die innerhalb der Messzeitpunkte. Dies gilt in stärkerem Maße für die 14-Tage-Gruppe. Es ist ein erster Hinweis auf die hypothesenkonforme niedrigere Stabilität der 14-Tage-Gruppe im Vergleich mit der 3-Monats-Gruppe. Der Effekt diesbezüglich fällt etwas deutlicher aus als bei Heckmann und zeigt sich in beiden Instruktionsgruppen. Gruppenunterschiede hinsichtlich der Situationseffekte lassen sich an den Daten der beiden Tabellen ebenfalls ablesen. Maßgeblich sind hierfür die Retestkorrelationen (unterstrichen).

Erwartungsgemäß und wie in der Vorgängerarbeit fallen diese für die 3-Monats-Gruppe etwas höher aus als für 14-Tagegruppe.

5.2 Latent-State-Trait-Analyse der Einzelgruppenmodelle

Der folgende Abschnitt widmet sich den Ergebnissen der Latent-State-Trait-Analysen für die Einzelgruppenmodelle. Diese wurden für beide Instruktionsgruppen getrennt berechnet. Zunächst geschah dies für die Gesamtstichprobe, danach noch einmal extra für die Untergruppe der Personen, welche die Kontrollfrage („Auf welchen Zeitraum haben Sie sich bei den Angaben zu Ihrem Lebensgefühl bezogen?“) korrekt beantwortet haben. Die Kontrollfrage war in der Vorgängerarbeit von Heckmann (2008, S. 66) ursprünglich als Ausschlusskriterium konzipiert worden. Personen, welche die zeitliche Instruktion nicht korrekt erinnern konnten und daher mutmaßlich nicht korrekt umgesetzt sollten nicht zu der im Rahmen der Analyse verwendeten Stichprobe gehören. Heckmann entschied sich jedoch aufgrund der hohen Anzahl falscher Antworten bei der Kontrollfrage dafür, die LST-Analysen sowohl für die Gesamtstichprobe als auch für Unterstichprobe der Personen, welche die Kontrollfrage korrekt beantwortet hatten zu rechnen. Dieses Vorgehen wird übernommen, damit die beiden Studien möglichst gut miteinander verglichen werden können. Die Möglichkeit eines Vergleichs wird allerdings durch die Tatsache limitiert, dass im Rahmen dieser Studie die Kontrollfrage nur zum zweiten Zeitpunkt abgefragt wurde. Von den 187 Versuchspersonen der 3-Monatsgruppe, die beide Messzeitpunkte absolvierten beantworteten 104 die Kontrollfrage korrekt, in der 14-Tage-Gruppe waren es 162 von 240.

5.2.1 Einzelgruppenanalysen für die Gesamtstichprobe

Fitstatistiken für die Latent-State-Trait-Analysen in den Instruktionsgruppen. Tabelle 5.4 gibt die errechneten Fitstatistiken für die beiden Einzelgruppenmodelle der Gesamtstichprobe wieder. Für die 3-Monats-Gruppe kann konstatiert werden, dass das angenommene Modell gut auf die erhobenen Daten passt. Alle Indizes liegen innerhalb des auf einen guten Modellfit hinweisenden Korridors (siehe Tabelle 4.1). Für die 14-

Tage-Gruppe kann ein guter Modellfit hinsichtlich des *p*-Werts des χ^2-Tests sowie hinsichtlich *NNFI*, *GFI* und *CFI* festgestellt werden. χ^2, χ^2/df und *RMSEA* liegen aber im akzeptablen Bereich, so dass die Fitstatistiken insgesamt dafür sprechen, das gewählte Modell für beide Instruktionsgruppen beizubehalten. Die Ursache dafür, dass der Allgemeinbefund hinsichtlich der Modellpassung für die 14-Tage-Gruppe etwas zurückhaltender ausfällt als für die 3-Monats-Gruppe dürfte hauptsächlich in der in Kapitel 5.1.2 beschriebenen Differenz der Varianzen der Testhälften *zwischen* den Messzeitpunkten zu suchen sein. Um dies zu prüfen wird im weiteren Verlauf der Analysen für die 14-Tage-Gruppe alternativ ein Modell getestet, bei dem unterschiedliche Werte für die beiden Methodenfaktoren geschätzt werden.

Tabelle 5.4: Fitstatistiken der Einzelgruppenmodelle (Gesamtstichprobe)

	χ^2_6	*p*	χ^2/df	*RMSEA* (CI)	*SRMR*	*NNFI*	*GFI*	*CFI*	*AIC*
3-Monate	3.98	.68	0.66	.00 (.00;.08)	.04	1.00	0.99	1.00	U: 638.50 Z: 11.98 S: 20.00
14-Tage	12.18	.06	2.03	.07 (.00;.12)	.09	0.99	0.98	0.99	U: 716.83 Z: 20.18 S: 20.00

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = standardisiertes Root Mean Square Residual, *NNFI* = Nonnormed Fit Index, *GFI* = Goodness of Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell)

Überprüfung der Teilstrukturen des Modells. Von einem guten Gesamtmodellfit kann nicht zwingend auf die Güte aller Teilstrukturen des Modells geschlossen werden (Schermelleh-Engel & Moosbrugger, 2002). In beiden Instruktionsgruppen wurden keine unmöglichen Schätzungen, wie z.B. negative Varianzen (sog. „Heywood-Cases“) aufgefunden. Die Beträge der *t*-Werte aller geschätzten Parameter lagen in beiden Instruktionsgruppen oberhalb von 1.96, wovon darauf geschlossen werden kann, dass die geschätzten Parameter auf 5%-Niveau signifikant von Null verschieden sind. Die geschätzte Varianz der manifesten Variablen lag in der 3-Monats-Gruppe beim Wert 49.44, in der 14-Tage-Gruppe beim Wert 53.37. Die Parameterschätzungen für die beiden Instruktionsgruppen sind in Tabelle 5.5 aufgeführt.

Tabelle 5.5: Parameterschätzungen für beide Instruktionsgruppen (Gesamtstichprobe)

	Trait	L-S-R	Methode	ε
3-Monate	32.43 (4.21)	9.20 (1.15)	4.35 (0.66)	3.46 (0.36)
14-Tage	31.04 (3.84)	12.28 (1.33)	5.74 (0.75)	4.31 (0.39)

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, Y_{ik} = Varianz der manifesten Variablen, Trait = Traitvarianz, L-S-R = Varianz der Latent-State-Residuen, Methode = Varianz der Methodenfaktoren, ε = Messfehleranteil; eingeklammert: Standardfehler des geschätzten Wertes.

Latent-State-Trait-Koeffizienten. Tabelle 5.6 beinhaltet die Schätzungen für die LST-Koeffizienten in der Gesamtstichprobe. Das BDI-V erzielte in beiden Instruktionsgruppen eine hohe Reliabilität. Die ermittelte Methodenspezifität ist gering und nimmt mit der Hochrechnung von der Ebene der Testhälften auf die des Gesamttests noch ab, so dass dem BDI-V eine hohe Homogenität attestiert werden kann. Diese Befunde entsprechen denen von Schmitt & Maes (2000) und Heckmann (2008, S. 48).

Tabelle 5.6: LST-Koeffizienten für die Einzelgruppenmodelle der Gesamtstichprobe

	TKon	ZSpe	MSpe	Rel
3-Monate (TH)	.66 (.71)	.19 (.21)	.09 (.05)	.94 (.97)
14-Tage (TH)	.58 (.64)	.23 (.25)	.11 (.07)	.92 (.96)

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, eingeklammert: Werte für den Gesamttest

Die Traitkonsistenz (TKon) liegt für beide Instruktionsgruppen recht hoch: für die 3-Monatsgruppe können annähernd zwei Drittel der Varianz der Testhälften durch stabile Merkmale erklärt werden. In der 14-Tage-Gruppe fällt dieser Anteil niedriger aus, erklärt aber immer noch deutlich mehr als die Hälfte der Varianz. Die Zeitspezifität (ZSpe) zeichnet für etwa ein Fünftel der Varianz der Testhälften verantwortlich – korrespondierend mit den Resultaten für die Traitkonsistenz liegt ihr Anteil in der 14-Tage-Gruppe etwas über dem für die 3-Monatsgruppe. Die auf den Gesamttest hochgerechnete Traitkonsistenz liegt für die 3-Monatsgruppe beim Wert .71 und damit um .07 über dem für die 14-Tage-Gruppe. Die Zeitspezifität ist in der 14-Tage-Gruppe mit .25 um .04 höher als in der 3-Monats-Gruppe. Die Methodenspezifität ist in beiden Gruppen relativ gering, die Gruppenunterschiede fallen diesbezüglich gering aus.

5.2.2 Prüfung des Einflusses der Testhälften in der 14-Tage-Gruppe

Fitstatistiken. Zwar zeigte das im vorangegangenen Abschnitt präsentierte Modell für beide Instruktionsgruppen einen mindestens akzeptablen Fit, jedoch blieb die Frage zu klären, warum dieser in so starkem Maße unterschiedlich ausfiel. Vor dem Hintergrund der Tatsache, dass wie in Kapitel 5.1.2 berichtet die Varianzen der Testhälften in der 14-Tage-Gruppe innerhalb eines Messzeitpunktes höchst unterschiedlich ausfielen wurde daher für die 14-Tage-Gruppe ein Modell gerechnet, das die Gleichheitsannahme für die Methodenfaktoren aufhob und diese frei schätzen ließ. Die Fitstatistiken für dieses Modell sind Tabelle 5.7 dargestellt. Deskriptiv zeigt sich eine Verbesserung aller Gütekennwerte mit Ausnahme des *GFI*, der auf dem

gleichen Wert verbleibt. Eine große Verbesserung zeigt sich vor allem hinsichtlich der χ^2-Statistiken. Der χ^2-Differenzentest (χ^2-Differenz = 4.84; *df* = 1) ist auf 5%-Niveau signifikant. Es zeigt sich also bereits bei den Einzelgruppen deutlich, dass es nicht gelungen ist, für beide Tests annähernd parallele Testhälften zu finden. Aufgrund des guten Modellfits des ursprünglich angenommenen Einzelgruppenmodells in der 3-Monats-Gruppe wurde auf eine Analyse des in diesem Abschnitt präsentierten Modells für diese Instruktionsgruppe verzichtet.

Tabelle 5.7: Fitstatistiken Einzelgruppenmodell 14-Tage-Gruppe mit unterschiedlichen Methodenfaktoren

	χ^2_5	*p*	χ^2/df	*RMSEA* (CI)	*SRMR*	*NNFI*	*GFI*	*CFI*	*AIC*
14-Tage	7.54	.18	1.51	.05 (.00;.11)	.07	1.00	0.98	1.00	U: 716.83 Z: 17.54 S: 20.00

Anmerkungen: N = 240, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = standardisiertes Root Mean Square Residual, *NNFI* = Nonnormed Fit Index, *GFI* = Goodness of Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell)

Teilstrukturen des Modells. Die Prüfungen der Teilstrukturen des Modells fielen zugunsten des Modells aus. Die geschätzte Varianz der manifesten Variablen betrug für die erste Testhälfte 50.11 und für die zweite Testhälfte 56.64. Die geschätzten Parameter für die beiden Testhälften sind in Tabelle 5.8 dargestellt.

Tabelle 5.8: Parameterschätzungen für die 14-Tage-Gruppe (unterschiedl. Methodenfaktoren)

	Trait	L-S-R	Methode1	Methode2	ε
14-Tage	31.05 (3.83)	12.27 (1.33)	2.47 (1.70)	9.01 (1.94)	4.31 (0.39)

Anmerkungen: N = 240, Y_{ik} = Varianz der manifesten Variablen, Trait = Traitvarianz, L-S-R = Varianz der Latent-State-Residuen, Methode = Varianz der Methodenfaktoren, ε = Messfehleranteil; eingeklammert: Standardfehler des geschätzten Wertes.

Latent-State-Trait-Koeffizienten. In Tabelle 5.9 sind die LST-Koeffizienten für die 14-Tage-Gruppe, getrennt nach Testhälften dargestellt. Die Methodenspezifität der zweiten Testhälfte fällt dabei deutlich höher aus. Dies geht vor allem zulasten der Traitkonsistenz.

Tabelle 5.9: LST-Koeffizienten für die 14-Tage-Gruppe (unterschiedliche Methodenfaktoren)

	TKon	ZSpe	MSpe	Rel.
14-Tage (TH1)	.62	.24	.05	.91
14-Tage (TH2)	.55	.22	.16	.93

Anmerkung: N = 240.

5.2.3 Einzelgruppenanalysen für die Kontrollfragenstichprobe

Mittelwerte, Kovarianzen und Korrelationen der Testhälften. In Tabelle 5.10 sind die Mittelwerte, Standardabweichungen, Varianzen und Korrelationen der Testhälften für die 3-Monatsgruppe in der Kontrollfragenstichprobe abgebildet. Die gleichen Statistiken enthält Tabelle 5.11 für die 14-Tage-Gruppe. Die Korrelationen fallen recht ähnlich aus wie in der Gesamtstichprobe. Auffällig ist, dass in der 3-Monats-Gruppe die Varianz der zweiten Testhälfte bei Erhebungszeitpunkt eins deutlich geringer ausfällt als die der ersten Testhälfte zum gleichen Zeitpunkt. Die für diese Gruppe festgestellte annähernde Parallelität der Testhälften lässt sich also für die Kontrollfragenstichprobe nicht im gleichen Umfang feststellen. Der für die

Gesamtstichprobe bereits beschriebene Abfall der Varianz von Testzeitpunkt eins zu zwei zeigt sich auch für die Kontrollfragenstichprobe. Insgesamt gilt für beide Instruktionsgruppen: Varianzen und Mittelwerte liegen unterhalb den für die Gesamtstichprobe ermittelten Werte. Ein ähnliches Muster hatte sich in Heckmanns (2008, S. 49) Studie ergeben.

Tabelle 5.10: Mittelwerte, Standardabweichungen, Varianzen (Diagonale), Kovarianzen (unterhalb der Diagonale) und Korrelationen (oberhalb der Diagonale) der beiden Testhälften des BDI-V für die 3-Monatsgruppe für beide Erhebungszeitpunkte in der Kontrollfragegruppe

	$BDI\text{-}V_{11}$	$BDI\text{-}V_{21}$	$BDI\text{-}V_{12}$	$BDI\text{-}V_{22}$
$BDI\text{-}V_{11}$	50.96	.85	.76	.65
$BDI\text{-}V_{21}$	40.01	43.91	.64	.71
$BDI\text{-}V_{12}$	36.46	28.42	45.06	.83
$BDI\text{-}V_{22}$	30.21	30.93	36.52	42.86
M	12.38	11.32	11.95	10.87
S	7.14	6.63	6.71	6.55

Anmerkungen: N = 104. Die erste Zahl des Index' der Variablen steht für die Testhälfte, die zweite für den Erhebungszeitpunkt.

Fitstatistiken für die LST-Modelle. Tabelle 5.12 enthält die Fitstatistiken für die nach Instruktionsgruppen getrennt berechneten Einzelmodelle der Kontrollfragenstichprobe. Diese entsprechen im Wesentlichen den für die Gesamtstichprobe ermittelten. Für die 3-Monats-Gruppe ergeben sich marginal bessere Gütekennwerte bei *SRMR, NNFI* und *AIC,* ebenso ein geringfügig schlechterer *GFI*-Wert. Die χ^2-Statistiken (inklusive *p*-Wert und den um die Freiheitsgrade relativierten χ^2-Wert) sind etwas besser, was angesichts der geringeren Stichprobengröße nicht verwunderlich ist. Insgesamt passt das Modell gut auf die Daten. Für die 14-Tage-Gruppe ergeben sich insgesamt etwas problematischere Werte: trotz der ebenfalls kleineren Stichprobe fallen die χ^2-Statistiken schlechter aus als in der Gesamtstichprobe. Dies gilt mit Ausnahme des *CFI* auch für alle anderen

Gütekennwerte. Der Unterschied zur Gesamtstichprobe fällt allerdings insgesamt nicht gravierend aus, so dass das Modell insgesamt beibehalten werden kann.

Tabelle 5.11: Mittelwerte, Standardabweichungen, Varianzen (Diagonale), Kovarianzen (unterhalb der Diagonale) und Korrelationen (oberhalb der Diagonale) der beiden Testhälften des BDI-V für die 14-Tagegruppe für beide Erhebungszeitpunkte in der Kontrollfragegruppe

	$BDI\text{-}V_{11}$	$BDI\text{-}V_{21}$	$BDI\text{-}V_{12}$	$BDI\text{-}V_{22}$
$BDI\text{-}V_{11}$	51.19	.81	.67	.61
$BDI\text{-}V_{21}$	45.02	59.23	.51	.68
$BDI\text{-}V_{12}$	32.02	25.66	43.40	.77
$BDI\text{-}V_{22}$	28.42	33.95	32.92	41.67
M	12.37	12.19	11.27	11.09
S	7.22	7.70	6.59	6.46

Anmerkungen: N = 162. Die erste Zahl des Index‘ der Variablen steht für die Testhälfte, die zweite für den Erhebungszeitpunkt.

Überprüfung der Teilstrukturen des Modells. Auch für die Kontrollfragenstichprobe wurden zusätzlich die Teilstrukturen des Modells überprüft. Auch hier zeigten alle *t*-Werte der Parameterschätzungen einen höheren Wert als |1.96| und es traten ebenfalls keine unmöglichen Schätzungen auf. Für die 3-Monats-Gruppe beträgt die geschätzte Varianz der manifesten Variablen 45.70, für die 14-Tage-Gruppe 48.87. Die Parameterschätzungen für beide Instruktionsgruppen können Tabelle 5.13 entnommen werden.

Tabelle 5.12: Fitstatistiken der Einzelgruppenmodelle in der Kontrollfragenstichprobe

	χ^2_6	*p*	χ^2/df	*RMSEA* (CI)	*SRMR*	*NNFI*	*GFI*	*CFI*	*AIC*
3-Monate	3.33	.77	0.57	.00 (.00;.09)	.06	1.01	0.98	1.00	U: 349.89 Z: 11.33 S: 20.00
14-Tage	12.86	.05	2.14	.08 (.01;.15)	.11	0.99	0.96	0.99	U: 463.79 Z: 20.86 S: 20.00

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = standardisiertes Root Mean Square Residual, *NNFI* = Nonnormed Fit Index, *GFI* = Goodness of Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell)

Tabelle 5.13: Parameterschätzungen für beide Instruktionsgruppen (Kontrollfragenstichprobe)

	Trait	L-S-R	Methode	ε
3-Monate	29.31 (5.19)	8.95 (1.48)	4.38 (0.85)	3.05 (0.43)
14-Tage	27.04 (4.22)	11.93 (1.57)	5.95 (0.91)	3.96 (0.44)

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, Y_{ik} = Varianz der manifesten Variablen, Trait = Traitvarianz, L-S-R = Varianz der Latent-State-Residuen, Methode = Varianz der Methodenfaktoren, ε = Messfehleranteil, eingeklammert: Standardfehler des geschätzten Wertes.

LST-Koeffizienten. Tabelle 5.14 enthält die für die Einzelmodelle der Kontrollfragenstichprobe berechneten LST-Koeffizienten. Diese liegen äußerst nahe an den für die Gesamtstichprobe ermittelten und weisen die für diese beschriebenen Muster in gleichem Maße auf. Die getrennte Analyse der Kontrollfragenstichprobe ergibt also für die Einzelgruppenmodelle, ähnlich wie in der Arbeit von Heckmann (2008, S. 52) keine bedeutsamen Unterschiede hinsichtlich der LST-Koeffizienten.

Tabelle 5.14: LST-Koeffizienten für die Einzelmodelle der Kontrollfragenstichprobe

	TKon	ZSpe	MSpe	Rel
3-Monate (TH)	.64 (.70)	.20 (.22)	.10 (.05)	.94 (.97)
14-Tage (TH)	.55 (.61)	.24 (.27)	.12 (.07)	.91 (.95)

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, eingeklammert: Werte für den Gesamttest

5.2.4 Alternativmodell für die 14-Tage-Gruppe

Zwar zeigte sich in der Kontrollfragenstichprobe auch in der 3-Monats-Gruppe eine Auffälligkeit hinsichtlich der Varianz der ersten Testhälfte zum ersten Messzeitpunkt. Aufgrund des insgesamt sehr guten Fits des angenommenen Modells in der 3-Monats-Gruppe wurde aber auf die Errechnung eines Alternativmodells verzichtet. Da die Fitstatistiken für die 14-Tage-Gruppe auch in der Kontrollfragenstichprobe deskriptiv erkennbar schlechter ausfallen wird auch für die Kontrollfragenstichprobe das Alternativmodell mit unterschiedlichen Varianzen der Methodenfaktoren geprüft. Die Fitstatistiken für dieses Modell sind in Tabelle 5.15 dargestellt.

Tabelle 5.15: Fitstatistiken Einzelgruppenmodell 14-Tage-Gruppe mit unterschiedlichen Methodenfaktoren

	χ^2_5	*p*	χ^2/df	*RMSEA* (CI)	*SRMR*	*NNFI*	*GFI*	*CFI*	*AIC*
14-Tage	11.82	.04	2.36	.09 (.02;.16)	.10	0.98	0.96	0.98	U: 463.69 Z: 21.82 S: 20.00

Anmerkungen: N = 162, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = standardisiertes Root Mean Square Residual, *NNFI* = Nonnormed Fit Index, *GFI* = Goodness of Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell)

Dabei kann konstatiert werden, dass anders als in der Gesamtstichprobe dieses Alternativmodell keine Verbesserung des Modellfits erbringt. Der χ^2-Wert sinkt zwar marginal, der *p*-Wert hingegen erhöht sich nicht, der χ^2/df-Wert steigt, der *RMSEA* erhöht sich, *NNFI* und *CFI* sinken und der *AIC*-Wert für das Zielmodell erhöht sich. All diese Tendenzen sprechen gegen das Modell. Es wird daher verworfen. Der für die Gesamtstichprobe in der 14-Tage-Gruppe festgestellte Einfluss der Testhälften auf die Modellgüte kann für die Kontrollfragenstichprobe nicht nachgewiesen werden. Auf die Ermittlung der LST-Koeffizienten wird daher verzichtet.

