universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

„Backtracking: Wie viel QUALI steckt in QUANTI?

Evaluierung von Messeigenschaften ausgewählter Items der Human Values nach Shalom Schwartz in der Fassung des European Social Surveys mittels Mixed-Methods Designs“

Verfasser

Georg Kessler

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften

(Mag. rer. soc. oec)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A121
Studienrichtung lt. Studienblatt:	Soziologie
Betreuer:	Univ. Prof. em. Dr. Peter Schmidt

Inhaltsverzeichnis

1 Werte? Auf jeden Fall! ... 1
1.1 Problemstellung – Zur Lücke zwischen Theorie und Praxis in der Erfassung von menschlichen Wertevorstellungen ... 2
1.2 Aufbau der Arbeit ... 4
2 Theorie ... 6
2.1 Werte in den Sozialwissenschaften – Ein Überblick ... 6
2.2 Theorie der human values nach Shalom Schwartz ... 10
2.3 Kritische Diskussion – Stand der wissenschaftlichen Forschung ... 22
2.3.1 Deskriptive Erfassung der Wertestruktur ... 22
2.3.2 Inferenzstatistische Tests ... 26
2.3.3 Davidov-Studien ... 30
2.3.4 Kultur- und Kontextabhängigkeit der Messung ... 33
2.3.5 Messqualität des PVQ ... 38
2.4 Zusammenfassung der Diskussion ... 43
3 Methode ... 45
3.1 Methodologische Probleme ... 45
3.1.1 Der Zugriff auf Werte – values as truisms ... 45
3.1.2 Skalierung von Werten ... 48
3.1.3 Doppelte Stimuli (double-barreledness) ... 50
3.1.4 Werte: normative Ideale oder persönliche Prioritäten ... 51
3.1.5 Messung von Kontext? ... 52
3.1.6 Soziale Erwünschtheit ... 53
3.1.7 Niedrige Reliabilitäten ... 59
3.2 Methodenmix ... 60
3.2.1 Warum überhaupt MM? ... 60
3.2.2 Wie lässt sich das Design klassifizieren? ... 62
3.3 Konfirmatorische Faktorenanalyse (CFA) ... 64
3.4 Kognitive Interviews ... 69
4 Empirie ... 75
4.1 Welche Version? Warum diese? ... 76
4.2 Beschreibung der Daten ... 76
4.3 Exploratorische Voruntersuchungen ... 78
4.3.1 Exploratorische Faktorenanalyse (EFA) ... 78
4.3.2 Multidimensionale Skalierung (MDS) ... 82

4.3.3 Missing-Analyse ... 84
4.4 Konvergierende Methodentriangulation ... 85
4.4.1 Beurteilung der Messmodelle ... 88
4.4.2 Vorläufige Synopsis ... 93
4.4.3 Paarweise Prüfung der Überlagerungen der Messung ... 94
4.4.4 Reliabilitäten ... 107
4.4.5 Bewertung der Messeigenschaft ausgewählter Items unter Zuhilfenahme kognitiver Interviews ... 109
4.5 Fazit ... 127
5 Schlussbetrachtung ... 133
5.1 Bewertung der Methode ... 133
5.2 Probleme und Grenzen der Arbeit ... 135
6 Literaturverzeichnis ... 137
7 Anhang ... 147
Tabelle 18: Labels, Wording und Wertetypen der ESS-PVQ-Items ... 148
Tabelle 19: Antwortkategorien für die ESS-PVQ-Items ... 148
Tabelle 20: Korrelaionsmatrix der 21 ESS-PVQ Items ... 149
Tabelle 21: Kommunalitäten (EFA, Extraktion: Hauptachsenanalyse; SPSS) ... 150
Tabelle 22: Deskriptive Statistiken der 21 ESS-PVQ-Items ... 151
Tabelle 23: Ladungsmatrix EFA (Rotation: geomin; Mplus) ... 152
Tabelle 24: Häufigkeiten und Mittelwerte fehlender Werte der 21 PVQ-Items ... 154
Tabelle 25: Ladungsmatrix EFA (Haupachsenanalyse; Rotation: oblique; SPSS) ... 155
Tabelle 26: Modifikationen für Kreuzladungen (7-Faktorenmodell, Mplus) ... 156
Tabelle 27: Modifikationen für Fehlerkovarianzen (7-Faktorenmodell, Mplus) ... 157
Abstract – Deutsche Fassung ... 158
Abstract – English Version ... 159
Curriculum Vitae ... 160

Tabellenverzeichnis

Tabelle 1: Unterschied Eigenschaften – Werte (aus: Roccas et al., 2002) 10
Tabelle 2: Antizipierte Effekte von Kovariaten auf zehn Wertetypen (aus: Meuleman et al. 2012) 36
Tabelle 3: Tatsächliche Effekte von Kovariaten auf Wertetypen (aus: Meuleman 2012) 37
Tabelle 4: Ergebnisse des paarweisen Vergleichs (aus: Knoppen und Saris 2009b) 42
Tabelle 5: „Avoiding pitfalls in asking probes“ (aus Willis, 2005: 141) 72
Tabelle 6: Quotenplan kognitive Interviews 73
Tabelle 7: EFA mittels Mplus (geomin-rotiert [oblique 0,5]) 80
Tabelle 8: Faktorkorrelationsmatrix Mplus – SPSS 82
Tabelle 9: Überschneidung in drei Stichproben (AT, CH und DE)
von Items mit überdurchschnittlich vielen fehlenden Werten 85
Tabelle 10: Vorläufige Einteilung relativ guter und mangelhafter Indikatoren 93
Tabelle 11: Paarweiser Vergleich benachbarter higher-order factors 94
Tabelle 12: Paarweiser Vergleich gegenüberliegender higher-order factors 94
Tabelle 13: Faktorkorrelationen der higher-order values 95
Tabelle 14: Faktorladungen und Reliabilitäten zu einem Modell mit vier und sieben Faktoren 108
Tabelle 15: Oberstes und unterstes Quartil gemessen an Reliabilität der Items 108
Tabelle 16: Übersicht Problemdimensionen für alle 21 Items 130
Tabelle 17: Problemkategorien für ausgewählte Items aus kognitiven Protokollen 131
Tabelle 18: Labels, Wording und Wertetypen der ESS-PVQ-Items 148
Tabelle 19: Antwortkategorien für die ESS-PVQ-Items 148
Tabelle 20: Korrelaionsmatrix der 21 ESS-PVQ Items 149
Tabelle 21: Kommunalitäten (EFA, Extraktion: Hauptachsenanalyse; SPSS) 150
Tabelle 22: Deskriptive Statistiken der 21 ESS-PVQ-Items 151
Tabelle 23: Ladungsmatrix EFA (Rotation: geomin; Mplus) 152
Tabelle 24: Häufigkeiten und Mittelwerte fehlender Werte der 21 PVQ-Items 154
Tabelle 25: Ladungsmatrix EFA (Haupachsenanalyse; Rotation: oblique; SPSS) 155
abelle 26: Modifikationen für Kreuzladungen (7-Faktorenmodell, Mplus) 156
Tabelle 27: Modifikationen für Fehlerkovarianzen (7-Faktorenmodell, Mplus) 157

Abbildungsverzeichnis

Abbildung 1: Zirkumplex mit Wertetypen und Wertetypen höherer Ordnung.....14
Abbildung 2: Doppelter Stimulus (aus: Schwartz, 2003).....51
Abbildung 3: SDR nach Paulhus (aus: Paulhus, 2002).....56
Abbildung 4: Logik des Backtracking.....63
Abbildung 5: EFA geplottet 80
Abbildung 6: Äquivalente Repräsentationen (aus: Knoppen und Saris, 2009b).....87
Abbildung 7: MDS mit unzentrierten Daten.....153
Abbildung 8: MDS mit zentrierten Daten.....153

Abkürzungsverzeichnis

CFA	Konfirmatorische Faktorenanalyse
EFA	Exploratorische Faktorenalayse
ESS	European Social Survey
MDS	Multidimensionale Skalierung
MM	Mixed-Methods
PVQ	Portrait Values Questionnaire
QUALI	qualitative Methoden
QUANTI	quantitative Methoden
SDR	Social Desirable Responding
SVS	Schwartz Value Survey

1 Werte? Auf jeden Fall!

Sehr schnell erkennt ein Forscher, der das Gebiet der menschlichen Werte zum ersten Mal betritt, dass zwischen dem, was man in der *scientific community* unter dem Begriff versteht, und dem, wie der Begriff alltäglich gebraucht wird, nicht unbedingt Konsens herrscht. Es kann durchaus passieren, dass man verwunderte Blicke erntet, wenn man unter Bekannten auf Unterschiede hinweist und Nuancen anspricht. In Gesprächen mit Menschen, die nicht in die wissenschaftliche Debatte rund um Werte vertieft sind, habe ich die Erfahrung gemacht, dass sie „Werte“ immer im Lichte einer laufenden Debatte sehen. Oft geht es in diesem Verständnis gleich um eine Bewertung im Sinne von „gut“ oder „schlecht“. In der Alltagssprache vermitteln Werte eine solide Basis in der Gesellschaft. Sie sind die Säulen, auf denen ein gedeihliches Zusammenleben fußt. In der Politik sind sie daher auch ein besonders beliebtes Thema: Wahlkämpfe münden oft und gerne in Wertedebatten. Schließlich betreffen Werte uns alle, aber niemand kann genau sagen, woher sie kommen und welche Auswirkungen sie tatsächlich auf unser Leben haben. Damit eignen sie sich hervorragend, um die Wähler emotional anzusprechen, bleiben aber trotzdem vage genug, um sich nicht positionieren zu müssen. So geschehen in der Bundespräsidentenwahl 2010, in der die Kandidaten für „Werte“ einstanden: Heinz Fischer, SPÖ, meinte auf seinen Plakaten: „Unser Handeln braucht Werte“, Barbara Rosenkranz, FPÖ, verlautbarte auf ihren: „Ohne Mut keine Werte“. Rudolf Gehring, bürgerlicher Parteiloser, wies hingegen darauf hin, dass er Werte nicht plakatiere, sie aber dafür lebe. Mit Werten lassen sich Wahlen gewinnen. Vielleicht auch zum Teil deshalb, weil Werte ein Versprechen darstellen, das in unserer entgrenzten, postmodernen, sich ständig wandelnden Welt Halt zu geben scheint.