5.3 Mehrgruppenvergleiche

Dieser Abschnitt widmet sich den Ergebnissen der Analyse der Mehrgruppenmodelle (Multi-sample-Analysen, siehe Kapitel 4.4.2), bei denen beide Instruktionsgruppen gleichzeitig getestet wurden. Auch diese werden für die Gesamtstichprobe und die Kontrollfragenstichprobe getrennt durchgeführt.

5.3.1 Mehrgruppenvergleiche in der Gesamtstichprobe

Modelltests und Fitstatistiken. Tabelle 5.16 enthält die Fitstatistiken für die in Kapitel 4.4.2 beschriebenen Modelle. Zu beachten ist bei den Fitstatistiken der Multi-sample-Analysen, dass *SRMR* und *GFI* in LISREL für die beiden Instruktionsgruppen getrennt ausgegeben werden. Diese lassen also eine nach Instruktionsgruppen getrennte Beurteilung zu, wie gut das gewählte Modell auf die Daten passt. Alle weiteren angegebenen Gütekennwerte beschreiben den Gesamtfit des jeweiligen Modells für beide Instruktionsgruppen. In Anlehnung an Heckmann (2008, S. 53) wird auf eine Untersuchung der Teilstrukturen der im folgenden berichteten Mehrgruppenmodelle verzichtet, da der Autor ebenfalls davon ausgeht, dass die Überprüfung der Einzelmodelle als ausreichend angesehen werden kann und unmögliche Schätzungen sowie nicht signifikant von Null verschiedene Parameterschätzungen nicht zu erwarten sind. Hinzu kommt, dass in diesem Abschnitt die vergleichende Betrachtung

zwischen den Modellen im Vordergrund steht und von einem Einfluss auf die Teilstrukturen nicht auszugehen ist.

Das Invarianz-Modell (In-M) weist insgesamt gute bis akzeptable Modellkennwerte auf. χ^2-Wert, χ^2/df, *CFI* und *NNFI* können als gut angesehen werden, der *p*-Wert des χ^2-Tests und der *RMSEA* liegen im akzeptablen Bereich. Für die nach Instruktionsgruppen getrennt berechneten Indizes gilt: der *GFI* ist für beide Instruktionsgruppen gut, der *SRMR* ist für die 3-Monatsgruppe ebenfalls als gut anzusehen, für die 14-Tage-Gruppe als akzeptabel. Wie beschrieben nimmt dieses Modell *keine* Unterschiede hinsichtlich der zu schätzenden Parameter zwischen den beiden Instruktionsgruppen an. Die gute bis akzeptable Passung dieses Modells spricht nicht für eine Auswirkung der zeitlichen Instruktionen auf das Antwortverhalten der Teilnehmerinnen und Teilnehmer im BDI-V.

Allerdings ist es umgekehrt auch nicht so, dass die Modellgütekennwerte des Invarianzmodells derart gut sind, dass eine Verbesserung ausgeschlossen erscheint. Deswegen werden die in Kapitel 4.4.2 beschriebenen Schritte 2 und 3 vollzogen und der Modellvergleich im Folgenden beschrieben. Schritt 2 beinhaltet die Analyse eines Modells, das Unterschiede zwischen den Instruktionsgruppen hinsichtlich der Traitvarianz und der Latent-State-Residuen zulässt. Die in Tabelle 5.16 zu entnehmenden Fit-Statistiken weisen auf deskriptiver Ebene tatsächlich auf eine Verbesserung des Modellfits hin. Mit Ausnahme des lediglich akzeptablen *SRMR* für die 14-Tage-Gruppe dürfen in diesem Modell alle ermittelten Modellkennwerte als gut zu bezeichnet werden. Allerdings fallen die Unterschiede zum Invarianzmodell nur geringfügig aus. Wie in Kapitel 4.4.2 beschrieben erlauben die Multi-sample-Analysen neben einem deskriptiven Vergleich der Modellgütekennwerte auch eine Absicherung vorgefundener Unterschiede durch einen Signifikanztest: den χ^2-Differenzentest. Der hierbei erzielte Wert der χ^2-Differenz (4.47, $df = 2$) ist auf 5%-Niveau nicht signifikant.

Tabelle 5.16: Fitstatistiken Mehrgruppenmodelle (Gesamtstichprobe)

		SRMR	*GFI*	χ^2	*p*-Wert	χ^2/ *df*	*RMSEA* (CI)	*NNFI*	*CFI*	*AIC*
In-M	3-M	.05	.97				.06			U: 1355.34
df = 16	14-T	.09	.97	26.99	.04	1.69	(.01; .09)	0.99	0.99	Z: 34.99 S: 40.00
Traitv/ L-S-R	3-M	.04	.98				.05			U: 1355.34
df=14	14-T	.09	.97	22.52	.07	1.61	(.00;.09)	1.00	0.99	Z: 34.52 S: 40.00
Un-M	3-M	.04	.99				.04			U: 1355.34
df=12	14-T	.09	.98	16.17	.18	1.35	(.00;.09)	1.00	1.00	Z: 32.17 S: 40.00

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, In-M = Invarianzmodell, Traitv/L-S-R = Modell mit Lockerungen der Gleichheitsrestriktionen für Traitvarianz und L-S-Residuen, Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt), *SRMR* = standardisiertes Root Mean Square Residual, *GFI* = Goodness of Fit Index, *RMSEA* = Root Mean Square Error of Approximation (CI = Konfidenzintervall), *NNFI* = Nonnormed Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell).

In Schritt 3 werden neben den in Schritt 2 gelockerten Parameter auch die Restriktionen für die Methodenvarianz und die Varianz der Messfehler aufgehoben. Mittels des daraus resultierenden Unabhängigkeitsmodells (Un-M) sollte ursprünglich überprüft werden, ob sich die in Schritt 2 angenommene signifikante Verbesserung des Modellfits durch Lockerung der für die zeitlichen Instruktion relevanten Parameter

von einer Veränderung durch hypothesenirrelevante Modellmodifikationen abgrenzen lässt (Heckmann, 2008, S. 54). Auf deskriptiver Ebene zeigen die meisten Kennwerte nur eine marginale Verbesserung gegenüber dem Modell aus Schritt 2. Die χ^2-Statistiken jedoch haben sich deutlich verbessert. Diese Veränderung wird durch den χ^2-Differenzentest (χ^2-Differenz = 6.08, *df* = 2), der auf 5%-Niveau signifikant ausfällt, bestätigt. Das Unabhängigkeitsmodell passt also von den drei Modellen am besten auf die Daten. Dieser Befund stützt den in den Hypothesen angenommenen Einfluss der zeitlichen Instruktionen auf das Antwortverhalten der Probandinnen und Probanden hinsichtlich des BDI-V *nicht*. Vielmehr scheint es anderweitige Merkmale zu geben, hinsichtlich derer sich die beiden Instruktionsgruppen unterscheiden, die das Antwortverhalten beim Ausfüllen des BDI-V beeinflussen. Dies könnten z.B. soziodemographische Eigenschaften sein, die entweder direkt auf das Antwortverhalten Einfluss nehmen oder die in Interaktion mit den im Rahmen dieser Arbeit verwendeten Testhälften dieses beeinflussen.

LST-Koeffizienten. Tabelle 5.17 enthält die LST-Koeffizienten für die Mehrgruppenmodelle der Gesamtstichprobe. Das Invarianzmodell impliziert, dass hinsichtlich der zu schätzenden Parameter keine Unterschiede zwischen den Gruppen bestehen. Dementsprechend werden für beide Gruppen die gleichen LST-Koeffizienten berechnet. Die dabei erzielten Werte liegen zwischen den Werten, die für die Einzelgruppenmodelle jeweils getrennt für die beiden Instruktionsgruppen berechnet wurden. Dieser Befund hat seine Ursache darin, dass LISREL zunächst die Parameterrestriktionen ignoriert und in einem ersten Schritt die Parameter getrennt für die beiden Gruppen berechnet, um danach im zweiten Schritt den Mittelwert aus beiden zu bilden (Jöreskog & Sörbom, 1996, S. 279). Die Lockerung der zeitlichen Instruktionen in Schritt 2 führt im Vergleich zum Invarianzmodell dazu, dass sich die Traitkonsistenz für die 3-Monatsgruppe erhöht und für die 14-Tage-Gruppe absenkt. Korrespondierend dazu liegt die Zeitspezifität für die 3-Monats-Gruppe unterhalb des auf Basis des Invarianzmodells berechneten Wertes, der für die 14-Tage-Gruppe darüber. Dieser Befund deckt sich mit dem für die Einzelgruppenmodelle festgestellten und es ist ein Ergebnis, das die aufgestellten Hypothesen erwarten ließen.

Tabelle 5.17: LST-Koeffizienten Mehrgruppenmodelle (Gesamtstichprobe)

		TKon	ZSpe	MSpe	Rel
In-M	3-M 14-T	.61 (.67)	.21 (.23)	.10 (.06)	.92 (.96)
Traitv./ L-S-R	3-M	.64 (.70)	.18 (.20)	.10 (.06)	.92 (.96)
	14-T	.59 (.64)	.24 (.26)	.10 (.06)	.93 (.96)
Un-M	3-M	.66 (.71)	.19 (.21)	.09 (.05)	.94 (.97)
	14-T	.58 (.64)	.23 (.25)	.11 (.07)	.92 (.96)

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, In-M = Invarianzmodell, Traitv/L-S-R = Modell mit Lockerungen der Gleichheitsrestriktionen für Traitvarianz und L-S-Residuen, Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt). Eingeklammert: Werte für den Gesamttest

Die im vorangegangenen Abschnitt beschriebene Tatsache, dass sich diese beiden Modelle nicht signifikant voneinander unterscheiden spricht allerdings dagegen, dies im Sinne der Hypothesen als eine Auswirkung der Manipulation der zeitlichen Instruktionen zu bezeichnen. Insgesamt bestätigen sich die für Einzelgruppenmodelle bereits beschriebenen Zusammenhänge: auch für die Mehrgruppenmodelle liegt die Traitkonsistenz für beide Gruppen auf einem deutlich niedrigeren Niveau als in der Vorgängerstudie (Heckmann, 2008, S. 57), die Zeitspezifität auf einem höheren. Der Einfluss der Methodenfaktoren fällt nur relativ gering ins Gewicht, er liegt allerdings ein wenig höher als in Heckmanns Studie.

Die zusätzliche Lockerung der Varianzen der Methodenfaktoren und der Messfehlervarianzen führt in etwa zum gleichen Befundmuster wie die vorangegangene Lockerung der Varianzen der Latent-State-Residuen und der Traitvarianz. Insgesamt verdichtet sich also das in Kapitel 5.2.1 beschriebene Bild eines

Unterschieds zwischen den beiden Instruktionsgruppen hinsichtlich des Traiteinflusses und der Zeitspezifität. Wie erläutert kann dieser Befund aber nicht oder nur bedingt im Sinne der Hypothesen interpretiert werden. Vielmehr gilt es analog zu den Analysen der Einzelgruppenmodelle zu prüfen, ob die Annahme gleicher Varianzen der Methodenfaktoren auf die erhobenen Daten passt und sich die ermittelten Unterschiede nicht besser durch den starken Effekt der Testhälften in der 14-Tage-Gruppe erklären lassen.

Einfluss der Testhälften. Zu diesem Zweck wurde ein Modell errechnet, dass nicht nur unterschiedliche Schätzungen der Methodenfaktoren *zwischen* den Gruppen zulässt, wie es das Unabhängigkeitsmodell bereits vorsah, sondern *zusätzlich* die Schätzung unterschiedlicher Werte für die Methodenfaktoren *innerhalb* der Instruktionsgruppen zuließ. Dieses Alternativmodell wird im Folgenden als Un-M-Meth bezeichnet. Die Fitstatistiken für dieses Modell sind in Tabelle 5.18 abgebildet. Zum Vergleich wurden zusätzlich die Modellgütekennwerte für das Unabhängigkeitsmodell (Un-M) in seiner ursprünglichen Fassung in die Tabelle aufgenommen. Bereits auf deskriptiver Ebene sind große Verbesserungen des Modellfits festzustellen. *SRMR* und *GFI* verbessern sich für die 14-Tage-Gruppe, der *GFI* auch leicht für die 3-Monats-Gruppe, der *RMSEA*-Wert verbessert sich ebenfalls, das Konfidenzintervall des *RMSEA* wird etwas kleiner, *CFI* und *AIC* verbessern sich ebenfalls. Die Verbesserungen hinsichtlich der χ^2-Statistiken sind beträchtlich. Der χ^2-Differenzentest (χ^2-Differenz = 11.00, df = 2) weist diese als auf 5% Niveau signifikant aus. Dies verdeutlicht, dass die ermittelten Gruppenunterschiede stark im Zusammenhang mit den verwendeten Testhälften stehen.

Tabelle 5.18: Vergleich Unabhängigkeitsmodelle

		SRMR	*GFI*	χ^2	*p*-Wert	χ^2/ *df*	*RMSEA* (CI)	*NNFI*	*CFI*	*AIC*
Un-M	3-M	.04	.98				.05			U: 1355.34
df=12	14-T	.09	.97	22.52	.07	1.61	(.00;.09)	1.00	0.99	Z: 34.52 S: 40.00
Un-M-Meth	3-M	.04	.99				.03			U: 1355.34
df=10	14-T	.07	.98	11.52	.32	1.15	(.00;.08)	1.00	1.00	Z: 31.52 S: 40.00

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt, nur ein Wert pro Gruppe für den Methodenfaktor), Un-M-Meth (wie Un-M, jedoch pro Gruppe zwei Werte für die Methodenfaktoren), *SRMR* = standardisiertes Root Mean Square Residual, *GFI* = Goodness of Fit Index, *RMSEA* = Root Mean Square Error of Approximation (CI = Konfidenzintervall), *NNFI* = Nonnormed Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell).

LST-Koeffizienten. Das Un-M-Meth-Modell macht es notwendig, dass die Latent-State-Trait-Koeffizienten nach Testhälften getrennt berechnet werden. In Tabelle 5.19 sind die dabei errechneten Werte aufgeführt. Während für die 3-Monats-Gruppe annähernd gleiche Parameter geschätzt werden und daher für die gerundeten LST-Koeffizienten keine Unterschiede zwischen Testhälften erkennbar sind, ergibt sich für die 14-Tage-Gruppe ein beträchtlicher Unterschied hinsichtlich der Methodenspezifität, deren Einfluss auf das Antwortverhalten in Testhälfte 2 deutlich gravierender ausfällt. Dies geht vor allem zulasten der Traitkonsistenz.

Tabelle 5.19: LST-Koeffizienten Unabhängigkeitsmodell mit zwei unterschiedlichen Methodenfaktoren (Gesamtstichprobe)

		TKon	ZSpe	MSpe	Rel
Un-M-Meth (TH 1)	3-M	.66	.19	.09	.94
	14-T	.62	.25	.05	.93
Un-M-Meth (TH 2)	3-M	.66	.19	.09	.94
	14-T	.55	.22	.16	.93

Anmerkungen: N (3-Monate) = 187, *N* (14-Tage) = 240, Un-M-Meth (Alle Parameter frei geschätzt, für alle Parameter außer den Methodenfaktoren gleiche Werte für die jeweilige Testhälfte)

5.3.2 Mehrgruppenvergleiche in der Kontrollfragenstichprobe

Im Folgenden werden die Ergebnisse der Multi-sample-Analysen für die Kontrollfragenstichprobe dargestellt.

LST-Koeffizienten. Tabelle 5.20 gibt die Fitstatistiken für die analysierten Multi-sample-Analysen der Kontrollfragenstichprobe wieder. Das Invarianzmodell weist in dieser Unterstichprobe auf eine gute Passung hindeutende Werte für die χ^2-Statistiken sowie *NNFI* und *CFI* auf. Der *GFI*-Wert für die Instruktionsgruppen ist jeweils gut, der *SRMR* für die 3-Monats-Gruppe akzeptabel, für die 14-Tage-Gruppe liegt er knapp außerhalb des akzeptabel zu bezeichnenden Korridors (siehe Tabelle 4.1), was bedeutet, dass die empirisch ermittelte Kovarianzmatrix recht stark von der modellimplizierten abweicht. Der *RMSEA*-Wert ist akzeptabel. Eine Gesamtbeurteilung des Modells erscheint daher nicht einfach. Da allerdings nur ein Wert knapp jenseits dessen liegt was als akzeptabel angesehen werden darf und manche sogar als gut bezeichnet werde müssen, kann das Modell wohl insgesamt als akzeptabel auf die erhobenen Daten passend angesehen werden. Die bereits beschriebenen Probleme bezüglich der niedrigeren Varianz zu Zeitpunkt 2, insbesondere in der 14-Tage-Gruppe, schlagen deutlich zu Buche.

Tabelle 5.20: Fitstatistiken Mehrgruppenmodelle Kontrollfragenstichprobe

		SRMR	*GFI*	χ^2	*p*-Wert	χ^2/ *df*	*RMSEA* (CI)	*NNFI*	*CFI*	*AIC*
In-M	3-M	.06	.96				.06			U: 813.68
df = 16	14-T	.11	.95	24.53	.08	1.53	(.00; .11)	0.99	0.99	Z: 32.53 S: 40.00
Traitv/ L-S-R	3-M	.05	.97				.06			U: 813.68
df=14	14-T	.11	.96	20.97	.10	1.50	(.00;.11)	0.99	0.99	Z: 32.97 S: 40.00
Un-M	3-M	.06	.98				.05			U: 813.68
df=12	14-T	.11	.96	16.18	.18	1.35	(.00;.11)	1.00	1.00	Z: 32.18 S: 40.00

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, In-M = Invarianzmodell, Traitv/L-S-R = Modell mit Lockerungen der Gleichheitsrestriktionen für Traitvarianz und L-S-Residuen, Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt), *SRMR* = standardisiertes Root Mean Square Residual, *GFI* = Goodness of Fit Index, *RMSEA* = Root Mean Square Error of Approximation (CI = Konfidenzintervall), *NNFI* = Nonnormed Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell).

Die Lockerungen der Parameterrestriktionen in Schritt 2 führen zu keinen großen Veränderungen. Der *SRMR* verbessert sich für die 3-Monats-Gruppe marginal, der sparsame Modelle bevorzugende *AIC* verschlechtert sich ein wenig, die die χ^2-Statistiken zeigen deskriptiv eine Verbesserung, die sich allerdings, wie schon bei der Gesamtstichprobe, als auf 5%-Niveau nicht signifikant erweist (χ^2-Differenz = 4.56, *df* = 2). Die Modelltests für die Kontrollfragenstichprobe zeichnen also ein ähnliches Bild wie für die Gesamtstichprobe: die Lockerung der Restriktionen für die Traitvarianz und die Latent-State-Residuen erbringt keine signifikante Verbesserung des Modellfits. Hingegen erbringt die Aufhebung der Gleichheitsrestriktionen für die Messfehlervarianz und die Varianz der Methodenfaktoren eine signifikante Verbesserung des χ^2-Wertes.

LST-Koeffizienten. Tabelle 5.21 gibt die LST-Koeffizienten für die Mehrgruppenmodelle der Kontrollfragenstichprobe wieder. Die Reliabilitäten sind nahezu identisch. Die Traitkonsistenz fällt insgesamt niedriger aus als in der Gesamtstichprobe, während Zeit- und Methodenspezifität höher ausfallen. Die Unterschiede zu den LST-Koeffizienten der Gesamtstichprobe sind allerdings marginal. Die Unterschiede zwischen den Instruktionsgruppen fallen für die Modelle aus Schritt 2 und 3 in etwa gleich aus wie für die Gesamtstichprobe.

Tabelle 5.21: LST-Koeffizienten Mehrgruppenmodelle (Kontrollfragenstichprobe)

		TKon	ZSpe	MSpe	Rel
In-M	3-M 14-T	.59 (.65)	.23 (.25)	.11 (.06)	.93 (.96)
Traitv./ L-S-R	3-M	.62 (.69)	.19 (.21)	.11 (.06)	.92 (.96)
	14-T	.57 (.63)	.25 (.27)	.11 (.06)	.93 (.96)
Un-M	3-M	.64 (.70)	.20 (.22)	.10 (.05)	.94 (.97)
	14-T	.55 (.61)	.24 (.27)	.12 (.07)	.91 (.95)

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, In-M = Invarianzmodell, Traitv/L-S-R = Modell mit Lockerungen der Gleichheitsrestriktionen für Traitvarianz und L-S-Residuen, Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt). Eingeklammert: Werte für den Gesamttest

Einfluss der Testhälften. Obwohl es für die Einzelgruppenmodelle der Kontrollfragenstichprobe keine bedeutsamen Modellverbesserungen erbracht hatte, die Bedingung der Gleichheit der Varianzen für die 14-Tage-Gruppen aufzulösen, wird hier zu Kontroll- und Vergleichszwecken dennoch das Unabhängigkeitsmodell mit verschiedenen Varianzen der Methodenfaktoren innerhalb der Gruppen (Un-M-Meth) präsentiert. Die Modellgütekennwerte sind Tabelle 5.22 zu entnehmen. Dabei wurde wieder der Vergleich zwischen dem Unabhängigkeitsmodell und dem Un-M-Meth-Modell dargestellt. Es zeigt sich, dass sich die Fitstatistiken dabei nur marginal verändern. Für die Einzelgruppenebene ergibt sich dabei nur eine minimale Verbesserung des *SRMR*, bei gleichbleibendem *GFI*. Überraschend hoch fällt die Verbesserung des *SRMR* für die 3-Monats-Gruppe aus. Auch der *GFI* verbessert sich für diese leicht. Die Gütekennwerte jenseits der χ^2-Statistiken verändern sich kaum.

Die wenigen feststellbaren Veränderungen (rechte Grenze des *CI* des *RMSEA*, *AIC*-Wert für das Zielmodell) sprechen zuungunsten des Modells. Das *AIC* bestraft dabei die geringere Sparsamkeit des Modells. Der χ^2-Differenzentest weist die minimalen Verbesserungen der χ^2-Statistiken als auf 5% Niveau nicht signifikant aus (χ^2-Differenz = 2,57; *df* = 2). Aus diesem Grund wird Un-M-Meth-Modell für die Kontrollfragenstichprobe verworfen und keine Berechnung der LST-Koeffizienten auf Basis dieses Modells durchgeführt.

Tabelle 5.22: Vergleich Unabhängigkeitsmodelle (mit und ohne Gruppenunterschiede bzgl. der Methodenfaktoren, Kontrollfragenstichprobe)

		SRMR	*GFI*	χ^2	*p*-Wert	χ^2/ *df*	*RMSEA* (CI)	*NNFI*	*CFI*	*AIC*
Traitv/ L-S-R	3-M	.06	.98				.05			U: 813.68
df=12	14-T	.11	.96	16.18	.18	1.35	(.00;.11)	1.00	1.00	Z: 32.18 S: 40.00
Un-M-Meth	3-M	.03	.99				.05			U: 813.68
df=10	14-T	.10	.96	13.61	.19	1.36	(.00;.12)	1.00	1.00	Z: 33.61 S: 40.00

Anmerkungen: N (3-Monate) = 104, *N* (14-Tage) = 162, In-M = Un-M = Unabhängigkeitsmodell (alle Parameter werden unabhängig geschätzt, nur ein Wert pro Gruppe für den Methodenfaktor), Un-M-Meth (wie Un-M, jedoch zwei Werte pro Gruppe für die Methodenfaktoren), *SRMR* = standardisiertes Root Mean Square Residual, *GFI* = Goodness of Fit Index, *RMSEA* = Root Mean Square Error of Approximation (CI = Konfidenzintervall), *NNFI* = Nonnormed Fit Index, *CFI* = Comparative Fit Index, *AIC* = Akaike Information Criterion (U = Unabhängigkeitsmodell, Z = Zielmodell, S = saturiertes Modell).

6 Diskussion

Das abschließende Kapitel dieser Arbeit widmet sich zunächst der Zusammenfassung und Interpretation der im vorangegangenen Kapitel dargestellten Ergebnisse (6.1). Im Anschluss daran werden Ursachen für die vorgefundenen Gruppenunterschiede bezüglich der LST-Koeffizienten jenseits des Einflusses der zeitlichen Instruktion diskutiert (6.2). Der dritte Abschnitt des Kapitels beleuchtet kritische Punkte bei der Durchführung (6.3). Im Anschluss werden geeignete Verfahren zur Findung von Testhälften dargestellt (6.4). Danach wird das Studiendesign diskutiert (6.5), bevor schließlich ein Fazit gezogen wird (6.6) und ein Ausblick angestellt wird.

6.1 Zusammenfassung und Interpretation der Ergebnisse

Als erstes bleibt festzuhalten, dass sich das BDI-V auch in dieser Studie wieder durch gute Messeigenschaften auszeichnete. Sowohl die interne Konsistenz als auch die durch die LST-Analysen geschätzten Reliabilitäten sind in hohem Maße überzeugend. Nahezu alle Items wiesen akzeptable bis gute Trennschärfen auf. Im Folgenden sollen nun die hypothesenrelevanten Ergebnisse der Studie zusammengefasst und interpretiert werden. Dabei wird zunächst die Frage nach der Auswirkung der Manipulation der zeitlichen Instruktionen besprochen (Kapitel 6.1.1). Anschließend werden die Auswirkungen des im Vergleich zur Vorgängerarbeit größer gewählten – Retest-Intervalls diskutiert (Kapitel 6.1.2).

6.1.1 Auswirkung der zeitlichen Instruktionen

Hinsichtlich der Auswirkungen der zeitlichen Instruktionen lassen die vorliegenden Ergebnisse *keine* klare Interpretation zu. Die Tatsache, dass bereits das Invarianzmodell einen guten bis akzeptablen Modellfit zeigt, spricht gegen die Annahme eines Einflusses der zeitlichen Instruktion. Der Modellfit konnte durch die Lockerung der Traitvarianz und der Varianz der Latent-State-Residuen (Schritt 2), anhand derer sich die Auswirkungen der zeitlichen Instruktionen hätten zeigen müssen, *nicht* verbessert werden. Eine signifikante Verbesserung der Modellgütekennwerte ergab sich durch die Lockerung der Restriktionen für die Messfehlervarianz und die Varianz der

Methodenfaktoren (Schritt 3). Die für die beiden zuletzt genannten Modelle errechneten LST-Koeffizienten lassen allerdings erhebliche Gruppenunterschiede erkennen: Die Traitkonsistenz liegt in der 3-Monats-Gruppe deutlich über der für die 14-Tage-Gruppe ermittelten, die Zeitspezifität hingegen ist in der 14-Tage-Gruppe höher. Um zu prüfen, ob möglicherweise die Testhälften für die Gruppenunterschiede verantwortlich zeichneten, wurde ein zusätzliches Modell gerechnet, dass die Annahme gleicher Varianzen der Methodenfaktoren innerhalb der Instruktionsgruppen verwarf und diese getrennt frei schätzen ließ. Dieses Modell erbrachte eine weitere signifikante Verbesserung des Modellfits. Auf Basis dieses Modell wurden für die 3-Monats-Gruppe die folgenden LST-Koeffizienten für *beide* Testhälften ermittelt: die Traitkonsistenz liegt bei .66, die Zeitspezifität bei .19, die Methodenspezifität bei .09 und die Reliabilität bei .94. In der 14-Tage-Gruppe liegt die Traitkonsistenz für die erste Testhälfte bei .62, für die zweite bei .55. Für die Zeitspezifität wurde in Testhälfte eins ein Wert von .25 ermittelt, in Testhälfte 2 von .22. Die Methodenspezifität für die erste Testhälfte beträgt .05, für die zweite .16. Die Reliabilität liegt in beiden Testhälften bei .93. Es ergeben sich also kaum Unterschiede hinsichtlich der Reliabilitäten, so dass *Hypothese 1b* hinsichtlich dieses Punktes als bestätigt angesehen werden kann. Verworfen werden muss sie hingegen hinsichtlich der Methodenspezifität, bei der sich große Unterschiede zeigten. Die Traitkonsistenz liegt in der 14-Tage-Gruppe für beide Testhälften unterhalb der für die 3-Monats-Gruppe ermittelten. Die Zeitspezifität hingegen liegt über dem Vergleichswert. Dieser mit *Hypothese 1a* konforme Befund zeigt sich also auch, wenn der unterschiedliche Einfluss der Methodenfaktoren berücksichtigt wird. Von einer Bestätigung der *Hypothese 1a* kann allerdings dennoch nicht gesprochen werden. Vielmehr gilt es im Folgenden zu prüfen, welche Eigenschaften der Versuchspersonen dazu geführt haben, dass für die einzelnen Testhälften v.a. in der 14-Tage-Gruppe derart unterschiedliche Varianzen erzielt wurden. Es kann nicht ausgeschlossen werden, dass diese Eigenschaften auch jenseits des Effekts der Testhälften dafür verantwortlich sind, dass die Traitkonsistenz und die Zeitspezifität für die Instruktionsgruppen unterschiedlich ausfallen. Auffällig ist darüber hinaus, dass in beiden

Instruktionsgruppen zu Messzeitpunkt 2 niedrigere Mittelwerte und Varianzen erzielt wurden, wobei auch dieser Effekt die 14-Tage-Gruppe in stärkerem Maße betraf.

Für die Kontrollfragenstichprobe konnte der inkrementelle Wert der Aufhebung der Gleichheitsannahme für die beiden Methodenfaktoren hingegen *nicht* festgestellt werden. Das Invarianzmodell weist zwar weitestgehend akzeptable Modellgütekennwerte auf, wobei der *SRMR*-Wert für die 14-Tage-Gruppe außerhalb des nach gängigen Konventionen als akzeptabel anzusehenden Korridors (siehe Tabelle 4.1) liegt. Dies mag tendenziell für einen Einfluss der zeitlichen Instruktionen sprechen, jedoch erbrachte die Lockerung der Varianz der Latent-State-Residuen und der Traitvarianz keine signifikante Verbesserung des Modellfits, was gegen eine Auswirkung der zeitlichen Instruktionen spricht. In der Kontrollfragenstichprobe ist es das Unabhängigkeitsmodell, welches am besten auf die Daten passt. Es zeigen sich dabei große Unterschiede hinsichtlich der LST-Koeffizienten, sowohl auf der Eben der Testhälften als auf der Ebene des Gesamttests. Für die Traitkonsistenz wurde in der 3-Monats-Gruppe ein Wert von .64 auf der Testhälftenebene ermittelt (.70 bei Testverlängerung), in der 14-Tage-Gruppe .55 (.61). Die Zeitspezifität liegt in der 3-Monats-Gruppe bei .20 (.22), in der 14-Tage-Gruppe bei .24 (.27). Die Methodenspezifität liegt bei .10 (.05) in der 3-Monats-Gruppe und .12 (.07) in der 14-Tage-Gruppe. Schließlich liegt die Reliabilität bei .94 (.97) für die 3-Monats-Gruppe und .91 (.95) für die 14-Tage-Gruppe. Auch wenn der Vergleich nicht unproblematisch ist, zeigt sich also ein ähnliches Muster hinsichtlich der Traitkonsistenz und der Zeitspezifität wie in der Gesamtstichprobe. Die Gruppenunterschiede hinsichtlich dieser beiden Faktoren konnten auch in dieser Unterstichprobe ermittelt werden. Die Befunde der Kontrollfragenstichprobe sprechen also auch keine klare Sprache hinsichtlich der Auswirkung der zeitlichen Instruktionen auf das Antwortverhalten. Vieles spricht dafür, dass die vorgefundenen Gruppenunterschiede hinsichtlich der LST-Koeffizienten durch hypothesenirrelevante Faktoren verursacht oder zumindest erhöht wurden. Die Kontrollfragenstichprobe spricht anders als die Gesamtstichprobe nicht dafür, dass diese Faktoren vornehmlich in Interaktion mit den Testhälften wirken. Vielmehr legt die Analyse dieser Unterstichprobe es nahe, Eigenschaften der

Versuchspersonen zu ermitteln, die in einer der Instruktionsgruppen deutlich stärker vertreten sind und bei denen davon ausgegangen werden kann, dass sie Traitkonsistenz und Zeitspezifität beeinflussen.

Heckmann (2008) hatte hinsichtlich der Traitkonsistenz einen Unterschied von fünf Prozentpunkten und von vier Prozentpunkten hinsichtlich der Zeitspezifität zwischen den beiden Instruktionsgruppen in der Gesamtstichprobe feststellen können (S. 62). Der χ^2-Differenzentest zwischen dem Unabhängigkeitsmodell und dem Modell, welches Gruppenunterschiede hinsichtlich der Traitvarianz und Varianz der Latent-State-Residuen zuließ erwies sich als auf 5%-Niveau signifikant, wohingegen die zusätzliche Lockerung der Restriktionen für die Messfehlervarianz und die Varianz der Methodenfehler keine weitere signifikante Verbesserung des Modellfits ergab (S. 54). Ein ähnliches Bild ergab sich für die Kontrollfragenstichprobe (S. 58). Heckmann konnte 2008 also einen wenn auch sehr geringen, so doch signifikanten Einfluss der zeitlichen Instruktionen feststellen (S. 61). Ein Vergleich von Heckmanns Resultaten mit den im Rahmen dieser Studie festgestellten ist nur schwer möglich, eine ähnlich klare Aussage kann nicht getroffen werden. Die Indizien sprechen allerdings insgesamt dafür, dass die in dieser Studie ermittelten Unterschiede hinsichtlich der LST-Koeffizienten nur in geringem Maße durch die zeitlichen Instruktionen verursacht wurden und in wesentlich stärkerem Maße auf andere Faktoren wie z.B. soziodemographische Eigenschaften der beiden Instruktionsgruppen, die verwendeten Testhälften oder die in Kapitel 4.3.2 beschriebene Intervention an Personen mit hohem BDI-V-Summenwert bzw. hohen Werten auf den Suizidalitätsitems zurückzuführen sind. Es können also gleiche Tendenzen wie bei Heckmann (2008) vermutet werden von einer Replikation der Befunde Heckmanns kann insgesamt aber nicht gesprochen werden. *Hypothese 1a*, die einen Unterschied der zeitlichen Instruktionen angenommen hatte kann nicht bestätigt werden.

6.1.2 Auswirkungen des Retest-Intervalls

Tabelle 6.1 gibt die LST-Koeffizienten des jeweils den besten Modellfit erzielenden Modells in der Gesamtstichprobe für die Vorgängerstudie von Heckmann (2008, S. 57) und diese Studie wieder. Es zeigen sich deutliche Unterschiede hinsichtlich der Traitkonsistenz, die in beiden Gruppen und für alle Testhälften unter den von Heckmann ermittelten Werten liegt. Korrespondierend dazu liegt die Zeitspezifität in dieser Untersuchung deutlich über der von Heckmann ermittelten. Dieser Befund spricht zugunsten von *Hypothese* 2. Allerdings muss dabei einschränkend festgehalten werden, dass es sich um einen Vergleich unterschiedlicher Modelle handelt. Der beschriebene Befund ergibt sich jedoch konsistent über alle gerechneten Modelle auch im direkten Vergleich. Die Werte der 14-Tage-Gruppe sind dabei allerdings nur begrenzt vergleichbar, da der Einfluss der Methodenfaktoren für die beiden Testhälften sehr unterschiedlich ausfällt. Es ist allerdings davon auszugehen, dass auch bei einem geringeren Einfluss des Methodenfaktors das Niveau der Traitkonsistenz für die zweite Testhälfte nicht dem von Heckmann erzielten entsprechen würde. Vor diesem Hintergrund kann *Hypothese* 2 für die Gesamtstichprobe als bestätigt angesehen werden.

Tabelle 6.1: Vergleich LST-Koeffizienten Mehrgruppenmodelle (Gesamtstichprobe)

		TKon	ZSpe	MSpe	Rel
Vorgänger-Studie	3-M	.78 (.84)	.07(.08)	.08(.04)	.92 (.96)
	14-T	.73 (.79)	.11 (.12)	.07 (.05)	.92 (.96)
Diese Studie	3-M	.66	.19	.09	.94
	14-T (TH1)	.62	.25	.05	.93
	14-T (TH2)	.55	.22	.16	.93

Anmerkungen: N (3-Monate) = 108 (Vorgängerstudie), *N* (14-Tage) = 117 (Vorgängerstudie); *N* (3-Monate) = 187 (Diese Studie), *N* (14-Tage) = 240; eingeklammert: Werte für den Gesamttest (nur in der Vorgängerstudie ermittelbar)

Tabelle 6.2 sind die entsprechenden Werte für die Kontrollfragenstichprobe zu entnehmen (Die Werte für die Vorgängerstudie wurden Heckmann, 2008, S.60 entnommen). Auch hier bestätigt der direkte Vergleich der am besten auf die Daten passenden Modelle die Annahmen von *Hypothese* 2. Die Unterschiede fallen insgesamt geringer aus als für die Gesamtstichprobe und für die 3-Monats-Gruppe geringer als für die 14-Tage-Gruppe. Beim letztgenannten Aspekt muss allerdings erneut auf die noch ausführlicher zu diskutierenden Probleme hinsichtlich der Stichprobe der 14-Tage-Gruppe hingewiesen werden. Einschränkend muss zu den Vergleichen der beiden Studien festgehalten werden, dass die Stichprobengröße bei Heckmann für beide Gruppen deutlich unter 100 lag, was wie in Kapitel 4.6.1 besprochen die Anwendung der ML-Methode als fragwürdig erscheinen lässt (siehe auch Heckmann, 2008, S. 50). Eine weitere gravierende Einschränkung der Vergleichbarkeit ergibt sich aus der Tatsache, dass für diese Studie die Kontrollfrage nur zu Messzeitpunkt zwei erhoben wurde, das Kriterium für die Aufnahme in diese Gruppe also weniger restriktiv war als bei Heckmann. Vor diesem Hintergrund sind die Befunde für die Kontrollfragenstichprobe nur bedingt vergleichbar.

Tabelle 6.2: Vergleich LST-Koeffizienten Mehrgruppenmodelle (Kontrollfrage)

		TKon	ZSpe	MSpe	Rel
Vorgänger-Studie	3-M	.67 (.75)	.11 (.12)	.09 (.06)	.87 (.93)
	14-T	.66 (.71)	.19 (.21)	.06 (.03)	.91 (.95)
Diese Studie	3-M	.64 (.70)	.20 (.22)	.10 (.05)	.94 (.97)
	14-T	.55 (.61)	.24 (.27)	.12 (.07)	.91 (.95)

Anmerkungen: N (3-Monate) = 44 (Vorgängerstudie), *N* (14-Tage) = 57, *N* (3-Monate) = 104 (Diese Studie), *N* (14-Tage) = 162 (Diese Studie); eingeklammert: Werte für den Gesamttest.

Rückschlüsse auf die Hypothesen erscheinen daher fragwürdig. Die Tendenz zeigt allerdings in die angenommene Richtung.

Eine eindeutige Aussage zu *Hypothese 2* ist vor diesem Hintergrund ebenfalls nur begrenzt möglich. Letztlich sprechen die Indizien allerdings dafür, dass die Traitkonsistenz in dieser Studie niedriger lag als in der Vorgängerstudie und die Zeitspezifität höher.

Dieser Befund legt den Schluss nahe, dass intraindividuelle Unterschiede hinsichtlich der Depressivität aufgrund des höheren Abstands zwischen den beiden Messzeitpunkten im Vergleich zur Vorgängerarbeit von Heckmann stärker zum Tragen kommen konnten. Allerdings bleibt auch festzuhalten, dass weiterhin weitaus mehr als die Hälfte der Varianz des Antwortverhaltens durch stabile Eigenschaften der Personen erklärt werden können. Auch Schmitt & Maes (2000) hatten für die Traitkonsistenz mit .64 einen Wert ermittelt, der deutlich über .50 lag. Auch wenn die Befunde dieser Studie nur stark eingeschränkt mit denen von Schmitt & Maes (2000) verglichen werden können, da die zitierten Autoren sowohl eine andere zeitliche Instruktion („Wie ist ihr gegenwärtiges Lebensgefühl?“; S. 39) als auch einen anderen

Retest-Abstand (2 Jahre, S. 40) gewählt hatten, so spricht in der Gesamtschau der Studien von Schmitt & Maes (2000), Heckmann (2008) und der vorliegenden doch vieles dafür, dass das BDI-V vorwiegend stabile Eigenschaften misst und dies unabhängig von der zeitlichen Instruktion.

6.1.3 Interpretation und Ausblick

Auch wenn die Befunde dieser Arbeit nicht eindeutig interpretierbar sind, spricht die Gesamtschau aus der Studie von Heckmann und der vorliegenden nur für einen sehr geringen Einfluss der zeitlichen Instruktionen auf das Antwortverhalten im BDI-V. Dies wirft die Frage auf, warum dem so ist. Heckmann sah den Übergang vom episodischen zum semantischen Emotionswissen, den die Versuchsteilnehmerinnen und –teilnehmer in den Studien von Robinson & Clore (2002a, b) bereits bei der Instruktion "last few weeks" erkennen ließen als plausibelste Erklärung (Heckmann, 2008, S. 63). Werden Probandinnen und Probanden dazu aufgefordert, sich auf die letzten 14 Tage oder die letzten 3 Monate zu beziehen, so wenden sie in beiden Fällen semantisches Emotionswissen zur Beantwortung der Frage an. Dies erklärt aus der Sicht von Heckmann sowohl die geringen Unterschiede zwischen den Instruktionsgruppen als auch den Allgemeinbefund der hohen Traitkonsistenz. Insofern auch der Umkehrschluss als gültig erachtet wird, dass zeitliche Instruktionen, die einen kürzeren Bezugszeitraum vorgeben zur verstärkten Anwendung von episodischem Emotionswissen führen und die Antworten daher weniger allgemeine Ansichten über die eigenen Emotionen reflektieren, müsste die Traitkonsistenz in diesem Fall deutlich niedriger liegen und die Zeitspezifität deutlich höher als bei Bezugszeiträumen, die hinreichend groß sind. Sofern diese Schlussfolgerungen gültig sind, ist Heckmanns Erklärung aus zwei Gründen zu hinterfragen. Erstens bietet sie keine Erklärung für den Einfluss des Retest-Abstands, da lediglich der Bezugszeitraum als relevanter Faktor benannt wird. Wie beschrieben liegen die im Rahmen dieser Arbeit ermittelten Koeffizienten für alle gerechneten Modelle deutlich unter den von Heckmann ermittelten. Zweitens müsste das Antwortverhalten im BDI-V bei einer Instruktion, die einen deutlich kürzeren Zeitraum als die von Robinson & Clore (2002a, b) verwendeten "last few weeks" erfragt vornehmlich durch das episodische

Emotionswissen geprägt sein und daher weitaus niedrigere Werte für Traitkonsistenz und höhere für Zeitspezifität erzielen. Schmitt & Maes (2000) hatten in ihrer Studie die Instruktion: „Wie ist Ihr gegenwärtiges Lebensgefühl?“ verwendet (S. 40). Dabei wurden Werte für Traitkonsistenz von .64 und für Situationsspezifität von .26 errechnet. Diese Werte liegen um .15 bis .20 über (*TKon*) bzw. um .14 bis .18 unter (*ZSpe*) den von Heckmann ermittelten. Vergleichspunkt ist hierbei das im 2. Schritt aufgestellte Modell in Heckmanns (2008) Arbeit (S. 57). Allerdings liegen sie unabhängig davon, welches Modell betrachtet wird in etwa auf dem Niveau, das für die besagten Koeffizienten in dieser Arbeit ermittelt wurde (vgl. Tabellen 5.14 und 5.15), in jedem Fall nicht darunter, wie vor dem genannten theoretischen Hintergrund zu erwarten wäre. Dies widerspricht der Erklärung Heckmanns. Jenseits der vergleichenden Perspektive erscheint auch die absolute Zahl von .64 (Schmitt & Maes, 2000) für die Traitkonsistenz im Widerspruch zu Heckmanns Erklärungsansatz, werden doch immerhin annähernd zwei Drittel der Varianz des Antwortverhaltens von stabilen Eigenschaften erklärt. Ein alternativer Erklärungsansatz wäre z.B. dass die zeitliche Instruktion, die Schmitt & Maes (2000) gewählt haben, letztlich doch eher allgemeine Überzeugungen über die eigenen Emotionen abfragt. Es kann an dieser Stelle nur gemutmaßt werden, aber möglicherweise wird das Wort „Lebensgefühl“ allgemein oder im Zusammenspiel mit den BDI-V-Items in einer Art und Weise verstanden, die von tagesaktuellen Emotionen ablenkt und zu einer Gesamtschau verleitet, in der Ereignisse und Emotionen der vergangenen Wochen mit einbezogen werden. Wird dies als plausibel erachtet, wären Untersuchungen mit Alternativinstruktionen wie z.B.: „Wie geht es Ihnen im Augenblick?“ oder „Wie fühlen Sie sich heute?“ in Erwägung zu ziehen. Für diese Instruktion müsste natürlich die Standardisierung des Antwortformats geändert werden. Statt von „nie“ bis „fast immer“ könnte z.B. „Trifft überhaupt nicht zu“ („0“) und „Trifft voll und ganz zu“ („5“) verwendet werden. Unabhängig von der Instruktion könnte auch das Antwortformat mitverantwortlich sein für den hohen Traitanteil der Antworten im BDI-V. So sind die Ankerpunkte „nie“ und „fast immer“ nach Auffassung des Autors *semantisch* nur begrenzt mit der Frage nach einem auf den aktuellen Tag oder gar nur die aktuelle Stunde bezogenen Befinden kompatibel. Auch dies spräche für das oben

vorgeschlagene alternative Antwortformat, falls eine höhere Situationsspezifität der Antworten erzielt werden soll.