Die akademische Auseinandersetzung mit dem Thema ist da weitaus trockener und betrachtet das Thema differenzierter und detaillierter; somit entfernt sich dieser Diskurs vom Alltagsverständnis. Latente Konstrukte werden zerlegt – analysiert – und immer weiter heruntergebrochen, bis nur noch wenig von der ursprünglichen Qualität vorhanden ist. Sich mit der Messqualität von Werten zu befassen, wie in der vorliegenden Arbeit, ist noch ein Schritt weiter Richtung Abstraktion. Zu Beginn einer solchen Arbeit stellt sich daher auch oft die Frage nach dem Nutzen und der Möglichkeit zur Anwendung oder der Anknüpfung an andere Lebenswelten. Ob meine Arbeit einen sinnvollen Beitrag zur wissenschaftlichen Diskussion bringt? Mag sein. Sicher ist, dass sich damit keine Wahlen gewinnen lassen.

Bei all der Abstraktion beschäftigten mich dennoch, während die Arbeit entstand, Gedanken, die einem Spurensucher, Detektiv oder Kriminalisten nicht unbekannt sein dürften: Wie lassen sich die vorgefundenen Zeichen deuten? Wie lassen sich Vorstellungen und Wirklichkeit verbinden? Wie lässt sich aus den Spuren, die ein Ereignis hinterließ, auf das Ereignis selbst schließen.

Mein Tatort sind Zahlen, und meine Fährten sind Aussagen von Leuten, die das Ereignis selber miterlebten. Wie ein Kriminalist versuche ich den Hergang der Tat sinnvoll zu rekonstruieren, indem ich von Tatort zu Tatort gehe und mir von den Zeugen beschreiben lasse, was vorgefallen ist. Das umfasst *backtracking*.

1.1 Problemstellung – Zur Lücke zwischen Theorie und Praxis in der Erfassung von menschlichen Wertevorstellungen

Werte sind für die Soziologie seit der Zeit Max Webers im 19. Jahrhundert bedeutsam. Seit damals besteht die Vorstellung, dass sie Handlungen bestimmen und damit soziales Verhalten erklären. Berühmt wurde Webers Theorie, dass die kapitalistische Produktionsweise durch eine religiöse Werthaltung – das protestantische Ethos – geprägt ist. Die Messung von Werten fand allerdings erst in den 1970er Jahren in Hofstedes Werteforschung in einem international tätigen Konzern und den Skalen von Inglehart und Rokeach ihren Höhepunkt, um anschließend abflachendes Interesse zu erfahren: Die Messung konnte mit den Erwartungen, die an sie gestellt wurden, nicht mithalten. 1992 belebte Schwartz die Diskussion mit seiner Version der *human values* neu. Im Gegensatz zu seinen Vorgängern misst er Werte in einer dynamischen Struktur, die Konflikte und Übereinstimmungen zwischen zehn universellen Werten vorhersagt. Das schafft die Möglichkeit, sowohl die dahinter stehende Theorie als auch Zusammenhänge mit anderen theoretischen Konstrukten empirisch zu testen. Die Grundlage dafür liefern Erhebungen der international vergleichenden Forschung, darunter auch der European Social Survey (kurz: ESS), dessen Daten die Basis für die vorliegende Arbeit sind. Der ESS erhebt eine verkürzte Version der Skala der *human values* von Schwartz alle zwei Jahre als Basismodul und macht sie damit einer großen Bandbreite an Forschungszwecken zugänglich. Darunter fallen Trendstudien zur Werteeinstellung innerhalb bestimmter Bevölkerungsgruppen, der Vergleich von Populationen in ihrer Wertestruktur und die Verwendung der Werte als erklärende Variablen in kausalen Beziehungen mit anderen theoretischen Konstrukten.

Damit solche Schlüsse auch richtig und verlässlich sind, muss eine Skala Qualitätskriterien erfüllen. Die wichtigsten sind ihre Validität und Reliabilität. Sie sagen aus, ob die gewünschten Überlegungen erfasst und wie treffsicher sie gemessen werden. Sind sie mangelhaft, so sind es auch die Schlüsse, die aus der Skala gezogen werden. Deshalb entspann sich um die Wertemessung von Schwartz eine lebendige Debatte, die von der ursprünglichen Version aus dem Jahr 1992 bis zu jener reicht, die in den ESS aufgenommen wurde. Dabei kristallisierte sich mit Fokus auf die im ESS verwendete Version heraus, dass sie mit gewissen Mängeln behaftet ist. Davidov z.B. stellte in seinen Studien (Davidov, 2008; 2010; Davidov et al., 2008) zur Vergleichbarkeit der Messung fest, dass die Skala nur in einer Auswahl an Ländern, die am ESS teilnehmen, konstant dasselbe misst. Außerdem ist die Messung anstatt mit den vorhergesagten zehn Werten nur mit sechs bis maximal acht – je nach Land, das gerade untersucht wurde – vergleichbar. Für nachfolgende Analysen mussten Werte, die sich nicht scharf genug voneinander trennen ließen, zusammengefasst werden, wodurch ein neuer „Wert“ entstand. Forscher auf dem Gebiet kritisierten, dass dadurch nicht mehr klar ist, was diese neuen Werte darstellen. Andere Autoren (Knoppen und Saris, 2009a; b) fanden heraus, dass diese Zusammenfassungen auf die spezielle Auswahl von Items zurückzuführen ist. Die Werte könnten wieder getrennt erkennbar sein, sobald bestehende Items durch andere, bessere ausgetauscht würden. Dies würde es erleichtern, die Theorie besser darzustellen. Dennoch bleibt die Frage bestehen, ob solche Maßnahmen ausreichen. Eine Verknüpfung von quantitativen und qualitativen Methoden könnte frischen Wind in die Diskussion bringen. Denn bisher behandelte die wissenschaftliche Debatte qualitative und quantitative Methoden im Zusammenhang mit der Werteskala von Schwartz nur getrennt voneinander: Im Zusammenhang mit ihren quantitativen Ergebnissen verzichteten die Forschergruppen bisher auf eine Bewertung der inhaltlichen Ebene und der kognitiven Prozesse bei der Beantwortung der Items.