6.2 Gruppenunterschiede

Vor dem Hintergrund der Tatsache, dass die Unterschiede zwischen den Instruktionsgruppen hinsichtlich der Latent-State-Trait-Koeffizienten nicht oder nur eingeschränkt im Sinne der Hypothesen auf die Manipulation der zeitlichen Instruktionen zurückgeführt werden können, gilt es Einflussfaktoren zu diskutieren, welche diese Unterschiede erzeugt bzw. erhöht haben. Dabei werden zunächst Faktoren im Zusammenhang mit der Stichprobe ins Auge gefasst (Kapitel 6.2.1), bevor die Testhälften einer näheren Betrachtung unterzogen werden (Kapitel 6.2.2). Schließlich wird auch die Frage nach potenziellen Auswirkungen der Intervention an hochdepressiven Personen diskutiert (Kapitel 6.2.3), bevor ein Fazit gezogen wird (6.2.4).

6.2.1 Stichprobe

In diesem Abschnitt wird es darum gehen zu prüfen, inwieweit die vorgefundenen Gruppenunterschiede hinsichtlich der LST-Koeffizienten post hoc auf Gruppenunterschiede zwischen 3-Monats- und 14-Tage-Gruppe zurückgeführt werden können. Wie in Kapitel 4.1.3 dargelegt unterscheiden sich die 14-Tage-Gruppe und die 3-Monats-Gruppe signifikant hinsichtlich des Geschlechts und der beruflichen Tätigkeit, welcher die Personen zum ersten Erhebungszeitpunkt nachgingen. In der 14-Tage-Gruppe zeigte sich hierbei ein deutlich höherer Anteil Studierender (43% gegenüber 28% in der 3-Monats-Gruppe). Hier könnte ein möglicher Ansatzpunkt für die vorgefundenen Gruppenunterschiede liegen. Dieser ergibt sich vor allem im Zusammenspiel mit dem Untersuchungszeitraum. Der erste Befragungszeitraum (Anfang Juli bis Ende August) ist für einen beträchtlichen Teil der Studierenden der Zeitraum, in dem sie Prüfungen zu absolvieren und sich auf diese vorzubereiten haben, während der zweite Erhebungszeitraum (Anfang Oktober bis Anfang Dezember) den gewöhnlich weniger belastenden Anfangszeitraum des Semesters

umfasst. Ein möglicher Erklärungsansatz wäre also, dass während des ersten Erhebungszeitraums insgesamt ein höherer Anteil an Personen in der 14-Tage-Gruppe aufgrund von Prüfungsängsten höhere Depressivitätswerte als zum zweiten Erhebungszeitpunkt aufwies. Da sicherlich nicht alle Personen in der 14-Tage-Gruppe Prüfungen zu absolvieren hatten bzw. Prüfungsangst nicht bei allen Personen gleich ausgeprägt ist würde dies auch erklären, warum die Varianz zum Messzeitpunkt 2 in der 14-Tage-Gruppe deutlich niedriger ausfällt. Dies deckt sich beispielsweise mit den Befunden von Bernice & Wilding (2004), die bei britischen Studierenden einen deutlichen Zuwachs hinsichtlich Depressivität und Ängstlichkeit zwischen dem Trimesterbeginn und den "mid term exams" nachweisen konnten. Vor diesem Hintergrund wäre hierin ein möglicher von der Manipulation der zeitlichen Instruktionen unabhängiger Faktor zu sehen, der die Zeitspezifität erhöht und die Traitkonsistenz niedriger ausfallen lässt. Für die Gruppe der Berufstätigen, die in der 3-Monats-Gruppe einen höheren Anteil aufwies können derartige Stressoren nur bedingt angenommen werden. Zwar könnte der umgekehrte Effekt postuliert werden, da der erste Erhebungszeitraum für die meisten Berufstätigen in die Urlaubszeit fallen dürfte, die in der Regel zu einem Rückgang der Depressivität und Ängstlichkeit führt (Rau et. al., 2008), allerdings sind Urlaubszeiträume in der Regel deutlich kürzer als Prüfungszeiträume und der Effekt von Urlaubszeiten auf die Depressivität kann auch ein negativer im Sinne verstärkter Depressivität sein (Baier, 1987). Der hohe Anteil weiblicher Versuchspersonen in der 14-Tage-Gruppe könnte in Interaktion mit dem hohen Studierendenanteil den beschriebenen Effekt noch verstärken, da Frauen insgesamt zu höherer Prüfungsangst neigen (Eum & Rice, 2011). Wäre diese Interpretation zutreffend, müssten die Depressivitätswerte in der 14-Tage-Gruppe zum Messzeitpunkt zwei allerdings signifikant unter den Werten von Messzeitpunkt eins liegen. Dies wurde mittels ANOVA überprüft. Die Mittelwerte der BDI-V-Summenscores der 14-Tage-Gruppe zu Messzeitpunkten zwei unterscheiden sich nicht signifikant von denen zu Messzeitpunkt eins (F (1, 480) = 1.79, $p > 0.05$). Dies widerspricht der Annahme eines Effekts des Prüfungszeitraums.

Unabhängig vom hohen Studierendenanteil könnte eine höhere Sensibilität von Frauen für depressivitätsrelevante Änderungen ihres persönlichen Umfelds oder ihres Befindens ebenfalls dazu führen, dass die Situationsspezifität in der 14-Tage-Gruppe erhöht ist. Dies kann nur vermutet werden, empirische Belege konnten in der Literatur nicht gefunden werden.

Eine alternative Erklärung wäre ein höherer Anteil Psychologie-Studierender in der 14-Tage-Gruppe. Es wäre denkbar, dass Psychologie-Studierende entweder in höherem Maße sensibel für Veränderungen ihres Umfelds und/oder ihres Befindens sind, auch wenn sich dafür keine Belege in der Literatur finden ließen. Wie in Kapitel 4.1.2 erwähnt wurden die Psychologie-Studierenden der zwei zum Erhebungszeitpunkt jüngsten Kohorten der Universität Koblenz-Landau, Campus Landau ebenfalls um ihre Teilnahme gebeten. Eine Kohorte wurde dabei der 14-Tage-Gruppe zugeordnet, die andere der 3-Monats-Gruppe. Studierende im Hauptstudium wurden ebenfalls angeschrieben, aber lediglich darum gebeten, den Link an Bekannte weiterzureichen. Es kann post-hoc weder ausgeschlossen noch festgestellt werden, ob sich dabei einer der kontaktierten Jahrgänge teilnahmefreudiger zeigte.

Sacco (1981) und Zimmerman (1986) hatten berichtet, dass die Stabilität des BDI-I bei studentischen Stichproben nur sehr gering ausfiele und daher davor gewarnt das BDI bei studentischen einzusetzen. Zwar handelt es sich bei der 14-Tage-Gruppe nicht um eine rein studentische Stichprobe, dennoch ist, wie dargelegt, der Anteil Studierender in dieser Gruppe signifikant höher als in der 3-Monats-Gruppe und mit 43% auch absolut gesehen sehr hoch. Dies wäre eine mögliche Erklärung für die erhöhte Zeitspezifität in der 14-Tage-Gruppe, allerdings erklärt es den Einfluss des Methodenfaktors in der Gesamtstichprobe nicht.

6.2.2 Testhälften

In Kapitel 5.1.2 wurde bereits dargestellt, dass im Rahmen dieser Studie nur eingeschränkt und in deutlich höherem Maße für die 3-Monats-Gruppe gelungen ist, annähernd parallele Testhälfte zu finden. Bei der 14-Tage-Gruppe zeigten sich stark unterschiedliche Varianzen der Testhälften *innerhalb* der Messzeitpunkte (siehe

Tabellen 5.2 und 5.3). Es gilt zu prüfen, ob dies ein möglicher Erklärungsansatz für die Instruktionsgruppenunterschiede hinsichtlich der LST-Koeffizienten ist. Dies wäre beispielsweise dann der Fall, wenn sich soziodemographische Eigenschaften identifizieren ließen, die Unterschiede in der Beantwortung der Items des BDI-V implizieren. Eine mögliche solche Variable wäre das Geschlecht. Zu diesem Zweck wurden für die gesamte Analysestichprobe Mittelwerte, Varianzen und part-whole korrigierte Trennschärfen der Items des BDI-V nach Geschlechtern getrennt ermittelt (siehe Tabellen B.7 und B.8 im Anhang). Hierbei konnten im deskriptiven Vergleich große Geschlechtsunterschiede festgestellt werden. Insbesondere die Items Nr. 16 („Ich habe Schlafstörungen."), Nr. 18 („Ich habe keinen Appetit") und Nr. 20 („Sex ist mir gleichgültig.") fielen dabei in starkem Maße auf, da Frauen bei diesen Items jeweils eine um .50 höhere Varianz erzielten. Beide Items gehören der zweiten Testhälfte an. Aufgrund des höheren Frauenanteils in der 14-Tage-Gruppe kann dies als ein möglicher Erklärungspunkt dafür angesehen werden, dass die zweite Testhälfte des BDI-V in der 14-Tage-Gruppe zu beiden Messzeitpunkten eine weitaus höhere Varianz erzielte als die erste. Dies wäre eine Erklärung, warum das Modell, welches unterschiedliche Werte für die Methodenfaktoren annahm in der 14-Tage-Gruppe in der Gesamtstichprobe signifikant besser auf die Daten passte als das Unabhängigkeitsmodell. In der Kontrollfragenstichprobe wurde für die 14-Tage-Gruppe ein anderes Bild vorgefunden: Hier wies zu Messzeitpunkt eins die zweite Testhälfte eine deutliche *höhere* Varianz auf als die erste (siehe Tabelle 5.11), während zu Messzeitpunkt zwei die beiden Testhälften relativ gleichmäßige Varianzen erzielten, wobei die der zweiten Testhälfte etwas *geringer* ausfiel. Um dies zu erklären wurden Itemstatistiken für die Kontrollfragengruppe zu Messzeitpunkt eins nach Geschlechtern getrennt ermittelt (siehe Tabelle B.9 und B.10 im Anhang). Hierbei zeigte sich, dass der Geschlechtsunterschied hinsichtlich der Items in der Kontrollfragenstichprobe deutlicher ausfiel und gleichmäßiger über die Testhälften verteilt war. Dies erklärt, warum das Modell mit verschiedenen Werten für die Methodenfaktoren in der Kontrollfragenstichprobe keinen besseren Modellfit erzielte. Über die Ursachen kann nur gemutmaßt werden. Chi^2-Test und ANOVA ergaben lediglich einen signifikanten Unterschied hinsichtlich des Bildungsniveaus für die

Kontrollfragenstichprobe. Während Haupt- und Realschüler die Kontrollfrage mehrheitlich falsch beantworteten, lag der Anteil richtiger Antworten bei Abiturienten und Hochschulabsolventen jenseits der 60%. Größere Sorgfalt und höheres Bildungsniveau scheinen also in diesem Fall dazu geführt zu haben, dass die Geschlechtsunterschiede deutlicher wurden und sich relativ gleichmäßig über beide Testhälften verteilten.

6.2.3 Intervention an hochdepressiven Versuchsteilnehmern

Die Intervention an Personen mit erhöhten BDI-V-Werten und befürchteter Suizidalität wäre eine weitere mögliche Erklärung, warum die Varianz in beiden Instruktionsgruppen zum Messzeitpunkt zwei niedriger lag als zum ersten, wenn davon ausgegangen wird, dass die betroffenen Personen beim zweiten Messzeitpunkt niedrigere Summenwerte erzielten. Eine mögliche Erklärung wäre es in diesem Fall, dass die betroffenen Personen einen erneuten Hinweis auf eine möglicherweise vorliegende Erkrankung vermeiden wollten und daher zur Dissimulation tendierten. Des Weiteren wäre es möglich, dass die betreffenden Personen aufgrund des Rats durch den Autor dieser Arbeit sich in therapeutische Hilfe begeben haben und daher eine Besserung erzielten. Beesdo & Wittchen (2006) beziffern die durchschnittliche Dauer einer unbehandelten depressiven Episode auf drei bis vier Monate und schätzen den Anteil der anschließend vollremittierten Personen auf etwa 70-80%. Vor diesem Hintergrund wäre eine Besserung unabhängig von der Intervention durch den Autor in gleichem Maße denkbar. Von den kontaktierten Personen nahmen 22 in der 14-Tage-Gruppe und 18 in der 3-Monats-Gruppe zu beiden Zeitpunkten an der Studie teil. Da sich Analysegruppe und Dropoutgruppe nicht signifikant hinsichtlich der Depressivität unterschieden und die Anzahl der betroffenen Personen nicht sehr groß war dürfte der Faktor nur wenig Erklärungskraft hinsichtlich der vorgefundenen Auffälligkeiten der Ergebnisse besitzen.

6.2.4 Fazit bezüglich der Gruppenunterschiede

Insgesamt zeigte sich, dass die deutlichen Instruktionsgruppenunterschiede hinsichtlich des Geschlechts in starkem Maße zu Buche schlagen. Diese erklären insbesondere die Auffälligkeiten hinsichtlich der verwendeten Testhälften in der Gesamtstichprobe. Inwieweit davon ausgegangen werden kann, dass Frauen insgesamt zu einer stärker situationssensiblen Beantwortung von Fragebögen neigen konnte durch Literaturrecherche nicht geklärt werden. Sofern dies angenommen werden könnte, wäre es eine Erklärung für die erhöhte Situationsspezifität in der 14-Tage-Gruppe jenseits eines möglichen Effekts der zeitlichen Instruktion. Studierende waren in der 14-Tage-Gruppe deutlicher stärker vertreten als in der 3-Monats-Gruppe. Sacco (1981) und Zimmerman (1986) hatten berichtet, dass die Stabilität des BDI-O bei studentischen Stichproben sehr niedrig sei. Sofern dieser Befund auf das BDI-V übertragen werden kann, wäre es eine naheliegende Erklärung für die erhöhten Werte bei der Situationsspezifität in der 14-Tage-Gruppe. Es wäre in weiteren Studien zu prüfen, inwieweit sich das BDI-V bei studentischen Stichproben tatsächlich eine höhere Zeitspezifität aufweist.

6.3 Kritische Punkte bei der Durchführung

In diesem Abschnitt sollen zwei Aspekte der Durchführung diskutiert werden, die als kritisch angesehen werden müssen. Dies sind die zusätzlich erhobenen Items (6.3.1) und die Intervention an hochdepressiven und potenziell suizidgefährdeten Personen (6.3.2).

6.3.1 Zusätzlich erhobene Items

Wie in Kapitel 4.2 bereits erwähnt wurden im Rahmen dieser Studie zwei zusätzliche Items im Anschluss an das BDI-V erhoben. Diese lauteten: „Ich sehne mich nach dem Tod.“ und „Ich bin des Lebens überdrüssig.“ Da diese erst im Anschluss an das BDI-V abgefragt wurden kann vermutlich von einer starken Beeinflussung der Versuchspersonen zum Messzeitpunkt eins nicht ausgegangen werden. Problematisch könnten die zusätzlichen Items allerdings hinsichtlich des Gesamteindrucks, den der

Fragebogen bei den Versuchspersonen hinterließ sein. Die gestellte offene Frage, was die Teilnehmerinnen und Teilnehmer als Gegenstand dieser Studie vermuten enthielt des Öfteren die Vermutung, es handle sich um eine Studie zur Suizidalität. Auch in die Kommentarzeile kritisierten einige Personen, die gestellten Fragen seien „zu düster". Über die Auswirkungen dieses Eindrucks kann nur gemutmaßt werden. Sofern davon ausgegangen wird, dass sich eine beträchtliche Zahl der Versuchspersonen auch nach drei Monaten noch an den zu Messzeitpunkt eins gewonnenen Eindruck erinnern, mag es sein, dass für einige von ihnen die Studie daher einen aversiven Charakter annahm und dadurch der Dropout erhöht wurde. Darüber hinaus kann nicht ausgeschlossen werden, dass einige Personen die Hypothese entwickelten, die erhobenen Items seien *alle* Indikatoren für Suizidalität und dadurch dazu verleitet wurden eher etwas niedrigere Werte anzugeben als eigentlich intendiert. Allerdings dürfte der große Retest-Abstand dem tendenziell entgegen gewirkt haben. Kommentare der Versuchspersonen wie: „Die Frage danach, ob ich mich nach dem Tod sehne, kam etwas plötzlich und passte nicht zu den anderen Fragen." oder: „Ich konnte mich an den ersten Teil der Studie kaum mehr erinnern." mögen Hinweise in diese Richtung sein, da die Kommentarmöglichkeit nur zum Messzeitpunkt zwei vorhanden war.

6.3.2 Intervention bei hohem BDI-V-Wert und befürchteter akuter Suizidalität

Der in Kapitel 6.2.3 bereits im Hinblick auf die Ergebnisse der Studie diskutierte Umgang mit Personen, die einen hohen BDI-V-Wert oder/und bei zwei der drei erhobenen Suizidalitätsitems mindestens den Wert vier angeben ist natürlich methodisch fragwürdig. Neben den bereits beschriebenen möglichen Dissimulations- oder Aggravationstendenzen kann nicht ausgeschlossen werden, dass im Rahmen dieser Studie Personen in den entsprechenden Foren von der Nachricht durch den Autor dieser Studie berichtet hatten und sich deswegen in bestimmten Gruppen weniger oder andere Personen zur Teilnahme an der Studie entschließen konnten und dadurch die Ergebnisse verzerrt wurden. Die Validität der Ergebnisse dieser Studie ist dadurch natürlich potenziell beeinträchtigt. Vor dem Hintergrund der Tatsache, dass wie dargestellt, die psychologische Diagnostik auf valide und reliable Instrumente

angewiesen ist, darf der Faktor der Validität dieser Studie dabei nicht als ethisch unbedeutend angesehen werden (siehe z.B. Bortz & Döring, 2006, S.113). Dem steht gegenüber, dass möglicherweise Menschen, die sich in einer depressiven Episode befanden auf diese Art und Weise dazu bewogen wurden sich Hilfe zu suchen und sich daher eine Besserung einstellen konnte. Vor diesem Hintergrund fällt die rückblickende Bewertung dieser ethischen Entscheidung nicht leicht. Ich tendiere letztlich dazu, für nachfolgende Studien mit dem BDI-V entweder von derartigen Eingriffen abzusehen oder aber wesentlich höhere Cut-off-Werte zu wählen.

6.4 Testhälften

Wie in den Kapiteln 6.2.1 und 6.2.2 diskutiert ist davon auszugehen, dass die Tatsache, dass sich das Auffinden geeigneter Testhälften, wie in Kapitel 4.5 beschrieben, derart aufwändig gestaltete mit der ungleichmäßigen Verteilung der diskutierten soziodemographischen Eigenschaften in den beiden Instruktionsgruppen zusammen hängt. Dennoch ist vor diesem Hintergrund auch die Frage nach geeigneteren Verfahren zur Bestimmung annähernd paralleler Testhälften zu diskutieren. Little, Cunningham, Shahar & Widaman (2002) sehen eine wesentliche Grundvoraussetzung zum adäquaten Einsatz von Itempäckchen ("parcels") in der Klärung der Frage, ob das zu modellierende Konstrukt uni- oder multidimensional ist. Bei multidimensionalen Konstrukten sehen sie das Problem, dass die zur Messung dieser Konstrukte verwendeten Items häufig selbst multidimensional sind, also auf mehrere Faktoren relativ hoch laden (S. 163). Werden solche Items verwendet, um Itempäckchen zu bilden, so besteht die Gefahr, dass diese Päckchen ihrerseits selbst multidimensional sind und die Varianzstruktur des durch sie gemessenen latenten Konstrukts daher schwer zu interpretieren ist: "when a latent variable is defined with multidimensional parcels, one can never be completely sure as to what the latent construct 'really' is" (S. 163). Sofern die Items klar unterschiedlichen Facetten des multidimensionalen Konstrukts zugeordnet werden können, schlagen sie mit Bezug auf die Arbeit von Kishton & Widaman (1994) zwei verschiedene Verfahren vor: das erste sieht die Facetten als Gruppierungsmerkmal vor, d.h. es werden die Items zu einem Päckchen

zusammengefasst, die auf eine Facette hoch laden. Pro Facette entsteht so ein Itempäckchen, die multidimensionale Natur des Konstrukts bleibt erhalten, die eigenständige Beitrag jeder Facette zum gemessenen Konstrukt höherer Ordnung ebenso (Little et. al., 2002, S. 167). Das zweite Verfahren beinhaltet, dass jede Facette innerhalb des Päckchens repräsentiert wird, bei drei Facetten und neun Items werden also drei Päckchen mit je einem Item pro Facette gebildet. Bezugnehmend auf die Studie von Kishton & Widaman (1994), bei der die beiden Alternativen gegeneinander mit einer Skala zur Messung des Lokus of Control getestet wurden ziehen Little et. al. (2002) die letztgenannte Alternative vor, da sich bei der erstgenannten instabile und inakzeptable Parameterschätzungen ergaben (S. 168). Die faktorielle Struktur des BDI ist umstritten. Richter, Werner & Bastine (1994) hatten berichtet, dass sich mehrfaktorielle Lösungen für das BDI-I bisher weder replizieren noch adäquat interpretieren lassen. Wishman, Perez & Ramel (2000) hingegen haben für das BDI-II zwei Faktoren extrahiert, von denen der erste kognitiv-affektive Symptome misst, der zweite somatische. Schmitt & Maes (2000) hingegen hatten für das BDI-V einen Hauptfaktor extrahiert, der 36% der Itemvarianz erklärte, während mehrfaktorielle Lösungen weder durch schiefwinklige noch durch orthogonale Rotation zu einer Einfachstruktur geführt haben (S. 39). Die Autoren bestätigten daher die Bewertung von Richter, Werner & Bastine (1994) für das BDI-V.

Sofern sich dieser Befund replizieren ließe, wäre also davon auszugehen, dass das BDI-V ein unidimensionales Konstrukt misst. Für unidimensionale Konstrukte sehen Little et. al. (2002) zwei wesentliche Größen, die es zu beachten gilt, nämlich die Itemschwierigkeit und die Faktorladungen (S. 166). Sofern keine großen Schwierigkeitsunterschiede bestehen, besteht die Aufgabe bei der Auffindung geeigneter Itempäckchen aus Sicht der Autoren darin, die Höhe der durchschnittlichen Faktorladungen der einzelnen Items in den Itempäckchen auszubalancieren. Zu diesem Zweck werden zunächst die Items mit den höchsten Faktorladungen auf die Päckchen verteilt, dann die nächsthöchsten, wobei die Zuteilung zu den Itempäckchen in umgekehrter Reihenfolge vollzogen wird, also das Item mit der höchsten Faktorladung wird dem dritten Päckchen zugeordnet usw. Sofern große

Schwierigkeitsunterschiede bestehen, gilt es gleichzeitig, die durchschnittlichen Schwierigkeiten in den Itempäckchen auszubalancieren. Das korrekte Vorgehen aus Sicht der Autoren wäre also bei jeder Untersuchung zunächst mit der Bestimmung der Faktorladungen der Items zu beginnen (siehe auch Little, 1997).

Die dritte Alternative wäre die Zuweisung der Items zu den Testhälften per Zufall (Little et. al., 2002, S. 165).

6.5 Bewertung des Designs

Im folgenden Abschnitt soll zunächst die Bevölkerungsrepräsentativität der Stichprobe diskutiert werden (6.5.1), bevor potenzielle Probleme von Online-Untersuchungen betrachtet werden (6.5.2). Der Abschnitt schließt mit Überlegungen zur Power der gerechneten Analysen der Strukturgleichungsmodelle (6.5.3).

6.5.1 Stichprobe

Das Ziel bestand darin, eine möglichst heterogene Stichprobe zu gewinnen. Vergleicht man die Gesamtstichprobe mit der bundesdeutschen Gesamtbevölkerung (Statistisches Jahrbuch, 2010), so ergibt sich in Bezug auf die Altersverteilung eine minimale Überrepräsentation der 18-20jährigen (2% Abweichung), die 21-39jährigen hingegen waren deutlich überrepräsentiert (33% Abweichung), die 40-59jährigen waren unterrepräsentiert (10% Abweichung), die 60-64jährigen waren leicht unterrepräsentiert (4% Abweichung), wohingegen die über 65jährigen deutlich unterrepräsentiert waren (22% Abweichung). Frauen waren in dieser Studie deutlich überrepräsentiert (17% Abweichung). Verheiratete waren deutlich unterrepräsentiert (17% Abweichung), Geschiedene leicht unterrepräsentiert (5% Abweichung), Ledige deutlich überrepräsentiert (20% Abweichung). Die im Rahmen dieser Studie untersuchte Stichprobe weicht also deutlich von der bundesdeutschen Bevölkerung ab. Das Ziel einer möglichst heterogenen Stichprobe kann nur bedingt als erreicht angesehen werden. Frauen und unter 40jährige sind deutlich überrepräsentiert.

6.5.2 Online-Untersuchung

Die gewählte Form der Online-Untersuchung hat eine Reihe von Vorteilen. So ist es durch die hohe Verbreitung des Internets in der Bevölkerung zeitlich und finanziell weniger aufwändig eine hohe Anzahl von Personen zu erreichen als durch Postversand oder Direktansprache möglich ist (Wright, 2005). Der zitierte Autor nennt aber auch einen gewichtigen Nachteil von Online-Untersuchungen: "For example, relatively little may be known about the characteristics of people in online communities, aside from some basic demographic variables, and even this information may be questionable" (Wright, 2005, ohne Seitenangabe). Dieses Problem ergab sich im Rahmen dieser Studie vor allem hinsichtlich der *yahoo!*- und *Google*-groups. Zwar sind Informationen über die Anzahl der durch den Verteiler erfassten Personen, der Anzahl der im vergangenen Zeitraum verschickten E-Mails und über das inhaltliche Thema der Gruppe erhältlich, doch bleibt auch eine Reihe von Fragen offen. Insbesondere kann nur sehr schwer abgeschätzt werden, wie viele Personen aus der jeweiligen Gruppe an der Untersuchung teilnehmen werden. Zwar reicht die Gruppengröße von einigen wenigen Personen bis zu mehreren Tausend, aber die Antwortrate einer Gruppe abzuschätzen ist beinahe unmöglich (Andrews et. al., 2003). Die Sorge darum, eine hinreichend große Stichprobe zu bekommen, lässt große Verteilergruppen attraktiv erscheinen. Allerdings ist davon auszugehen, dass Personen, die in der gleichen Gruppe sind einander hinsichtlich weiterer Eigenschaften neben dem Interesse am Thema der Gruppe ähnlicher sind als Personen in der Gesamtbevölkerung, was die Gefahr birgt, dass diese Eigenschaften in der Stichprobe überrepräsentiert sind. Dieses Problem könnte gelöst werden, wenn anstelle der Gruppe die einzelnen Mitglieder der Gruppe angeschrieben werden könnten. Dann wäre es möglich, eine zufällige Zuteilung zu den Untersuchungsgruppen auf Ebene der Einzelpersonen vorzunehmen. Allerdings erlauben Mitglieder von Online-Foren ihren Administratoren häufig nicht, die E-Mail-Adressen an Dritte weiterzugeben, so dass die Anzahl der potenziell erreichbaren Gruppen verkleinert würde bzw. innerhalb von Gruppen Selbstselektionsprozesse auftreten können, welche die Zusammensetzung der Untersuchungsgruppe ebenfalls verzerren könnten (Wright,

2005; Thompson et. al., 2003). Eine weitere denkbare Alternative wäre es, wenn großer Verteilerlisten in einer ähnlichen Studie angeschrieben werden, im Anschreiben die Hyperlinks zu beiden Untersuchungsgruppen mitzuschicken und die Versuchspersonen zu bitten, sich für eine der beiden zu entscheiden. Dies wäre eine Möglichkeit der Überrepräsentation der Personen aus einem bestimmten Verteiler in einer Untersuchungsgruppe entgegen zu wirken. Allerdings besteht dabei die Gefahr, dass die Personen beide Links weiterverfolgen, die Unterschiede zwischen den beiden Onlinefragebögen herausfinden und sich daher Hypothesen über den Gegenstand der Untersuchung bilden, was zu einer Verfälschung der Ergebnisse führen könnte. Ein weiteres Problem der Untersuchungen in Onlineforen ergibt sich aus der Tatsache, dass es möglich ist, dass einzelne Mitglieder eines Online-Forums nach dem Ausfüllen des Fragebogens einen Beitrag an dieses Forum senden, in dem sie diesen beschreiben und Mutmaßungen über den Gegenstand der Untersuchung äußern. Dies könnte weitere Selbstselektionsprozesse auslösen, beispielsweise könnte es Personen, die ein höheres Interesse am vermuteten Untersuchungsgegenstand haben in höherer Zahl an der Studie teilnehmen, während Personen, die das Thema nicht für interessant halten weniger an der Studie teilnehmen. Da die Items des BDI-V durchaus Rückschlüsse auf Themenbereiche wie negative Emotionen, Depressionen und Suizidalität zulassen wäre es außerdem denkbar, dass Personen, die hohe Depressivitätswerte bei sich vermuten aus Scham oder Angst vor negativen Bewertungen, Abstand von der eigentlich intendierten Teilnahme nehmen oder umgekehrt, dass sie in verstärktem Maße teilnehmen, weil sie sich Hilfe versprechen. Bortz & Döring (2006) weisen in diesem Zusammenhang darauf hin, dass „Probanden bei psychologischen Untersuchungen meist automatisch einen ‚Psychotherapeuten' oder gar ‚Psychiater' als Adressaten vermuten und somit (…) eine Individualdiagnose befürchten." (S. 232). Schließlich wäre es auch denkbar, dass innerhalb der Gruppe eine intensive inhaltliche Diskussion über das Thema geführt wird, was ebenfalls dazu führen kann, dass die Ergebnisse verzerrt werden. Zwar wäre es prinzipiell kontrollierbar, ob die Mitglieder einer angeschriebenen Mailingliste über die Untersuchung diskutiert haben, indem der Durchführende allen angeschriebenen Listen beitritt. Der zeitliche Aufwand der

Kontrolle, die das Lesen aller im Untersuchungszeitraum über die Verteilerlisten geschickten E-Mails impliziert hätte wäre allerdings nicht leistbar gewesen.

Abschließend zum Thema Online-Untersuchungen soll die Frage nach „Senioren im Internet“ kurz betrachtet werden. Wie in Kapitel 4.1.2 bereits dargestellt wurden Gruppen, die explizit ältere Menschen ansprechen oder deren thematische Ausrichtung ein älteres Klientel vermuten ließ bevorzugt angeschrieben. Zwar ist ein Wachstum bei der Altersgruppe der Menschen über 70, die das Internet nutzen zu verzeichnen, dennoch ist diese Altersgruppe im Internet weiterhin tendenziell unterrepräsentiert im Vergleich zu ihrem Anteil an der Gesamtbevölkerung (Yoon, Yoon & George, 2011). Vor diesem Hintergrund wäre es für nachfolgende Studien erwägenswert, gezielt ältere Menschen anzusprechen und dabei Papier-und-Bleistift-Versionen des Fragebogens zu verwenden.

6.5.3 Post hoc Power Analysen

Heckmann (2008) hatte unter Bezugnahme auf die Ausführungen von Kim (2005) ermittelt, dass für die Einzelgruppenmodelle, die jeweils sechs Freiheitsgrade aufwiesen für eine Power von .80 für einen *RMSEA*-Wert von .05 der kritische Wert für den Nonzentralitätsparameter bei 13.62 liege und daher auf Basis von Kims Formel [$N = (\delta_{1-\beta} / RMSEA^2 df) + 1$] ein kritisches N von 981 ermittelt (S. 68). Für das Invarianzmodell ($df = 16$) der Multi-sample-Analysen hatte sie 481 als Mindeststichprobe errechnet. Für die weiteren Multi-sample-Analysen liegen die kritischen Werte entsprechend höher, da sie weniger Freiheitsgrade aufweisen (Kim, 2005, zitiert nach Heckmann). Diese Anforderung an die Stichprobe wurde auch im Rahmen dieser Studie deutlich verfehlt, was die ermittelten Ergebnisse bis zu einem gewissen Grad zusätzlich hinterfragenswert erscheinen lässt. Allerdings ist auch darauf zu verweisen, dass genau aus diesem Grund in den beiden Studien eine Vielzahl von Fitindizes parallel überprüft wurde, von denen der Großteil nur in geringem Maße von der Stichprobengröße abhängig ist (Heckmann, 2008, S. 68). Wie in Heckmanns Studie war es auch in dieser Studie meistenteils so, dass die verwendeten Fitindizes unisono in die gleiche Richtung wiesen.

Heckmann (2008) gibt unter Berufung auf Kaplan (1995, zitiert nach Heckmann) als weiteren Einflussfaktor auf die Power die Unterschiedlichkeit der Stichprobengröße an (S. 68). Stevens (2002, zitiert nach Heckmann) nennt Gruppengrößen ausreichend ähnlich groß, wenn der Quotient aus der größeren und der kleineren den Wert von 1.5 nicht überschreitet. In der Gesamtstichprobe bestand die 14-Tage-Gruppe aus 240 Personen, die 3-Monats-Gruppe aus 187. Der Quotient beträgt also 1.22, weswegen die Gruppen als ausreichend ähnlich groß betrachtet werden können. Für die Kontrollfragenstichprobe (14-Tage-Gruppe 162 Personen, 3-Monats-Gruppe 104 Personen) beträgt der Quotient 1.55, wurde also knapp überschritten, was die Interpretierbarkeit der Befunde für diese Gruppe weiter einschränkt.

6.6 Fazit

Im Rahmen dieser Studie hat sich das BDI-V erneut als Messinstrument mit einer hohen Reliabilität ausgezeichnet. Die Tatsache, dass es zu einem hohen Anteil stabile Eigenschaft misst, aber gleichzeitig änderungssensitiv ist, macht es zu einem attraktiven Instrument für die klinische Praxis, da die klinische Diagnostik in hohem Maße darauf angewiesen ist festzustellen, ob vorgefundene Veränderungen tatsächlich als stabil anzusehen sind oder ob sie eher in tagesaktuellen Ereignissen ihre Ursache haben (Yousfi & Steyer, 2006). Aus diesem Grund bietet sich das BDI-V vor allem als Instrument für die Verlaufsdiagnostik, Therapieevaluation sowie für Katamneseuntersuchungen an. Wie in Kapitel 2.1.5 dargestellt ist es zweifelhaft, ob das Zweiwochenkriterium für Depressionen hinsichtlich des nosologischen Kriteriums der klinisch relevanten Funktionsbeeinträchtigung und des Leidens der Betroffenen als valide angesehen werden kann (Howland et. al., 2008; Nierenberg et. al., 2010). Die Ergebnisse der Studie von Heckmann und der hier vorgestellten werfen auch die Frage auf, in wie weit dieses Kriterium tatsächlich im diagnostischen Prozess reliabel umsetzbar ist, da in beiden Studien ein beträchtlicher Anteil der Probanden den abgefragten Bezugszeitraum nicht korrekt wiedergeben konnten. Ob ein Einfluss der zeitlichen Instruktionen besteht, konnte im Rahmen dieser Studie nicht klar festgestellt werden. Zwar wurden beträchtliche Unterschiede hinsichtlich der Latent-State-Trait-

Koeffizienten zwischen den Gruppen festgestellt, jedoch können diese nicht eindeutig auf die zeitlichen Instruktionen zurückgeführt werden. Vielmehr scheinen soziodemographische Unterschiede zwischen den beiden Stichproben in starkem Maße das Antwortverhalten beeinflusst zu haben. Als ein wesentlicher Faktor hat sich dabei das Geschlecht herauskristallisiert. Beide Untersuchungsgruppen wiesen einen Überhang weiblicher Versuchsteilnehmer auf, für die 14-Tage-Gruppe galt dies allerdings in wesentlich stärkerem Maße. Dies mag ein Faktor sein, der dazu beigetragen hat, dass sich die Kontraste zwischen Gruppen verstärkten. Ein generelles Fazit hinsichtlich der Auswirkung des zeitlichen Bezugsrahmens muss daher unter Vorbehalt ausgesprochen werden. In der Gesamtschau mit den Ergebnissen der Vorgängerstudie ergeben sich jedenfalls erste Hinweise, dass Kriterien wie das in DSM-IV und ICD-10 geforderte Zweitwochenkriterium für die depressiven Erkrankungen und das Zweijahreskriterium für die Dysthymie fragwürdig sind. Hinsichtlich der Auswirkungen des Retest-Intervalls kann trotz der genannten Einschränkungen letztlich konstatiert werden, dass das BDI-V bei größer gewähltem Retest-Abstand stärker durch situative Einflüsse und die Person-Situations-Interaktion beeinflusst wird.

6.7 Ausblick

Wie in den vorangegangenen Kapiteln dargestellt sind die Ergebnisse dieser Untersuchung insgesamt nicht klar interpretierbar und daher nur von begrenztem Wert. Es wäre daher erwägenswert, die Studie zu wiederholen. Wichtig wäre dabei, eine möglichst annähernde Gleichverteilung der Geschlechter zu erzielen. Darüber hinaus wäre eine bessere Repräsentation der Personen mit Haupt- und Realschulabschluss als höchstem Bildungsabschluss wünschenswert. Eine größere Stichprobe wäre ebenfalls anzustreben, damit eine bessere Power bei der Berechnung der Strukturgleichungsmodelle zu erzielt werden kann.

Des Weiteren wäre es interessant zu prüfen, wie hoch der Trait- und State-Anteil für das BDI-V bei der zeitlichen Anweisung, die Schmitt & Maes (2000) verwendet hatten

mit einem kleineren Retest-Abstand als die von den Autoren gewählten 2 Jahre wäre. Alternativ bzw. ergänzend dazu könnten andere, noch stärker den Momentaufnahmecharakter der Untersuchung betonende Anweisungen wie beispielsweise: „Wie geht es Ihnen im Augenblick?“ oder „Wie fühlen Sie sich heute“ getestet werden. Zum einen wäre dies interessant, um zu sehen, ob die Items des BDI-V bis zu einem gewissen Grad inhärent, also unabhängig von der zeitlichen Instruktion, vorwiegend stabile Eigenschaften messen. Zum anderen wäre es beispielsweise im Sinne der Feststellung von Allen & Potkay (1981), die Traits als Summation von State-Einheiten sahen, konsequent, Traitausprägungen durch Aggregation von State-Messungen zu messen. Sofern durch eine andere zeitliche Anweisung ein höherer State-Anteil für das BDI-V erzielt werden könnte, wäre es auch bei dieser Konzeption von Depressivität als State als Maß geeignet.

Des Weiteren wäre es von Interesse, ob sich das BDI-V ähnlich wie BDI-O als deutlich instabiler bei studentischen Stichproben im Vergleich mit der Gesamtbevölkerung erweist. Angesichts der ermittelten Geschlechtsunterschiede bei manchen Items stellt sich die Frage nach deren Validität. Messen Items wie „Sex ist mir gleichgültig.“ oder „Ich habe keinen Appetit“ tatsächlich Depressivität der Versuchsperson oder „messen“ sie gar deren Geschlecht?

Für die klinische Psychologie wären neben weiteren Untersuchungen zum Einfluss zeitlicher Instruktionen auf das Antwortverhalten bei klinischen Diagnoseinstrumenten, weitere Forschung zum Zusammenhang zwischen der Dauer depressiver Episoden und der psychosozialen Funktionsbeeinträchtigung von Interesse. Studien wie die von Howland et. al. (2008) und Nierenberg et. al. (2010) scheinen insgesamt den Schluss nahe zu legen, dass auch vergleichsweise kurze depressive Episoden beträchtliche Beeinträchtigungen hervorrufen können. Die Vorgängerstudie und in Grenzen auch diese Studie werfen Zweifel an der Diagnostizierbarkeit dieses Zweiwochenkriteriums für depressive Erkrankungen auf. Vor dem Hintergrund dieser Einschränkungen hinsichtlich Validität und Reliabilität erscheint das Zweiwochenkriterium insgesamt fragwürdig.

Literatur

Ahava, G. W., Iannone, C., Grebstein, L. & Schirling, J. (1998). Is the Beck Depression Inventory Reliable Over Time? An Evaluation of Multiple Test-Retest-Reliability in a Nonclinical College Student Sample. *Journal of Personality Assessment, 70,* 222-231.

Akaike, H. (1974). A new look at statistical model identification. *IEEE transactions on automatic control*, *19*, 716-723.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

Allen, B. P. & Potkay, C. R. (1981). On the arbitrary distinction between states and traits. *Journal of Personality and Social Psychology, 41,* 916-928.

Amelang, M. & Bartussek, D. (1997). *Differentielle Psychologie und Persönlichkeitsforschung*. Stuttgart: Kohlhammer.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention.* Heidelberg: Springer.

Andrews, D., Nonnecke, B. & Preece, J. (2003). Electronic survey methodology: A case study in reaching hard-to-involve Internet users. *International Journal of Human-Computer Interaction, 16,* 185-210.

Andrews, G., Anderson, T. M., Slade, T. & Sunderland, M. (2008). Classification of Anxiety and Depressive Disorders. *Depression and Anxiety, 25,* 274-281.

Angst, J. & Merikangas, K. R., (2001). Multi-dimensional criteria for the diagnosis of depression, *Journal of Affective Disorders, 62,* 7-15.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2006). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* Berlin: Springer.

Baier, M. (1987). The "holiday blues" as a stress reaction. *Perspectives in Psychiatric Care, 24,* 64-68.

Beck, A. T. (1978). *The depression inventory.* Philadelphia: Center for Cognitive Therapy.

Beck, A. T., Steer, R. A. & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A. & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77-100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561-571.

Beesdo, K. & Wittchen, H.-U. (2006) Depressive Störungen: Major Depression und Dysthymie. In: H.-U. Wittchen & J. Hoyer (Hrsg.), *Klinische Psychologie & Psychotherapie* (S. 731-762). Heidelberg: Springer.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin,* 88, 588–606.

Bernice, A. & Wilding, J. M. (2004). The relation of depression and anxiety to life-stress and achievement in students. *British Journal of Psychology, 95,* 509-521.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Brown, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J.S. Long (Hrsg.), *Testing structural equation models* (S. 136-162). Newbury Park, CA: Sage.

Brown, T. A. (2007). Temporal course and structural relationships among dimensions of temperament and DSM–IV anxiety and mood disorder constructs. *Journal of Abnormal Psychology, 116,* 313–328.

Chou, C.-P. & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Hrsg.), *Structural equation modeling: Concepts, issues and application (*S. 37-57). Thousand Oaks, CA: Sage.

Cooper, C. & McConville, C. (1990). Interpreting mood sources: Clinical implications of individual differences in mood variability. *British Journal of Medical Psychology, 63,* 215-225.

Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C. & Cole, D. A. (2008). Analyzing the Convergent and Discriminant Validity of States and Traits: Development and Applications of Multimethod Latent State-Trait Models. *Psychological Assessment, 20,* 270-280.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Curran, P. J., West, S. G. & Finch, J. F. (1996). The robustness of test statistics to nonnormality and error in Confirmatory Factor Analysis. *Psychological Methods, 1,* 16-29.

Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., Ostendorf, F. & Neubauer, A. (1995). Situational effect in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality, 9,* 1-23.

Dilling, H., Mombour, W., Schmidt, M. H. & Schulte-Markwort, E. (Hrsg.) (2006). *Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) – Diagnostische Kriterien für Forschung und Praxis.* Bern: Huber.