Ein anderes Problem stellt die Auswahl der Analysemethode dar. Während die Theorie die Beziehungen der Wertestruktur klar beschreibt, liefern verschiedene statistische Auswertungsmethoden (exploratorische und konfirmatorische Faktorenanalyse, multidimensionale Skalierung) voneinander und von der Theorie abweichende Ergebnisse. Sind dadurch sämtliche Methoden gleich zu bewerten, oder gibt es für die Analyse der Werte solche, die besser oder schlechter funktionieren? Die vorliegende Arbeit thematisiert diese Unterschiede, und die Wahl der Analysemethode wird Teil der Analyse.

Aufgrund mangelnder Messqualität und kritischer Befunde in der wissenschaftlichen Diskussion bestehen Zweifel, ob die verwendeten Items das theoretische Konstrukt der *human*

values nach Schwartz angemessen wiedergeben können. Welchen Beitrag eine Integration aus qualitativen und quantitativen Methoden in die Diskussion einbringen kann, soll die vorliegende Arbeit zeigen.

1.2 Aufbau der Arbeit

Der theoretische Teil (Kapitel 2) spannt den Rahmen auf, innerhalb dessen Werte in den Sozialwissenschaften verortet werden (können) und umfasst eine generelle Definition von Werten sowie eine Abgrenzung zu anderen Konzepten; es folgt ein Aufriss der Wertetheorie von Schwartz, und die theoretische Diskussion endet mit dem aktuellen Stand der wissenschaftlichen Debatte: Kritische und bestätigende Studien zur Wertemessung von Schwartz kommen zu Wort.

Daran schließt das Kapitel 3 rund um methodische Belange an: (a) Probleme, die bei der Messung von Werten auftreten, sind zahlreich. Sie reichen von der Art, wie Werte im Menschen verortet sind, bis hin zu Problemen, die die Konstruktion von Items betreffen. (b) Diese Vielzahl an Problemen macht ein spezielles Design zur Bewertung der Messung nötig, das qualitative und quantitative Methoden vereint. Es werden sowohl die Methoden einzeln als auch ihr Zusammenspiel beschrieben.

Die Empirie (Kapitel 4) umfasst (a) einen deskriptiven, (b) einen exploratorischen und (c) einen konfirmatorischen Teil, der schrittweise in eine Methodenkombination übergeht. Beschreibende Maße der Verteilung der Indikatoren vermitteln ein erstes Bild, das bereits bestimmte Charakteristika aufweist: Die Wertemessung unterliegt aufgrund ihrer Operationalisierung Antworttendenzen. Ausnahmen sind bereits erste Indizien für die spätere Detailanalyse. Strukturaufdeckende Verfahren zeigen, dass die Wahl der Methode Einfluss auf die Anzahl herausgerechneter latenter Faktoren hat. Gründe dafür werden erläutert. Die Analyse der fehlenden Werte fällt ebenso in dieses Teilkapitel, da analysierte Zusammenhänge als strukturentdeckend aufgefasst werden. Im konfirmatorischen Teil werden Messmodelle zuerst einzeln und dann in einem paarweisen Vergleich getestet. Die Tests bestätigen Annahmen aus der exploratorischen Analyse und der wissenschaftlichen Debatte über die theoretischen Zusammenhänge der *human values* nach Schwartz. In einem Mixed-Methods-Design ergänzen Ergebnisse aus qualitativen Interviews die quantitative Analyse. Letztere wählt jene Items aus, die im quantitativen Paradigma gemäß gültiger Standards problematische Messeigenschaften besitzen. Qualitative Befunde helfen, Ursachen und theoretische Verknüpfung zu verstehen und zu erklären.

Abschließend gibt Kapitel 5 einen zusammenfassenden Überblick über die empirischen Ergebnisse, und Nutzen und Mängel der angewandten Methode werden reflektiert.

2 Theorie

2.1 Werte in den Sozialwissenschaften – Ein Überblick

Werte, wie sie in den Sozialwissenschaften verwendet werden, sind schwierig zu definieren. Ein Blick auf die umfassende Literatur zum Thema „menschliche Werte“ eröffnet ein Gebiet, das je nach Disziplin und Erkenntnisinteresse Werte unterschiedlich erfasst. Am ehesten lassen sich widersprechende Ansichten über Werte durch eine negative Definition vereinen, also das, was Werte nicht sind. Bevor daher im folgenden Kapitel auf die spezielle Definition von Schwartz eingegangen wird, scheint es nützlich, vorweg Abgrenzungen und kritischen Stimmen Raum zu geben. Sie führen dem Forscher vor Augen, wie ungenau und vage Aussagen über Werte sind. Schließlich muss auch Platz für die fundamentale Frage sein: Können wir Werte tatsächlich als einen eigenständigen Erkenntnisgegenstand betrachten?