Eid, M. (1995). *Modelle der Messung von Personen in Situationen.* Weinheim: Beltz.

Eum, K. & Rice, K. G. (2011). Test anxiety, perfectionism, goal orientation, and academic performance. *Anxiety, Stress & Coping, 24,* 167-178.

Gershuny, B. S. & Sher, K. J. (1998). The relation between personality and anxiety: Findings from a 3-year prospective study. *Journal of Abnormal Psychology, 107,* 252-262.

Hank, P., Schwenkmezger, P. & Schumann, J. (2001). Daily mood reports in hindsight: Results of a computer-assisted time sampling study. In J. Fahrenberg & M. Myrtek (Hrsg.), *Progress in ambulatory assessment* (S. 143-156). Seattle WA: Hogrefe & Huber Publishers.

Hatzenbuehler, L. C., Parpal, M. & Matthews, L. (1983). Classifying college students as depressed or nondepressed using the Beck Depression Inventory: An empirical analysis. *Journal of Consulting and Clinical Psychology, 51,* 360-366.

Hau, K.-T. & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology, 57,* 327-351.

Hautzinger, M. (2010). *Akute Depression.* Göttingen: Hogrefe.

Hautzinger, M., Bailer, M., Worall, H. & Keller, F. (1994). *Beck-Depressions-Inventar (BDI).* Bern: Huber.

Hautzinger, M., De Jong-Meyer, R. (1996). Depression. *Zeitschrift für Klinische Psychologie, 26,* 76-160.

Hautzinger, M., Keller, F. & Kühner, C. (2006). *BDI II – Beck Depressions-Inventar – Manual.* Frankfurt am Main: Harcourt Test Services.

Hautzinger, M. & Meyer, T. D. (2002). *Diagnostik Affektiver Störungen.* Göttingen: Hogrefe.

Heckmann, N. (2008). *Einflüsse des zeitlichen Bezugsrahmens auf Angaben zur eigenen depressiven Befindlichkeit.* Nicht veröffentlichte Studienabschlussarbeit, Universität Koblenz-Landau, Landau.

Hogan, R. (2009). Much ado about nothing: The person-situation-debate, *Journal of Research in Personality, 43,* 249.

Howland, R. H., Schettler, P., Rapaport, M. H., Mischoulon, D., Schneider, T., Fasiczka, A., Delrahiem, K., Maddux, R., Lightfoot, M. & Nierenberg, A. A. (2008). Clinical Features and Functioning of Patients with Minor Depression. *Psychotherapy and Psychosomatics, 77,* 384-389.

Hu, L. & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle: *Structural equation modeling* (S. 76-99). Newbury Park, CA: Sage.

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model specification. *Psychological Methods, 3,* 424-453.

Janke, W. & Hüppe, M. (1991). Emotionalität. In: W. D. Oswald, W. M. Herrmann, S. Kanowski, U. Lehr & H. Thomae (Hrsg.), *Gerontologie*. Stuttgart: Kohlhammer.

De Jong-Meyer, R. (2005). Depressive Störungen: Klassifikation und Diagnostik. In M. Perrez & U. Baumann (Hrsg.), *Lehrbuch Klinische Psychologie – Psychotherapie* (S. 852-891). Bern: Huber.

Jöreskog, K. G. & Sörbom, D. (1984). *LISREL Vl user's guide.* Mooresville, IN: Scientific Software.

Jöreskog, K. G & Sörbom, D. (1993). *Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Lawrence Erlbaum.

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Jöreskog, K. G. & Sörbom, D. (2006a). LISREL Student Edition (Version 8.80) [Computer Software]. Chicago, IL: Scientific Software International.

Jöreskog, K. G. & Sörbom, D. (2006b). PRELIS Student Edition (Version 2.80) [Computer Software]. Chicago, IL: Scientific Software International.

Judd, L. L., Rapaport, M. H., Paulus, M. P. & Brown, J. L. (1994). Subsyndromal symptomatic depression: a new mood disorder? *Journal of Clinical Psychiatry, 55,* 18-28.

Kammer, D. (1983). Eine Untersuchung der psychometrischen Eigenschaften des Beck-Depressionsinventars (BDI). *Diagnostica, 29,* 48-60.

Kaplan, D. (2000). Statistical Power in Structural Equation Modeling. In R. H. Hoyle (Hrsg.), *Structural equation modeling* (76-99). Thousand Oaks: Sage.

Kelava, A. & Schermelleh-Engel, K. (2007). Latent-State-Trait-Theorie (LST-Theorie). In: H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion.* Heidelberg: Springer.

Kelley, K. & Lai, K. (2011). Accuracy in Parameter Estimation for the Root Mean Square Error of Approximation: Sample Size Planing for Narrow Confidence Intervals. *Multivariate Behavioral Research, 46,* 1-32.

Kessler, R. C. (1997). The effects of major life events on depression. *Annual Review of Psychology,* 48, 191-214.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E. & Wang, P.S. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association, 289,* 3095-3105.

Kessler, R. C., Zhao, S., Blazer, D. G. & Swartz, M. (1997). Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *Journal of Affective Disorder, 45,* 19-30.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling, 12,* 368-390.

Kishton, J. M. & Widaman, K. F. (1994). Unidimensional versus domaine representative parceling questionnaire items: An empirical example. *Education and Psychological Measurement, 54*, 757-765.

Krohne, H.-W., Egloff, B., Kohlmann, C.-W. & Tausch, A. (1996). Untersuchungen mit der einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*, 139-156.

Kühner, C., Bürger, C., Keller, D. & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II): Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt, 78*, 651-656.

Lei, M. & Lomax, R. G. (2005). The Effect of Varying Degress of Nonnormality in Structural Equation Modeling, *Structural Equation Modeling, 12*, 1-27.

LimeSurvey Version 1.86 [Computer Software] (2010). Schmitz, C. Zugriff am 19.6.2011. Verfügbar unter http://www.limsurvey.org.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling, 9*, 151-173.

Lovejoy, M. C. & Steuerwald, B. L. (1997). Subsyndromal unipolar and bipolar disorders II: comparison on daily stress levels. *Cognitive Therapy and Research, 21*, 607-618.

Lucas, R. & Donnellan, M. (2009). If the person-situation debate is really over, why does it still generate so much negative affect? *Journal of Research in Personality, 43*, 146-149.

Lukesch, H. (1974). Testkriterien des Depressionsinventars von A. T. Beck. *Psychologische Praxis, 18*, 60-78.

MacCallum, R. C., Browne, M. W. & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypothesis. *Psychological Methods, 11*, 19-35.

MacCallum, R. C. & Hong, S. (1997). Power Analysis in Covariance Structure Modeling Using GFI and AGFI. *Multivariate Behavioral Research, 32*, 193-210.

Marsh, H. W., Antill, J. K. & Cunningham, J. D. (1989) Masculinity and femininity: A bipolar construct and independent constructs. *Journal of Personality, 57*, 625-663.

Menon, G. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Sudman (Hrsg.), *Autobiographical memory and the validity of retrospective reports* (S. 161-172). New York: Springer.

Michael, T. & Margraf, J. (2003). Klassifikationssysteme. In: K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz.

Mohiyeddini, C., Hautzinger, M. & Baer, S. (2002). Eine Latent-State-Trait-Analyse zur Bestimmung der dispositionellen und zustandsbedingten Anteile dreier Instrumente zur Erfassung von Depressionen: ADS, BDI und SDS. *Diagnostica, 48*, 12-18.

Moosbrugger, H. & Schermelleh-Engel, K. (2007). Exploratische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In: Moosbrugger, H., Kelava, A. (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 307-324). Heidelberg: Springer.

Mulaik, S. A., James, L. R.., Van Alstine, J., Bennett, N., Lind, S. & Stillwell, C. D. (1989). An evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Nierenberg, A. A., Rapaport, M. H., Schettler, P. J., Howland, R. H., Smith, J. A., Edwards, D., Schneider, T. & Mischoulon, D. (2010). Deficits in Psychological Well-Being and Quality-of-Life in Minor Depression: Implications for DSM-V. *Neuroscience & Therapeutics, 16*, 208-216.

Random Sequence Generator [Computer Software] (1998). Haahr, M. Zugriff am 30.6.2011. Verfügbar unter http://www.random.org/sequences

Rapaport, M. H. & Judd, L. L. (1998). Minor depressive disorder and subsyndromal depressive symptoms: functional impairment and response to treatment. *Journal of Affective Disorders, 48*, 227-232.

Rau, R., Hoffmann, K., Metz, U., Richter, P. G., Rösler, U. & Stephan, U. (2008). Gesundheitsrisiken bei Unternehmern. *Zeitschrift für Arbeits- und Organisationspsychologie, 52*, 115-125.

Richter, P. (1991). *Zur Konstruktvalidität des Beck-Depressionsinventars bei der Erfassung depressiver Verläufe.* Regensburg: Roderer.

Richter, P., Werner, J. & Bastine, R. (1994). Psychometrische Eigenschaften des Beck-Depressionsinventars (BDI): Ein Überblick. *Zeitschrift für klinische Psychologie, 23*, 3-19.

Robinson, M. D. & Clore, G. L. (2002a). Belief and feeling: Evidence for an accesibilty model of emotional self-report. *Psychological Bulletin, 128*, 943-960.

Robinson, M. D. & Clore, G. L. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence from two judgement processes. *Journal of Personality and Social Psychology, 83*, 198-215.

Sacco, W. P. (1981). Invalid use of the Beck Depression Inventory to identify depressed college student subjects: A methodological comment. *Cognitive Therapy and Research, 5*, 143-147.

Saris, W. E., Satorra, A. & van der Feld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling, 16*, 561-582.

Saß, H., Wittchen, H.-U., Zaudig, M. & Houden, I. (2003). *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-IV: Textrevision – DSM-IV-TR: übersetzt nach der Textrevision der vierten Auflage des Diagnostic and statistical manual of mental disorders der American Psychiatric Association.* Göttingen: Hogrefe.

Schermelleh-Engel, K. & Moosbrugger, H. (2002). Beurteilung der Modellgüte von Strukturgleichungsmodellen. *Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität Frankfurt am Main, Heft 4/2002.*

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodnes-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74.

Schmitt, M. (1992). Interindividuelle Konsistenzunterschiede als Herausforderung für die Differentielle Psychologie. *Psychologische Rundschau, 43*, 30-45.

Schmitt, M. (2005). Interaktionistische Ansätze. In: H. Weber & T. Rammsayer (Hrsg.), *Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie.* Göttingen: Hogrefe.

Schmitt, M., Altstötter-Gleich, C., Hinz, A., Maes, J. & Brähler, E. (2006). Normwerte für das vereinfachte Beck-Depressions-Inventar (BDI-V) in der Allgemeinbevölkerung. *Diagnostica, 52,* 51-59.

Schmitt, M., Beckmann, M., Dusi, D., Maes, J., Schiller, A & Schonauer, K. (2003). Messgüte des vereinfachten Beck-Depressions-Inventars (BDI-V). *Diagnostica, 49*, 147-156.

Schmitt, M. & Maes, J. (2000.). Vorschlag zur Vereinfachung des Beck-Depressions-Inventars (BDI). *Diagnostica, 46,* 38-46.

Schmitt, M., Maes, J. & Schmal, A. (1999). Ungerechtigkeitserleben im Vereinigungsprozeß: Folgen für das emotionale Befinden und die seelische Gesundheit. In M. Schmitt & L. Montada (Hrsg.), *Gerechtigkeitserleben im wiedervereinigten Deutschland* (S. 169-212). Opladen: Leske + Budrich.

Schmitt, M. J. & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences, 14,* 519-529.

Schumacker, R. E. & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Erlbaum.

Schwarz, N. & Sudman, S. (Hrsg.). (1994). *Autobiographical memory and the validity of retrospective reports.* New York: Springer.

Statistisches Bundesamt (2010) (Hrsg.), *Statistisches Jahrbuch für die Bundesrepublik Deutschland 2010*. Wiesbaden

Steiger, J. H. (1990). Structural model evaluation and modification: An Interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modelling, 7,* 149-162.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. Mahwah: Lawrence Erlbaum.

Steyer, R. & Eid, M. (1993). *Messen und Testen.* Berlin: Springer.

Steyer, R., Ferring, D. & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 79-98.

Steyer, R. & Schmitt, M. J. (1990). The effects of aggregation across and within occasions on consistency, specifity and reliability. *Methodika, 4,* 58-94.

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent State-Trait Theory and Research in Personality and Individual Differences. *European Journal of Personality, 13,* 389-408.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). *Thinking about answers.* San Francisco, CA: Jossey-Bass.

Sumner, J. A., Griffith, J. W., Mineka, S., Rekart, K. N., Zinbarg, R. E. & Craske, M. G. (2011). Overgeneral autobiographical memory and chronic interpersonal stress as predictors of the course of depression in adolescents. *Cognition and Emotion, 25,* 183-192.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Van Heck, G. L. (1984). The construction of a general taxonomy of situations. In: H. Bonarius, G. L. Van Heck & N. Smid (Hrsg.), *Personality Psychology in Europe: Theoretical and Empirical Developments, Vol. 1* (S. 149-164). Lisse: Sweets & Zeitlinger.

Van Heck, G. L. (1989). Situation concepts: definition and classification. In: P. J. Hettema (Hrsg.), *Personality and Environment, Assessment of Human Adaptation* (pp. 241-259). Chichester: Wiley.

Watson, D. (1988). The vicissitudes of mood measurement: Effects of variying desprictors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology, 55,* 128-141.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, *54,* 1063-1070.

West, S. G., Finch, J. F. & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Hrsg.), *Structural equation modeling: Concepts, issues and application (*S. 56-75). Thousand Oaks, CA: Sage.

Wetter, E. K., & Hankin, B. L. (2009). Mediational pathways through which positive and negative emotionality contribute to anhedonic symptoms of depression: A prospective study of adolescents. *Journal of Abnormal Child Psychology*, *37*, 507–520.

Widiger, T. A. & Clark, L. A. (2000). Toward DSM-V and the classification of psychopathology. *Psychological Bulletin, 126,* 946-963.

Wishman, M. A., Perez, J. E. & Ramel, W. (2000). Factor Structure of the Beck Depression Inventory-Second Edition (BDI II) in a Student Sample. *Journal of Clinical Psychology, 56,* 545-551.

Wittchen, H.-U. & Jacobi, F. (2005). Size and burden of mental disorders in Europe: A critical review and appraisal of 27 studies. *European Neuropsychopharmacology, 15,* 357-376.

Wright, K. B. (2005). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services, *Journal of Computer-Mediated Communication, 10,* ohne Seitenangabe.

Yoon, J., Yoon, T. E. & George, J. F. (2011). Anticipating information needs for senior portal contents. *Computers in Human Behavior, 27,* 1012-1020.

Yousfi, S. & Steyer, R. (2006). Latent-State-Trait-Theorie. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 346-357). Göttingen: Hogrefe.

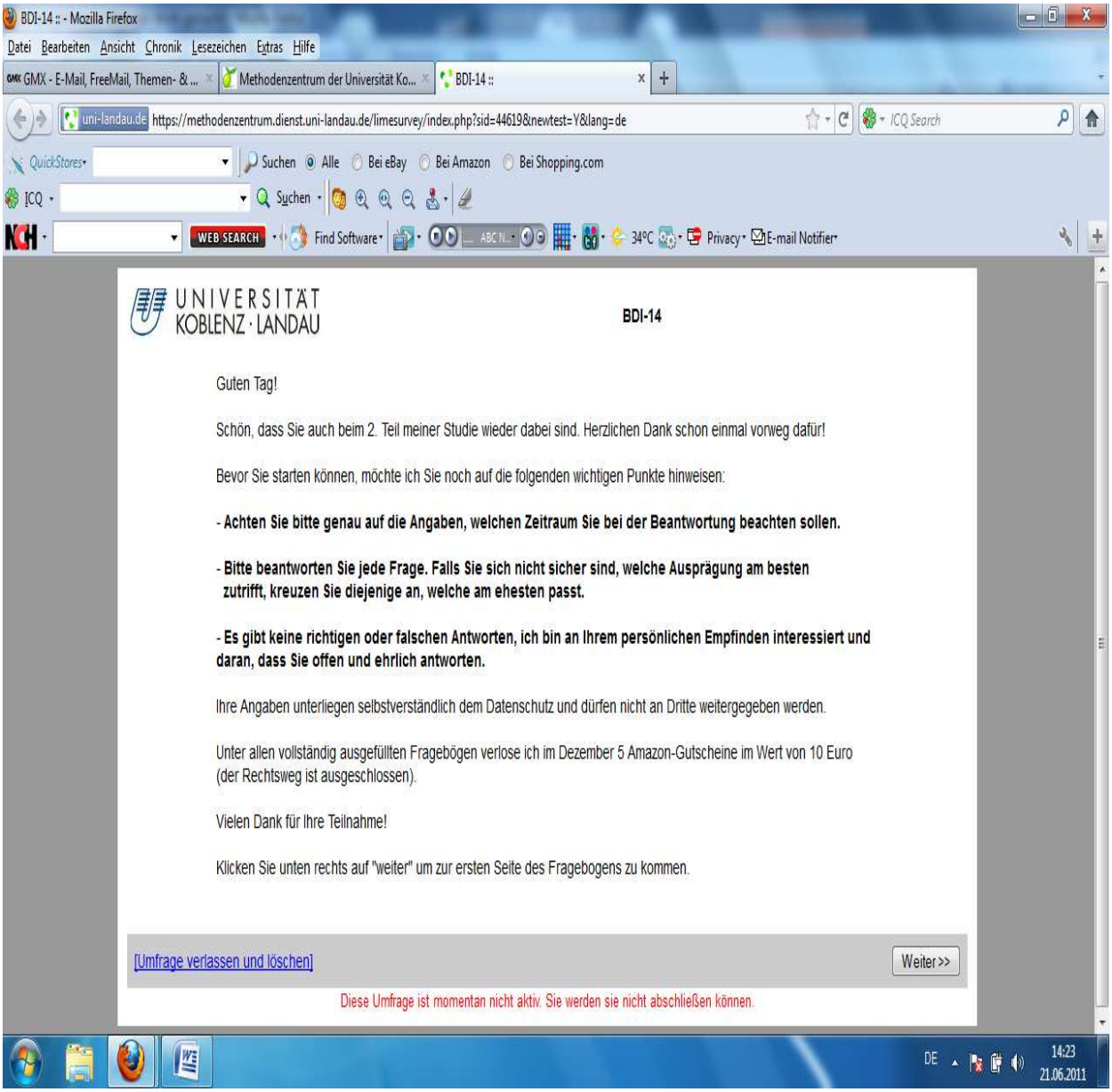
Zimmerman, M. (1986). The stability of the Revised Beck Depression Inventory in college student: Relationship with life events. *Cognitive Therapy and Research, 10,* 37-43.