Hechter (1992) nennt vier Eigenschaften von Werten, die die Beschäftigung mit ihnen so unüberschaubar machen:

a) Werte können verschiedene Formen annehmen, sind vielschichtig und daher nicht direkt beobachtbar. Sie können sich durch das Maß der Kontrolle, das man über sie hat, unterscheiden, oder in ihrer Anwendung, wenn z.B. einige nur in bestimmten Situationen aktiviert werden, während andere jederzeit ein gültiges Motiv darstellen. Sie sind verschieden in dem Maße, wie sehr sie innerhalb einer Gesellschaft von den Mitgliedern geteilt werden. Hechter, aus der Handlungstheorie – genauer: der Rational-Choice-Theorie – kommend, betrachtet Werte deshalb auch in zwei großen Kategorien: *Einerseits* gibt es jene, die als instrumentell gesehen werden. Diese werden von den meisten Menschen geteilt, weil sie sich an der Vermehrung austauschbarer Güter orientieren. *Andererseits* gibt es jene, die den Individuen innewohnen. Während erstere das Gemeinsame darstellen, ermöglichen es letztere den Menschen, sich als eigene Personen wahrzunehmen.

b) Der derzeitige Stand der Theorie gibt wenig Orientierungshilfe, wie Werte das Verhalten tatsächlich beeinflussen könnten.[1] Werte werden in der Psychologie, Sozialpsychologie und der Ökonomie auf Bedürfnisse, Instinkte und das Konzept des Nutzens zurückgeführt, das dann das individuelle Verhalten erklären soll.

1 Eine mögliche Verknüpfung wird in der *Theory of Planned Behavior* (Ajzen, 1991) dargelegt.

Daneben gibt es auch soziologische Erklärungsansätze, wie z.B. Webers Theorie eines kausalen Zusammenhangs zwischen der Religion, den damit transportierten Werten und einer Ethik der Arbeit und des Wirtschaftens. Diese Forschungstraditionen bestehen derzeit gleichwertig nebeneinander.

c) Solange nicht geklärt ist, wie Werte selbst entstehen, ist ihre Erklärungskraft für Verhalten nicht überzeugend. Neben den vorherrschenden Erklärungen, dass Werte Optimierungsstrategien für die Maximierung austauschbarer Güter (Reichtum, Status, Macht) sind, gibt es auch solche, die belegen, dass Menschen auch ohne Zweckrationalität handeln. Nämlich dann, wenn sie moralisch handeln oder Ritualen nachgehen.

d) Aus genannten Gründen ergibt sich auch, dass sich die Messung von Werten problematisch gestalten kann. Werte sind nicht oder nur schwer von außen beobachtbar. Daher verlässt sich die Forschung zumeist auf Selbsteinschätzungen der Befragten. Das setzt allerdings voraus, dass diese auch klare Vorstellungen über die erfragten Konzepte haben[2]. Manche Kritiker gehen so weit zu sagen, dass Werte vollständig kontextbasiert sind und daher auch nur in Kontexten erfasst werden können (Fischhoff, Slovic und Lichtenstein, 1980).

Hitlin und Piliavin (2004) ergänzen Hechters Überlegungen um zwei weitere Punkte:

e) Werte werden in der wissenschaftlichen Diskussion oft mit anderen sozialpsychologischen Phänomenen (z.B. Einstellungen, Charakterzügen und Normen) in einen Topf geworfen.

f) Der Gehalt von Werten ist historischer und kultureller Veränderung unterworfen. Sie als Konstante zu betrachten, tut ihrer Erforschung keinen Dienst.

Notwendig für das Verständnis scheint vor allem die Abgrenzung von anderen Phänomenen, wie sie in Punkt e) angesprochen wird. Es gibt zu diesem Thema mehr oder weniger eindeutige Parameter, die die Konzepte klären. Von Hitlin und Piliavin (2004) werden Einstellungen, Charakterzüge und Normen angeführt.

Einstellungen

Betrachten wir zuerst die nächsten Verwandten – die Einstellungen – , die am häufigsten für Verwechslungen sorgen (vgl. Hitlin und Piliavin, 2004). Unterschiede sollten beachtet wer-

2 Wie Bernard et al. (2003) in ihrer Studie zeigen, handelt es sich bei Werten tatsächlich um Konzepte, die nicht verfestigt und konkret in den Befragten schlummern, sondern sehr vage sind.

den, weil sie Auswirkungen auf die Messbarmachung von Werten haben. Zunächst wäre zu nennen, dass Werte abstraktere Vorstellungen sind als Einstellungen. Im Gegensatz zu Werten, die immer etwas Wünschenswertes formulieren, können Einstellungen sowohl positive als auch negative Bewertungen von Objekten darstellen. Werte stellen Ideale, Einstellungen konkrete soziale Objekte in den Fokus der Bewertung (Rokeach, 1973).

Weil sich die vorliegende Arbeit im Speziellen mit der Werteskala von Schwartz beschäftigt, dessen Arbeit besonders von Milton Rokeach geprägt wurde, soll nun Rokeachs Abgrenzung von Werten und Einstellungen gesondert hervorgehoben werden: Rokeach definiert einen Wert als eine einzige Glaubensvorstellung einer ganz speziellen Form: „… [a value] refers to a single belief of a very specific kind“ (1973: 18). Eine Einstellung hingegen: „[…] refers to an organization of several beliefs around a specific object or situation“ (ebd.: 18). Diese Definition von Einstellungen deckt sich auch in gewissem Maß mit der von Tourangeau, Rips und Rasinski (2000). Nach ihrem Verständnis sind Einstellungen Konzepte, bei denen oft mehrere Vorstellungen zu einem Thema gleichzeitig vorherrschen, die auch im Konflikt miteinander stehen können. Eine Einstellung betrifft damit immer eine Gruppierung mehrerer Überzeugungen zu einer jeweils spezifischen Situation. Am Beispiel der Likertskala erklärt bedeutet das, dass Überzeugungen zu einem bestimmten Thema bewertet werden und schließlich zu einem Index zusammenfließen. Werte hingegen beziehen sich auf eine einzige spezifische Überzeugung. „It [i.e. a value] concerns a desirable mode of behaviour or end-state that has a transcendental quality to it, guiding actions, attitudes, judgments, and comparisons across specific objects and situations and beyond immediate goals to more ultimate goals“ (Rokeach, 1973: 18).

Werte sind Standards, Einstellungen nicht. Eine Vielzahl von Einstellungen lässt sich auf eine überschaubare Zahl von Werten zurückführen. Werte beziehen eine zentralere Position in der Zusammensetzung der Persönlichkeit und können Einstellungen und Handlungen mehr oder weniger direkt anregen (Rokeach, 1973).

Weitere Unterschiede, zu denen ein grober Konsens in der wissenschaftlichen Diskussion herrscht – so Hitlin und Piliavin (2004) – , sind:

- Wenn Standards des internen Bewertungsprozesses sich von Werten herleiten lassen, stehen sie in der Hierarchie höher als solche, die auf Einstellungen basieren.
- Werte sind damit von zentralerer Bedeutung in der „Zusammensetzung“ der Persönlichkeit.

- Dadurch sind sie auch weiter entfernt von der direkten Einflussnahme auf tatsächliches Verhalten.
- Das macht sie beständiger gegenüber Veränderungen im Laufe eines Lebens.

Dass es einen empirischen Unterschied zwischen den Konzepten „Werte“ und „Einstellungen“ gibt, haben Maio und Olson (2000) herausgefunden, indem sie das verbindende Glied zwischen ihnen gefunden haben (sie nennen es *goal-expressive attitudes*). Andere Forscher vertreten die Ansicht, dass Werte eine spezielle Form oder Funktion von Einstellungen seien. Kristiansen und Zanna (1991) hingegen sind der Meinung, dass Einstellungen Werte ausdrücken oder aber deren Wahrnehmung verändern können (diesen Effekt nennen sie deshalb *halo effect*). Man sieht: Viele Forscher vertreten zwar viele Ansätze, die aber alle anerkennen, dass es einen prinzipiellen Unterschied zwischen den Konstrukten gibt.

Charakterzüge

Traits (Charakterzüge/Eigenschaften) sind ein weiteres Konzept, das oft mit Werten verwechselt wird. Vereinfacht lässt sich eine Eigenschaft definieren, wenn eine bestimmende Qualität beschrieben wird („Man *ist* schön, klein, glücklich etc.“). Rokeach vertritt die Überzeugung, dass Verhalten, das von Charakterzügen ausgeht, weniger bewusster Kontrolle durch die Person unterliegt als jenes, das durch Werte angeregt wird. Rokeach sieht in der Unterscheidung aber mehr eine ideologische Trennung als eine empirisch oder theoretisch fundierte. Er meint, man könne den Charakter einer Person als Bündel von Charakterzügen betrachten und ihre Unabänderlichkeit hinnehmen. Genauso gut aber könnte man ihn als System von Werten betrachten: Das, was wir von außen sehen (z.B. einen autoritären Charakterzug), kann von der Person selber so betrachtet werden, dass sie besonderen Wert auf bestimmte Endzustände legt (z.B. Gehorsam, Sauberkeit, Höflichkeit) (Rokeach, 1973).

Roccas, Sagiv, Schwartz und Knafo (2002) finden eine vertiefende Trennung der Konzepte zielführend und kommen zu der Gegenüberstellung in Tabelle 1:

Charakterzüge	Werte
dauerhafte Veranlagungen	dauerhafte Ziele
„wie Menschen sind“	„was Menschen für wichtig erachten“
verändern sich durch Häufigkeit und Intensität	verändern sich in ihrer Wichtigkeit
können positiv oder negativ sein	sind immer wünschenswert
	dienen als Maßstab zur Bewertung von eigenem und fremdem Verhalten
	rechtfertigen die Wahl oder das Verhalten an sich
Sowohl Werte als auch Eigenschaften können herangezogen werden, um Verhalten zu erklären.	

Tabelle 1: Unterschied Eigenschaften – Werte (aus: Roccas et al., 2002)

Die Forschergruppe hat sinnvolle Zusammenhänge zwischen den Konzepten festgestellt. Da sie sich nicht vollständig überlappen, muss es sich um eigenständige Konstrukte handeln.