Anhang

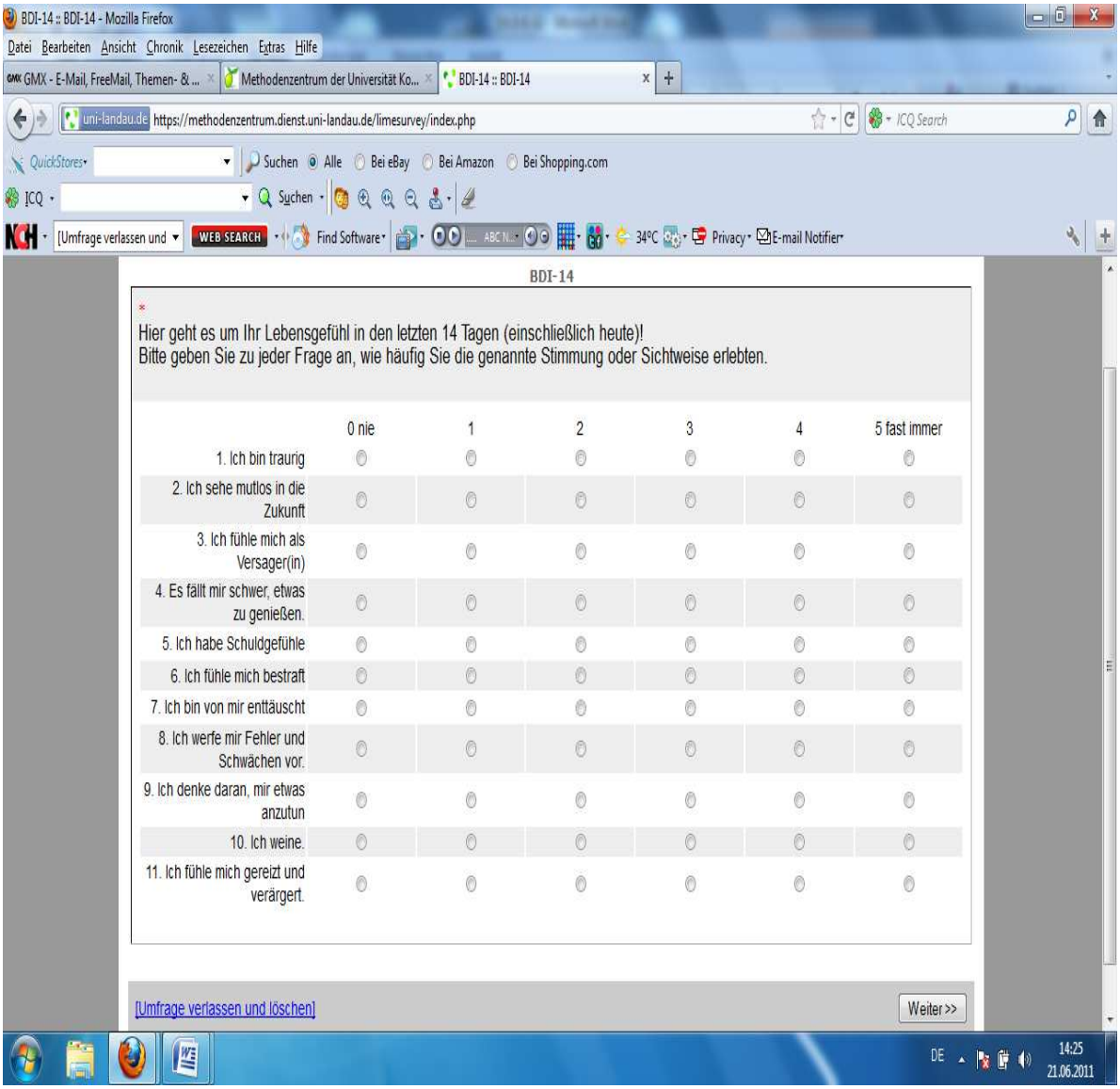
Anhangsverzeichnis

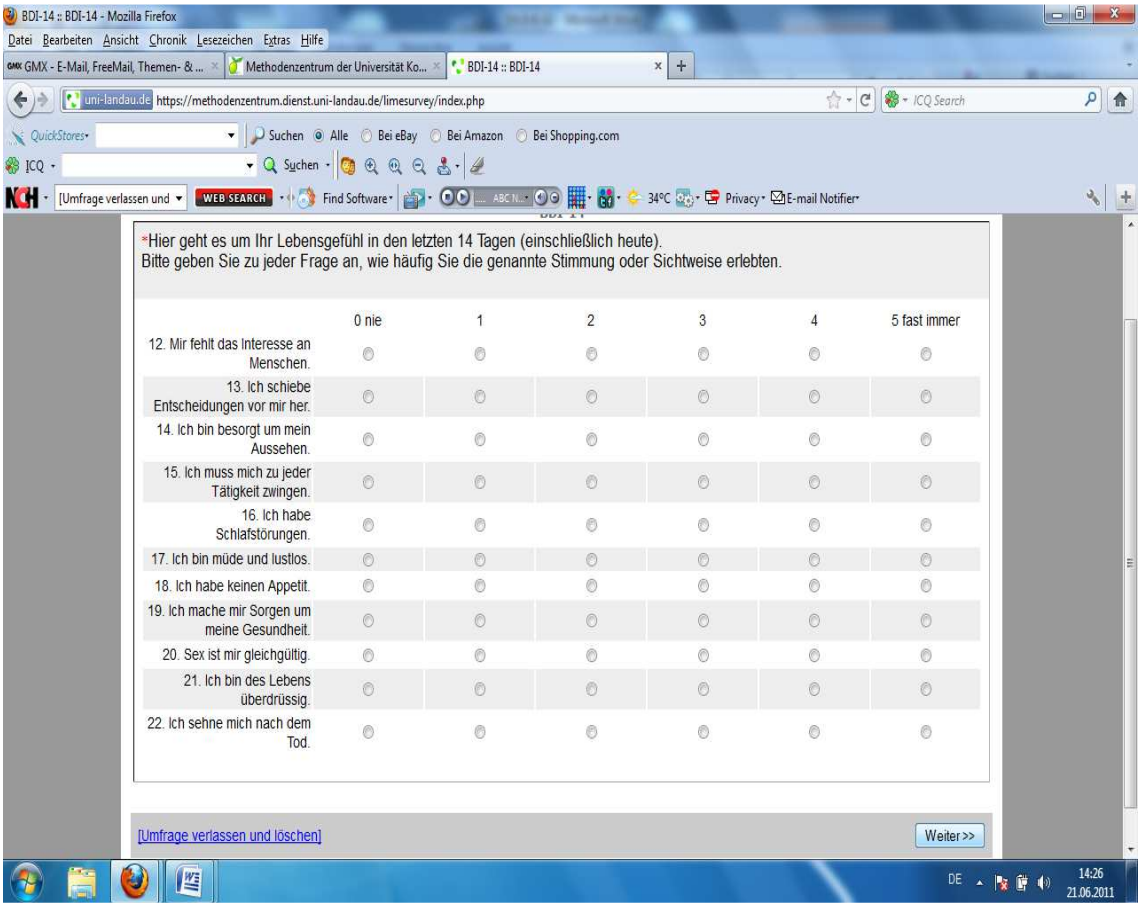
A Untersuchungsmaterial

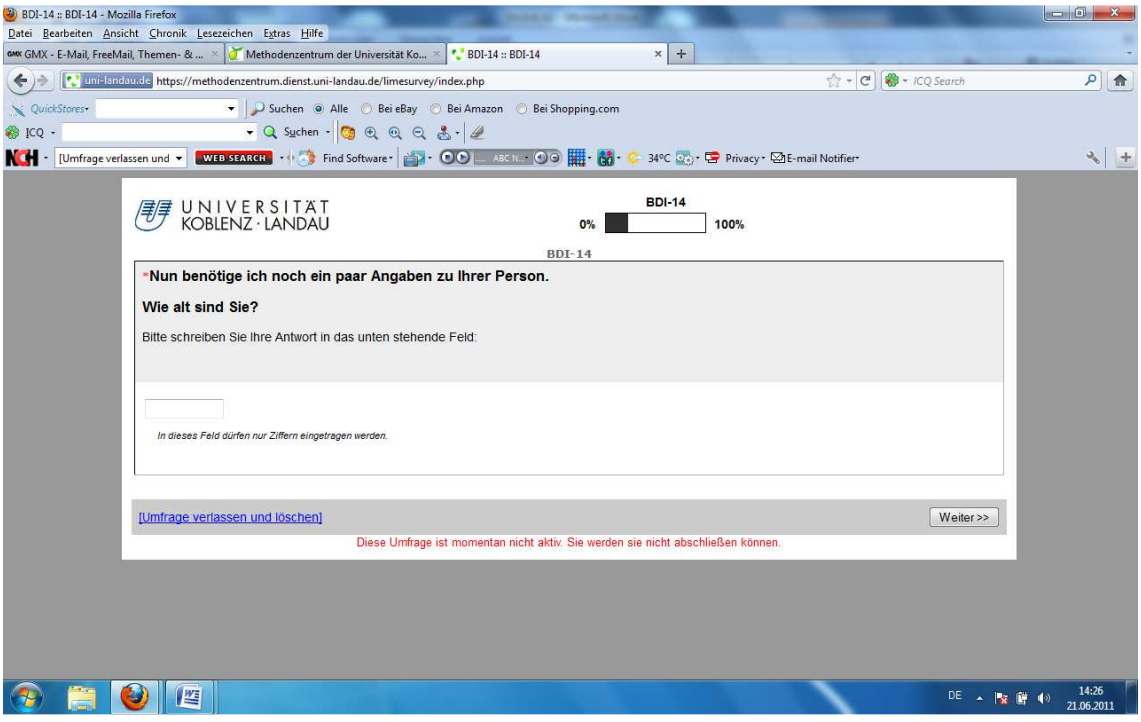
A.1 Online-Fragebogen

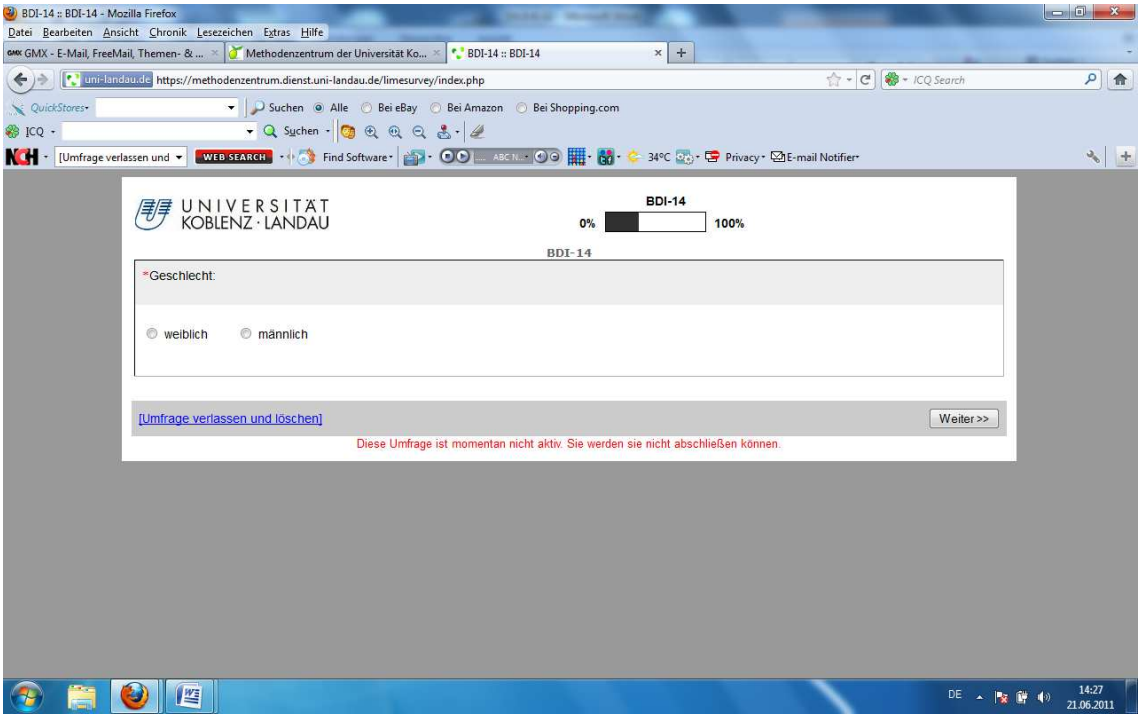


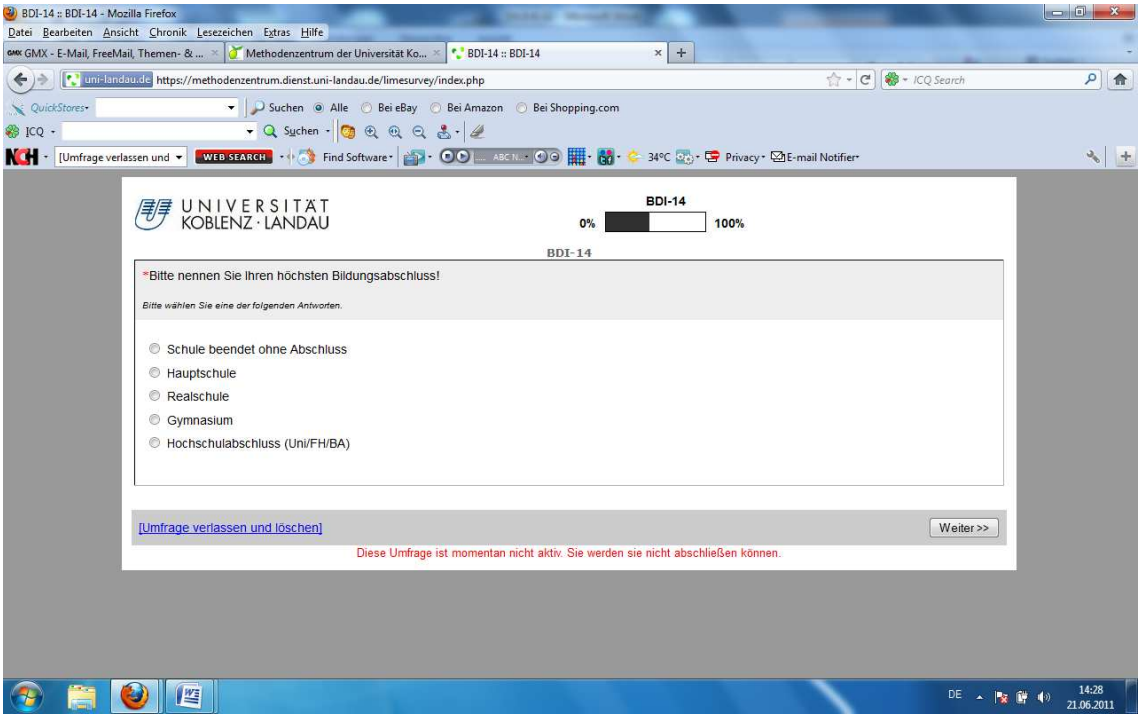
Anmerkung: Die Screenshots geben den zum zweiten Messzeitpunkt verwendeten Fragebogen wieder. Dieser unterscheidet sich von dem zum ersten Messzeitpunkt eingesetzten hinsichtlich Details der Begrüßung sowie hinsichtlich der Kontrollfrage, der Frage nach dem Gegenstand der Untersuchung sowie der Kommentarmöglichkeit. Der Screenshot wurde nach Deaktivierung der Umfrage erstellt. Sämtliche rot gedruckten Textteile und Zeichen waren nicht Teil des Originalfragebogens.

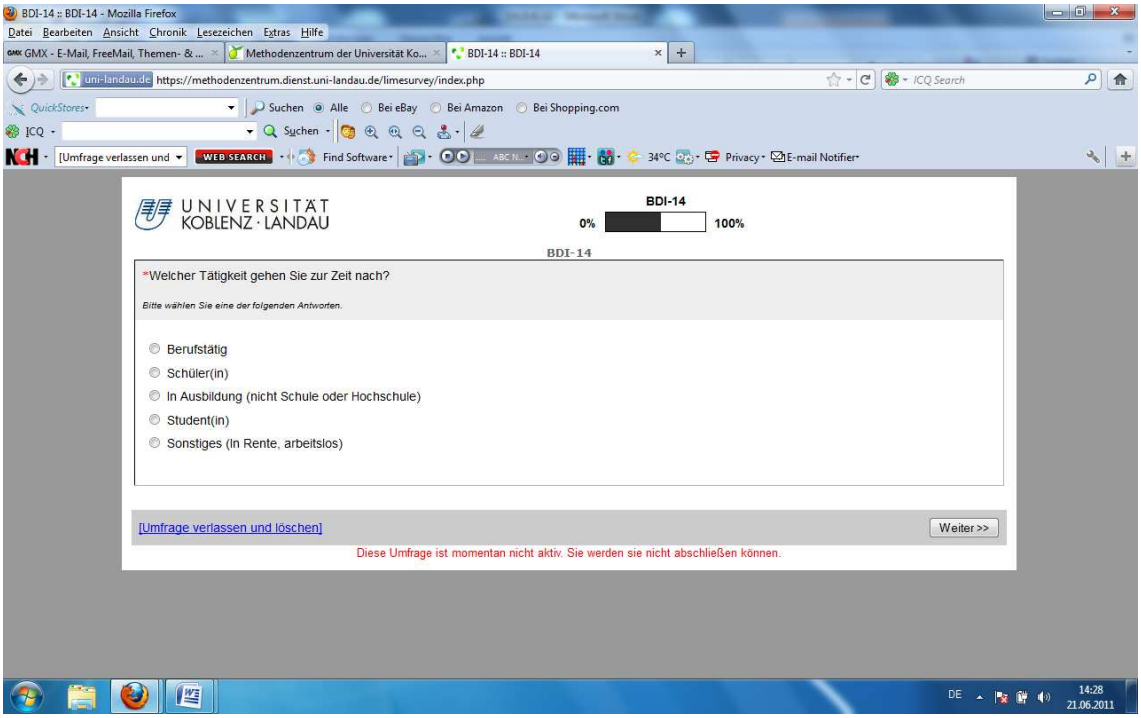

BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
GMX - E-Mail, FreeMail, Themen- & ...
Methodenzentrum der Universität Ko...
BDI-14 :: BDI-14
uni-landau.de https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
ICQ Search
Suchen Alle Bei eBay Bei Amazon Bei Shopping.com
Suchen
[Umfrage verlassen und
WEB SEARCH
Find Software
34°C
Privacy
E-mail Notifier
BDI-14
*
Hier geht es um Ihr Lebensgefühl in den letzten 14 Tagen (einschließlich heute)!
Bitte geben Sie zu jeder Frage an, wie häufig Sie die genannte Stimmung oder Sichtweise erlebten.
0 nie
1
2
3
4
5 fast immer
1. Ich bin traurig
2. Ich sehe mutlos in die Zukunft
3. Ich fühle mich als Versager(in)
4. Es fällt mir schwer, etwas zu genießen.
5. Ich habe Schuldgefühle
6. Ich fühle mich bestraft
7. Ich bin von mir enttäuscht
8. Ich werfe mir Fehler und Schwächen vor.
9. Ich denke daran, mir etwas anzutun
10. Ich weine.
11. Ich fühle mich gereizt und verärgert.
[Umfrage verlassen und löschen]
Weiter >>
DE
14:25
21.06.2011

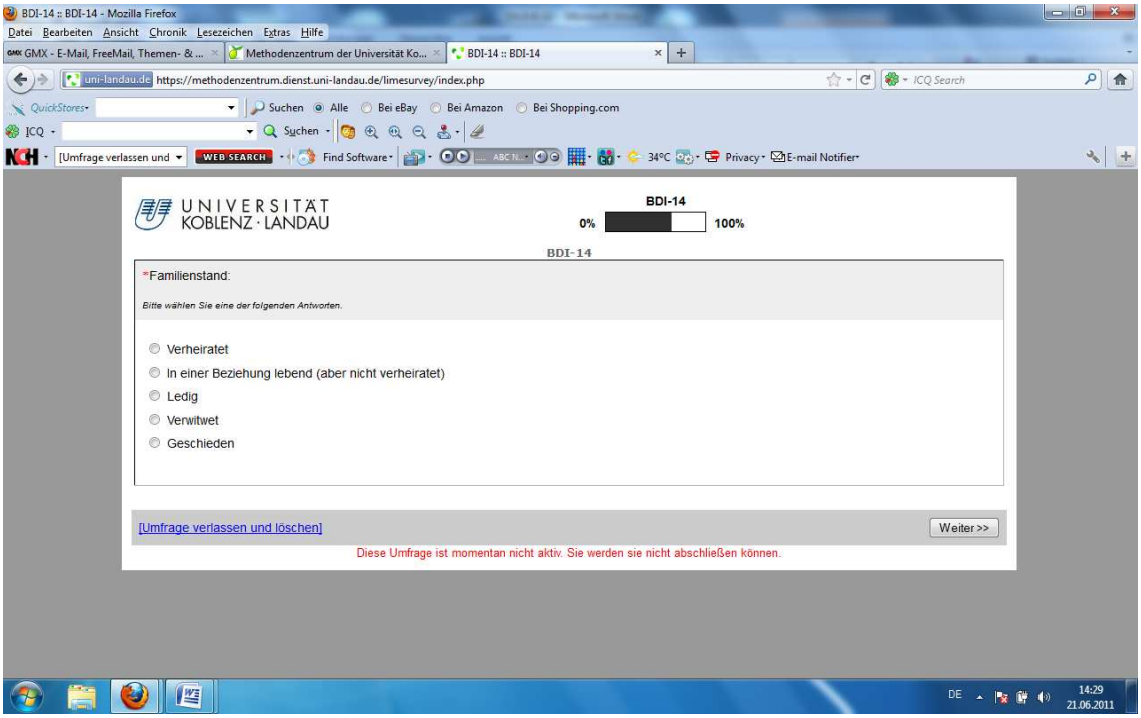

BDI-14 :: BDI-14 - Mozilla Firefox
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
*Hier geht es um Ihr Lebensgefühl in den letzten 14 Tagen (einschließlich heute).
Bitte geben Sie zu jeder Frage an, wie häufig Sie die genannte Stimmung oder Sichtweise erlebten.
0 nie
1
2
3
4
5 fast immer
12. Mir fehlt das Interesse an Menschen.
13. Ich schiebe Entscheidungen vor mir her.
14. Ich bin besorgt um mein Aussehen.
15. Ich muss mich zu jeder Tätigkeit zwingen.
16. Ich habe Schlafstörungen.
17. Ich bin müde und lustlos.
18. Ich habe keinen Appetit.
19. Ich mache mir Sorgen um meine Gesundheit.
20. Sex ist mir gleichgültig.
21. Ich bin des Lebens überdrüssig.
22. Ich sehne mich nach dem Tod.
[Umfrage verlassen und löschen]
Weiter >>
14:26
21.06.2011

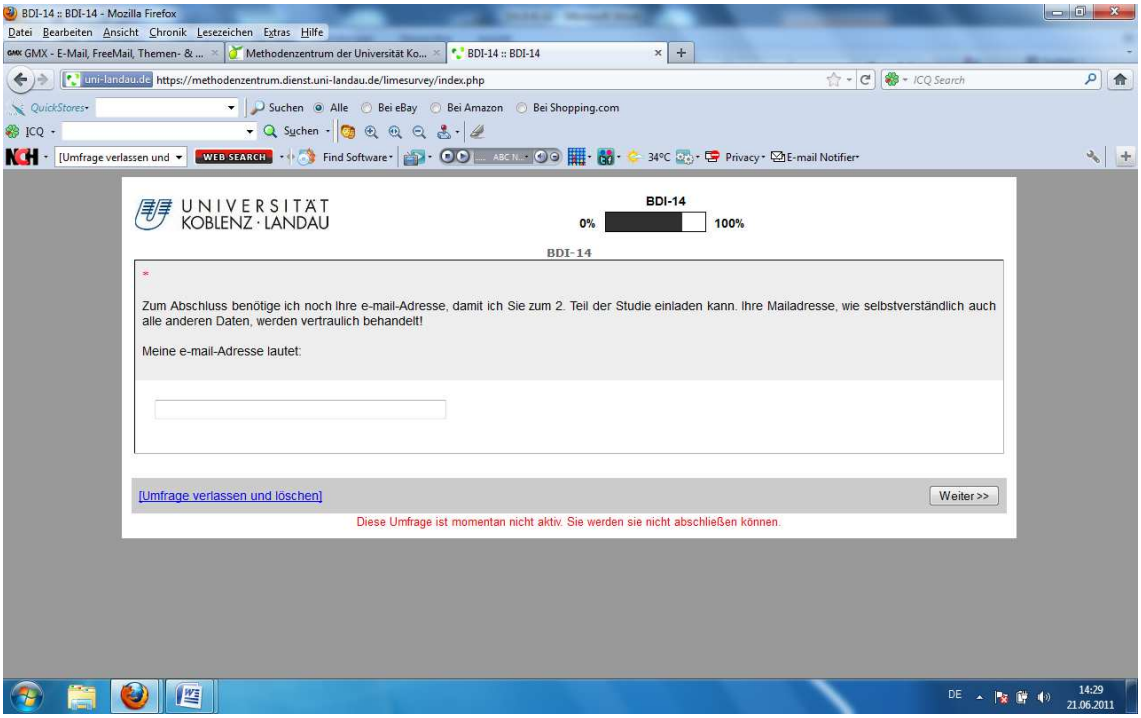

BDI-14 :: BDI-14 - Mozilla Firefox
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Nun benötige ich noch ein paar Angaben zu Ihrer Person.
Wie alt sind Sie?
Bitte schreiben Sie Ihre Antwort in das unten stehende Feld:
In dieses Feld dürfen nur Ziffern eingetragen werden.
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.
14:26
21.06.2011