Normen

Um Normen handelt es sich, wenn ein: „Man *sollte* …“ gedanklich dahinter steht. Befragten berichten über ein Gefühl des Zwangs, wenn sie einen Stimulus als Norm wahrnehmen, anders als bei Werten. Werden Normen an einen bestimmten Kontext gebunden wahrgenommen (vgl. Hitlin und Piliavin, 2004), definiert u.a. Schwartz Werte als transsituational. Sie liegen ohne Kontext in den Menschen vor. Ein anderer Unterschied betrifft die Ebene, auf der die Konstrukte erhoben werden: Werte werden auf einer individuellen Ebene erhoben, Normen gelten für die gesamte Gesellschaft. Dennoch gilt auch für Werte, dass sie eine Basis besitzen, die durch den Austausch mit anderen geteilt wird: „Values, like norms, are a group-level phenomenon requiring shared agreement“ (Hitlin und Piliavin, 2004: 361).

Hechter fasst den Unterschied so zusammen: „Norms emanate from sources that are external to the individual and must therefore be associated with sanctions in order to be effective“ (Hechter et al., 1999: 409). Werte hingegen sind in den Individuen intrinsisch vorhanden und kommen ohne Sanktionen aus. Sie handeln vom Wünschenswerten, während Normen das Akzeptable verkörpern.

2.2 Theorie der human values *nach Shalom Schwartz*

Aufbauend auf den theoretischen Vorarbeiten von Schwartz und Bilsky (1987; 1990) bringt Schwartz (1992) eine empirische Untersuchung seiner Theorie der *human values* in die wissenschaftliche Debatte ein. Vorläufer seiner Arbeit im Bereich menschlicher Werteforschung in den Sozialwissenschaften sind Rokeach (1973), Williams (1968) und Kluckhohn (1951).

Werte werden von ihnen als Hilfsmittel definiert, *einerseits* um eigene Handlungen zu legitimieren, während sie *andererseits* dazu dienen, andere Menschen, sich selbst und Ereignisse zu bewerten. Schwartz übernimmt diese Definition für die Entwicklung seines Konzepts. Die forschungsrelevante Neuheit seiner Tätigkeit ist die Entwicklung einer Skala, die auf einem theoretischen Konstrukt beruht, das Vorhersagen über Konflikte und Übereinstimmungen von Werten vorhersagt, und es ermöglicht, unterschiedliche Kulturen miteinander zu vergleichen. Das ist möglich, da Schwartz von der Hypothese ausgeht, dass die von ihm beschriebene Struktur der Werte universell ist und damit von allen Menschen geteilt wird.

Theorie und Entwicklung

Schwartz (2007) charakterisiert Werte folgendermaßen:

- Werte sind Konzepte oder Glaubensvorstellungen, die mit Emotionen verknüpft sind.
- Sie zielen auf wünschenswerte Endzustände oder Verhaltensweisen ab und motivieren zu solchen.
- Sie sind nicht auf spezielle Situationen beschränkt, sondern in ihrer Anwendung allgemeiner zu verstehen.
- Sie werden von Individuen als Entscheidungshilfen bei der Auswahl adäquater Verhaltensweisen herangezogen und dienen als Standards oder Kriterien, wenn die eigenen oder fremde Verhaltensweisen bewertet werden. Sie werden ebenfalls herangezogen, wenn es um die Einschätzung von Situationen geht.
- Werte sind in einer Hierarchie der relativen Wichtigkeit zueinander geordnet.
- Die relative Wichtigkeit *multipler* Werte ist ausschlaggebend für die Wahl einer Handlung. Stimuli aktivieren die gesamte Wertestruktur auf einmal. Je nach Kontext variieren die relativen Wichtigkeiten (Schwartz, 2007: 170).

In einem Satz zusammengefasst, bringt Schwartz Werte auf den Punkt: Sie sind „transsituational goals, varying in importance that serve as guiding principles in people's lives" (Schwartz und Bardi, 2001: 269). Als Abgrenzungskriterium gegenüber Einstellungen sieht Schwartz die Punkte (3) „Generalisierbarkeit" und (5) „Rangordnung" ausschlaggebend.

Bestimmend für den „Inhalt" des Wertes ist das Ziel, auf das dieser Wert gerichtet ist. Diese Ziele werden auch *motivationale Typen* genannt, weil die Motivation zur Erreichung

des Ziels sie beschreibt. Hergeleitet hat Schwartz seine Werte aus Überlegungen zu drei universellen Bedingungen, an die er die menschliche Existenz geknüpft sieht. Das sind:

a. Bedürfnisse von Individuen als biologische Organismen

b. Erfordernisse für koordinierte soziale Interaktion

c. Bedürfnisse zum Überleben und Wohlergehen von Gruppen (Schwartz, 1992: 4).[3]

Aus den drei Bedingungen ließen sich ursprünglich acht motivationale Typen ableiten (prosocial, Self-direction, restrictive confirmative, enjoyment, Achievement, maturity, Security, Power) (Schwartz, 1992: 4). Diese erste Version des Schwartz Value Surveys wurde abgeändert und mündete in einer mit elf Wertetypen, die einer empirischen Untersuchung unterzogen wurden. Sie zeigte, dass für eine universelle Theorie zehn Wertetypen angemessen sind. Diese werden im Folgenden näher beschrieben.