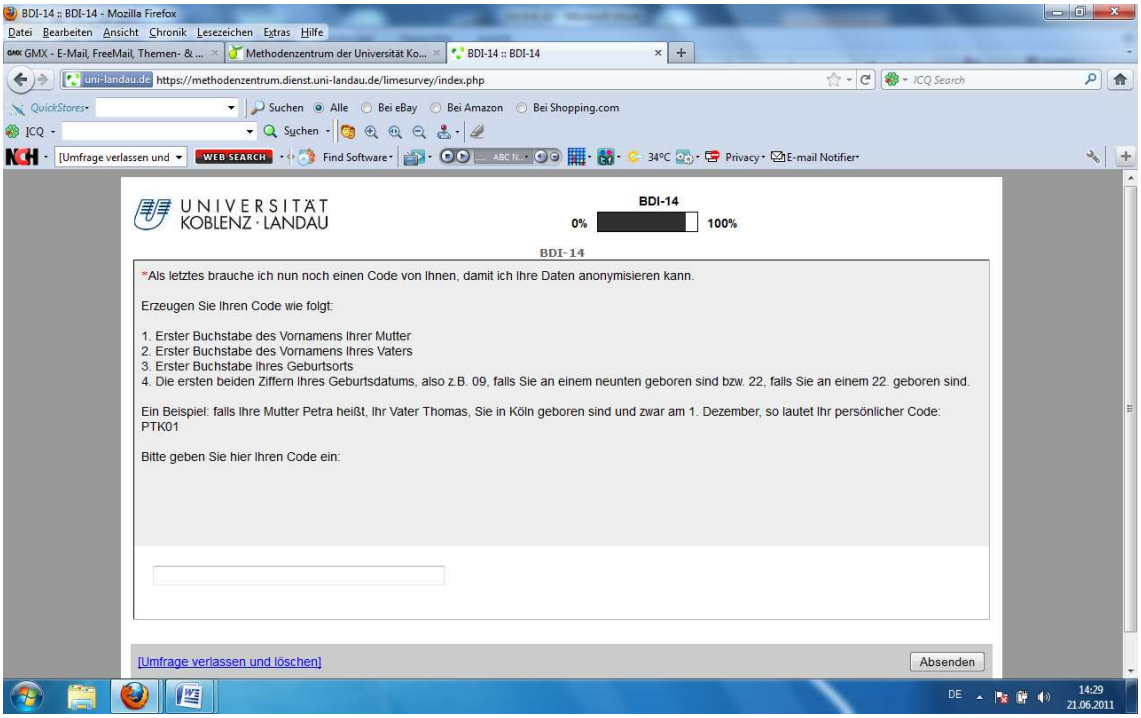

BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
GMX - E-Mail, FreeMail, Themen- & ...
Methodenzentrum der Universität Ko...
BDI-14 :: BDI-14
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Geschlecht:
weiblich
männlich
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.
14:27
21.06.2011

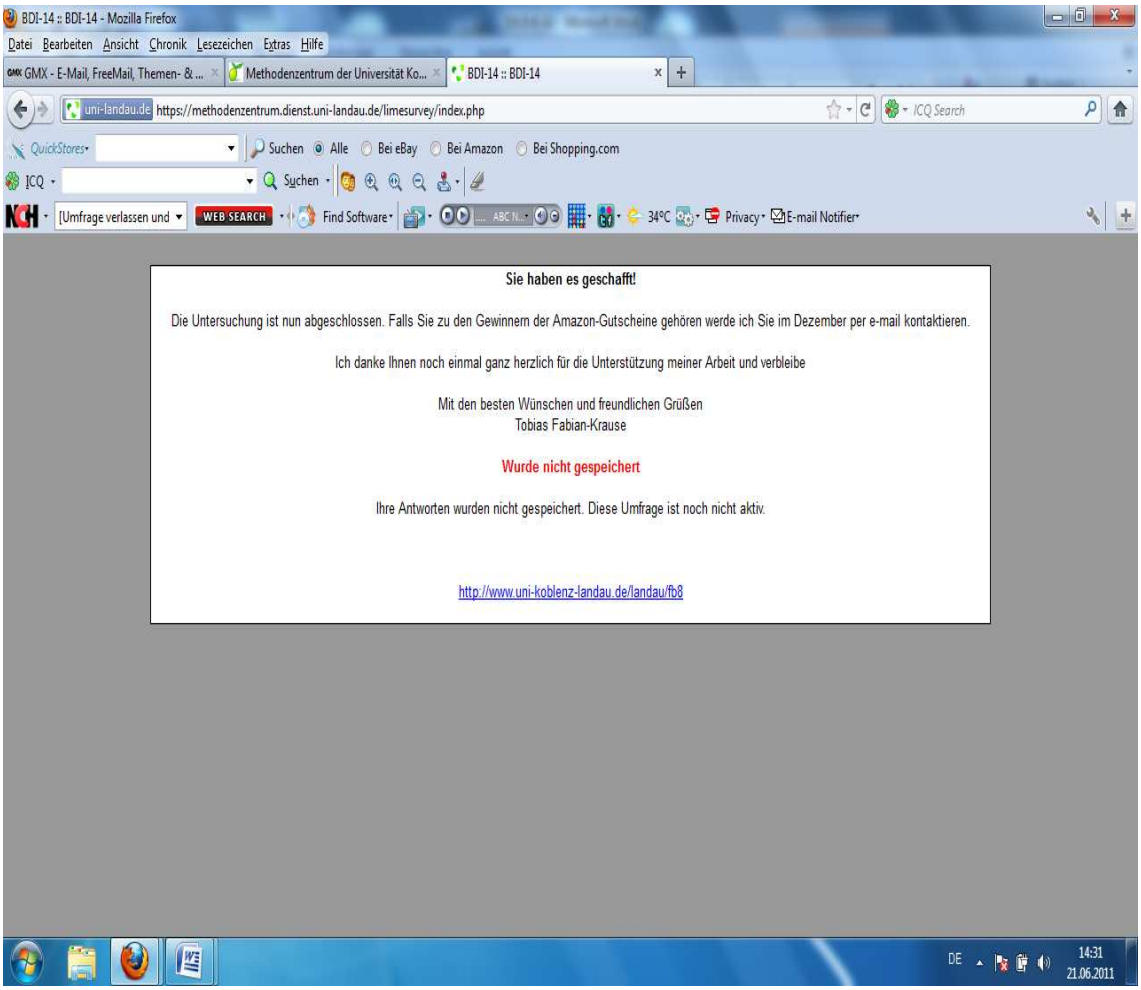

BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
GMX - E-Mail, FreeMail, Themen- & ...
Methodenzentrum der Universität Ko...
BDI-14 :: BDI-14
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Bitte nennen Sie Ihren höchsten Bildungsabschluss!
Bitte wählen Sie eine der folgenden Antworten.
Schule beendet ohne Abschluss
Hauptschule
Realschule
Gymnasium
Hochschulabschluss (Uni/FH/BA)
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.
14:28
21.06.2011


BDI-14 :: BDI-14 - Mozilla Firefox
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Welcher Tätigkeit gehen Sie zur Zeit nach?
Bitte wählen Sie eine der folgenden Antworten.
Berufstätig
Schüler(in)
In Ausbildung (nicht Schule oder Hochschule)
Student(in)
Sonstiges (In Rente, arbeitslos)
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.


BDI-14 :: BDI-14 - Mozilla Firefox
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Familienstand:
Bitte wählen Sie eine der folgenden Antworten.
Verheiratet
In einer Beziehung lebend (aber nicht verheiratet)
Ledig
Verwitwet
Geschieden
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.


BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*
Zum Abschluss benötige ich noch Ihre e-mail-Adresse, damit ich Sie zum 2. Teil der Studie einladen kann. Ihre Mailadresse, wie selbstverständlich auch alle anderen Daten, werden vertraulich behandelt!
Meine e-mail-Adresse lautet:
[Umfrage verlassen und löschen]
Weiter >>
Diese Umfrage ist momentan nicht aktiv. Sie werden sie nicht abschließen können.


BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
UNIVERSITÄT KOBLENZ · LANDAU
BDI-14
0%
100%
BDI-14
*Als letztes brauche ich nun noch einen Code von Ihnen, damit ich Ihre Daten anonymisieren kann.
Erzeugen Sie Ihren Code wie folgt:
1. Erster Buchstabe des Vornamens Ihrer Mutter
2. Erster Buchstabe des Vornamens Ihres Vaters
3. Erster Buchstabe Ihres Geburtsorts
4. Die ersten beiden Ziffern Ihres Geburtsdatums, also z.B. 09, falls Sie an einem neunten geboren sind bzw. 22, falls Sie an einem 22. geboren sind.
Ein Beispiel: falls Ihre Mutter Petra heißt, Ihr Vater Thomas, Sie in Köln geboren sind und zwar am 1. Dezember, so lautet Ihr persönlicher Code: PTK01
Bitte geben Sie hier Ihren Code ein:
[Umfrage verlassen und löschen]
Absenden


BDI-14 :: BDI-14 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
GMX - E-Mail, FreeMail, Themen- & ...
Methodenzentrum der Universität Ko...
BDI-14 :: BDI-14
uni-landau.de https://methodenzentrum.dienst.uni-landau.de/limesurvey/index.php
Suchen Alle Bei eBay Bei Amazon Bei Shopping.com
[Umfrage verlassen und
WEB SEARCH
Find Software
Privacy
E-mail Notifier
Sie haben es geschafft!
Die Untersuchung ist nun abgeschlossen. Falls Sie zu den Gewinnern der Amazon-Gutscheine gehören werde ich Sie im Dezember per e-mail kontaktieren.
Ich danke Ihnen noch einmal ganz herzlich für die Unterstützung meiner Arbeit und verbleibe
Mit den besten Wünschen und freundlichen Grüßen
Tobias Fabian-Krause
Wurde nicht gespeichert
Ihre Antworten wurden nicht gespeichert. Diese Umfrage ist noch nicht aktiv.
http://www.uni-koblenz-landau.de/landau/fb8
DE
14:31
21.06.2011

B Zusätzliche Ergebnisse

B.1 Tabelle B.1

Tabelle B.1: Dropout-Analyse *Chi*²-Tests

Variable		Anzahl	*df*	χ^2-Test [1]	*p* (2-seitig)
Geschlecht	Weiblich	257 (137)	1	2.73	.10
	Männlich	170 (118)			
Höchster Schulabschluss	Hauptschule	14 (16)	3	11.45	.01
	Mittlere Reife	34 (35)			
	Abitur	169 (102)			
	Hochschulabschluss	210 (102)			
Tätigkeit	Berufstätig	218 (135)	4	9.33	.049
	Schule	5 (6)			
	Ausbildung	7 (2)			
	Studium	157 (74)			
	Sonstiges	40 (38)			

Anmerkungen: N = 682; Bei Anzahl steht vor der Klammer der in der Analysestichprobe ermittelte Wert, in der Klammer der Wert der Dropoutgruppe;

[1]In den Tabellen B.1 und B.3 sind *Chi*²- und *p*-Werte bei Tätigkeit exakte Werte nach Fisher, da in beiden Fällen Zellen auftraten, die einen Erwartungswert unter 5 aufwiesen (Bortz, 2005, S. 170). Alle anderen Werte sind asymptotische Werte nach Pearson.

B.2 Tabelle B.2

Tabelle B.2: Dropoutanalyse ANOVA

Variable	$df_{Zähler}$	df_{Nenner}	*F-Wert*	*p*	η^2
Alter	1	680	1.83	.18	.00
BDI-V-Wert	1	681	0.53	.47	.00

Anmerkungen: N = 682; Eine Person in der Dropoutgruppe gab ihr Alter nicht an.

B.3 Tabelle B.3

Tabelle B.3: Vergleich der Instruktionsgruppen (Analysestichprobe) *Chi*²-Tests

Variable		Anzahl	*df*	χ^2-Test	*p* (2-seitig)
Geschlecht	Weiblich	99 (158)	1	7.29	.01
	Männlich	88 (82)			
Höchster Schul-abschluss	Hauptschule	8 (6)	3	7.09	.07
	Mittlere Reife	20 (14)			
	Abitur	63 (105)			
	Hochschul-abschluss	96 (115)			
Tätigkeit	Berufstätig	105 (113)	4	11.50	.016
	Schule	3 (2)			
	Ausbildung	5 (2)			
	Studium	53 (103)			
	Sonstiges	21 (20)			

Anmerkungen: N = 427; Bei Anzahl steht vor der Klammer der Wert der 3-Monats-gruppe, in der Klammer der Wert der 14-Tage-Gruppe

B.4 Tabelle B.4

Tabelle B.4: Vergleich der Instruktionsgruppen (Analysestichprobe) ANOVA

Variable	$df_{Zähler}$	df_{Nenner}	*F-Wert*	*p*	η^2
Alter	1	426	.05	.82	.00
BDI-V-Wert	1	426	.01	.92	.00

Anmerkungen: N = 427

B.5 Tabelle B.5

Tabelle B.5: Soziodemographische Eigenschaften der 3-Monatsgruppe

Geschlecht	Weiblich	Männlich			
	53%	47%			
Schulabschluss	Hauptschule	Mittlere Reife	Abitur	Hochschul-abschluss	
	4%	11%	34%	53%	
Familienstand	Verheiratet	Feste Beziehung	Keine Beziehung	Verwitwet	Geschieden
	33%	33%	29%	1%	4%
Tätigkeit	Berufstätig	Schule	Ausbildung	Studium	Sonstiges
	56%	2%	3%	28%	11%
Alter	Mittelwert	Minimum	Maximum		
	34,35	16	75		

Anmerkung: N = 187

B.6 Tabelle B.6

Tabelle B.6: Soziodemographische Eigenschaften der 14-Tagegruppe

Geschlecht	Weiblich	Männlich			
	66%	34%			
Schulabschluss	Hauptschule	Mittlere Reife	Abitur	Hochschul-abschluss	
	3%	6%	44%	48%	
Familienstand	Verheiratet	Feste Beziehung	Keine Beziehung	Verwitwet	Geschieden
	28%	39%	29%	0%	5%
Tätigkeit	Berufstätig	Schule	Ausbildung	Studium	Sonstiges
	48%	1%	1%	43%	8%
Alter	Mittelwert	Minimum	Maximum		
	34,35	16	75		

Anmerkung: N = 240

B.7 Tabelle B.7

Tabelle B.7: Mittelwerte, Varianzen und Trennschärfen der Items des BDI-V für weibliche Teilnehmer in der Gesamtstichprobe

	Item	*M*	*Var*	r_{it}
1.	Ich bin traurig. (2)	1.80	1.07	.62
2.	Ich sehe mutlos in die Zukunft. (2)	1.18	1.36	.70
3.	Ich fühle mich als Versager(in). (2)	.96	1.25	.70
4.	Es fällt mir schwer, etwas zu genießen. (2)	1.37	1.88	.58
5.	Ich habe Schuldgefühle. (1)	1.30	1.35	.55
6.	Ich fühle mich bestraft. (1)	.65	1.08	.57
7.	Ich bin von mir enttäuscht. (2)	1.33	1.37	.68
8.	Ich werfe mir Fehler und Schwächen vor. (1)	1.73	1.59	.67
9.	Ich denke daran, mir etwas anzutun. (1)	.32	.64	.53
10.	Ich weine. (1)	1.33	1.36	.45
11.	Ich fühle mich gereizt und verärgert. (2)	1.92	1.31	.52
12.	Mir fehlt das Interesse an Menschen. (2)	.83	1.24	.56
13.	Ich schiebe Entscheidungen vor mir her. (1)	2.04	1.87	.49
14.	Ich bin besorgt um mein Aussehen. (1)	1.95	1.62	.36
15.	Ich muss mich zu jeder Tätigkeit zwingen. (1)	1.26	1.49	.62
16.	Ich habe Schlafstörungen. (2)	1.39	2.05	.49
17.	Ich bin müde und lustlos. (1)	1.69	1.83	.73
18.	Ich habe keinen Appetit. (2)	.75	1.26	.40
19.	Ich mache mir Sorgen um meine Gesundheit. (1)	1.34	1.51	.39
20.	Sex ist mir gleichgültig. (2)	1.42	2.09	.43

Anmerkungen: N = 257, die eingeklammerte Zahl hinter den Items gibt an, zu welcher Testhälfte das Item gehört.

B.8 Tabelle B.8

Tabelle B.8: Mittelwerte, Varianzen und Trennschärfen der Items des BDI-V für männliche Teilnehmer in der Gesamtstichprobe

	Item	*M*	*Var*	r_{it}
1.	Ich bin traurig. (2)	1.43	1.13	.63
2.	Ich sehe mutlos in die Zukunft. (2)	.92	1.15	.62
3.	Ich fühle mich als Versager(in). (2)	.85	1.14	.68
4.	Es fällt mir schwer, etwas zu genießen. (2)	1.28	1.34	.62
5.	Ich habe Schuldgefühle. (1)	1.00	1.14	.56
6.	Ich fühle mich bestraft. (1)	.59	1.05	.47
7.	Ich bin von mir enttäuscht. (2)	1.03	1.04	.69
8.	Ich werfe mir Fehler und Schwächen vor. (1)	1.53	1.14	.69
9.	Ich denke daran, mir etwas anzutun. (1)	.25	.72	.47
10.	Ich weine. (1)	.49	.68	.45
11.	Ich fühle mich gereizt und verärgert. (2)	1.68	1.11	.52
12.	Mir fehlt das Interesse an Menschen. (2)	1.11	1.10	.56
13.	Ich schiebe Entscheidungen vor mir her. (1)	2.00	1.30	.61
14.	Ich bin besorgt um mein Aussehen. (1)	1.31	1.24	.34
15.	Ich muss mich zu jeder Tätigkeit zwingen. (1)	1.26	1.10	.68
16.	Ich habe Schlafstörungen. (2)	1.12	1.25	.36
17.	Ich bin müde und lustlos. (1)	1.37	1.16	.69
18.	Ich habe keinen Appetit. (2)	.52	.82	.40
19.	Ich mache mir Sorgen um meine Gesundheit. (1)	1.25	1.09	.40
20.	Sex ist mir gleichgültig. (2)	.94	1.17	.42

Anmerkungen: $N = 170$, die eingeklammerte Zahl hinter den Items gibt an, zu welcher Testhälfte das Item gehört.

B.9 Tabelle B.9

Tabelle B.9: Mittelwerte, Varianzen und Trennschärfen der Items des BDI-V für weibliche Teilnehmer in der Kontrollfragenstichprobe (erster Messzeitpunkt)

	Item	*M*	*Var*
1.	Ich bin traurig. (2)	1.80	1.22
2.	Ich sehe mutlos in die Zukunft. (2)	1.21	1.37
3.	Ich fühle mich als Versager(in). (2)	.96	1.24
4.	Es fällt mir schwer, etwas zu genießen. (2)	1.41	1.84
5.	Ich habe Schuldgefühle. (1)	1.24	1.19
6.	Ich fühle mich bestraft. (1)	.64	1.19
7.	Ich bin von mir enttäuscht. (2)	1.36	1.32
8.	Ich werfe mir Fehler und Schwächen vor. (1)	1.80	1.70
9.	Ich denke daran, mir etwas anzutun. (1)	.28	.59
10.	Ich weine. (1)	1.31	1.48
11.	Ich fühle mich gereizt und verärgert. (2)	2.02	1.44
12.	Mir fehlt das Interesse an Menschen. (2)	.83	1.34
13.	Ich schiebe Entscheidungen vor mir her. (1)	2.03	1.88
14.	Ich bin besorgt um mein Aussehen. (1)	1.94	1.63
15.	Ich muss mich zu jeder Tätigkeit zwingen. (1)	1.37	1.58
16.	Ich habe Schlafstörungen. (2)	1.32	1.94
17.	Ich bin müde und lustlos. (1)	1.69	1.74
18.	Ich habe keinen Appetit. (2)	.72	1.33
19.	Ich mache mir Sorgen um meine Gesundheit. (1)	1.30	1.51
20.	Sex ist mir gleichgültig. (2)	1.32	1.79

Anmerkungen: N = 166, die eingeklammerte Zahl hinter den Items gibt an, zu welcher Testhälfte das Item gehört.

B.10 Tabelle B.10

Tabelle B.10: Mittelwerte, Varianzen und Trennschärfen der Items des BDI-V für männliche Teilnehmer in der Kontrollfragenstichprobe (erster Messzeitpunkt)

	Item	*M*	*Var*
1.	Ich bin traurig. (2)	1.34	.91
2.	Ich sehe mutlos in die Zukunft. (2)	.83	.93
3.	Ich fühle mich als Versager(in). (2)	.69	.70
4.	Es fällt mir schwer, etwas zu genießen. (2)	1.16	1.43
5.	Ich habe Schuldgefühle. (1)	.96	1.01
6.	Ich fühle mich bestraft. (1)	.45	.65
7.	Ich bin von mir enttäuscht. (2)	.96	.75
8.	Ich werfe mir Fehler und Schwächen vor. (1)	1.51	1.08
9.	Ich denke daran, mir etwas anzutun. (1)	.15	.33
10.	Ich weine. (1)	.36	.31
11.	Ich fühle mich gereizt und verärgert. (2)	1.56	.88
12.	Mir fehlt das Interesse an Menschen. (2)	1.00	.87
13.	Ich schiebe Entscheidungen vor mir her. (1)	1.91	1.28
14.	Ich bin besorgt um mein Aussehen. (1)	1.42	1.56
15.	Ich muss mich zu jeder Tätigkeit zwingen. (1)	1.13	.92
16.	Ich habe Schlafstörungen. (2)	1.08	1.53
17.	Ich bin müde und lustlos. (1)	1.28	1.15
18.	Ich habe keinen Appetit. (2)	.49	.66
19.	Ich mache mir Sorgen um meine Gesundheit. (1)	1.19	1.04
20.	Sex ist mir gleichgültig. (2)	.91	1.23

Anmerkungen: $N = 100$, die eingeklammerte Zahl hinter den Items gibt an, zu welcher Testhälfte das Item gehört.

Erklärung

Hiermit versichere ich gemäß § 18 Abs. 8 der Diplomprüfungsordnung Psychologie der Universität Koblenz-Landau, Campus Landau, in der Fassung vom 18.02.1993, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsausschuss vorgelegen.

Landau, Juli 2011