Die Wertetypen

Die Wertetypen werden entsprechend ihrer motivationalen Ziele, ihrer Ableitung von den drei Bedingungen definiert und durch die zugehörigen Werte, die den Typ erfassen, gemessen. Es soll an dieser Stelle ausdrücklich darauf hingewiesen werden, dass es sich bei den „Werten“ um Typen handelt. Die einzelnen Werte, die sich unter einem Wertetyp zusammenfassen lassen, werden durch Indikatoren gemessen. Je nach Breite der Ziele jedes Wertetyps werden mehr oder weniger Werte erhoben. Die englische Bezeichnung der Wertetypen wurde im Text beibehalten, da die meiste Forschung auf dem Gebiet auf Englisch verfasst ist und die Übertragung und ein Vergleich dann leichter fallen. Die deutschen Übersetzungen werden beigefügt. Tabelle 1 fasst die von Schwartz postulierten zehn Wertetypen, die ihnen zugrunde liegenden motivationalen Ziele und die sie repräsentierenden Einzelwerte (in Klammern) zusammen:*

3 Schwartz zitiert Buss (1986), wenn er sich darauf bezieht, dass diese drei Bedingungen von einem evolutionären Standpunkt aus sehr prägend waren.

POWER/MACHT: Sozialer Status und Prestige, Kontrolle oder Dominanz über Menschen und Ressourcen (soziale Macht, Autorität, Reichtum, mein öffentliches Ansehen wahren).
ACHIEVEMENT/LEISTUNG: Persönlicher Erfolg durch die Demonstration von Kompetenz bezüglich sozialer Standards (erfolgreich, fähig, ehrgeizig, einflussreich).
HEDONISM/HEDONISMUS: Vergnügen und sinnliche Belohnungen für einen selbst (Vergnügen, das Leben genießen).
STIMULATION/STIMULATION: Aufregung, Neuheit und Herausforderungen im Leben (wagemutig, ein abwechslungsreiches Leben, ein aufregendes Leben).
SELF-DIRECTION/SELBSTBESTIMMUNG: Unabhängiges Denken und Handeln, schöpferisch tätig sein, erforschen (Kreativität, Freiheit, unabhängig, neugierig, eigene Ziele auswählen).
UNIVERSALISM/UNIVERSALISMUS: Verständnis, Wertschätzung, Toleranz und Schutz des Wohlergehens aller Menschen und der Natur (tolerant, Weisheit, soziale Gerechtigkeit, Gleichheit, eine Welt in Frieden, eine Welt voll Schönheit, Einheit mit der Natur, die Umwelt schützen).
BENEVOLENCE/BENEVOLENZ: Bewahrung und Erhöhung des Wohlergehens der Menschen, zu denen man häufigen Kontakt hat (hilfsbereit, ehrlich, vergebend, treu, verantwortungsbewusst).
TRADITION/TRADITION: Respekt vor, Verbundenheit mit und Akzeptanz von Gebräuchen und Ideen, die traditionelle Kulturen und Religionen für ihre Mitglieder entwickelt haben (fromm, meine Stellung im Leben akzeptieren, demütig, Achtung vor der Tradition, gemäßigt).
CONFORMITY/KONFORMITÄT: Beschränkung von Handlungen, Neigungen und Impulsen, die andere beleidigen oder verletzen könnten oder gegen soziale Erwartungen und Normen verstoßen (Höflichkeit, Gehorsam, Selbstdisziplin, ehrerbietig gegenüber Eltern und älteren Menschen).
SECURITY/SICHERHEIT: Sicherheit, Harmonie und Stabilität der Gesellschaft, von Beziehungen und des Selbsts (familiäre Sicherheit, nationale Sicherheit, soziale Ordnung, sauber, niemandem etwas schuldig bleiben).
*Da es in ihrer Bedeutung kulturelle Unterschiede gibt, werden folgende Werte des SVS bei der Indexbildung der Wichtigkeit der zehn Wertetypen nicht berücksichtigt: soziale Anerkennung, Intelligenz, Selbstrespekt, innere Harmonie, wahre Freundschaft, ein spirituelles Leben, reife Liebe, Sinn im Leben, Loslösung, Gefühl der Zugehörigkeit, Gesundheit.

Tabelle 1: Wertetypen (aus: Schmidt et al., 2007)

Spiritualität war in der ersten Version des SVS vorhanden, weil sich dieser Wertetyp aus den universellen Bedürfnissen ableiten ließ. Als universeller Wertetyp stellte er sich dennoch als problematisch heraus. *Zum einen* setzen diesbezügliche Erhebungen eines Mindestmaßes bei den Befragten differenzierte, aufwändige Überlegungen voraus, mit denen die meisten Menschen nur selten in Berührung kommen. Stattdessen können die Wertetypen Tradition, Sicherheit und Konformität das Spektrum von Spiritualität abdecken. *Zum anderen* repräsentieren in unterschiedlichen Gruppen sehr verschiedene Werte Spiritualität. Es kann daher nicht auf eine bestimmte Auswahl an universell charakteristischen spirituellen Werten zu-

rückgegriffen werden, auch wenn ein abstrakter Spiritualitätstyp von Werten universell zu sein scheint. Aus genannten Gründen wurde dieser Wertetyp aus der weiteren Analyse und damit der Theoriebildung von Schwartz ausgeschlossen.

Dynamische Struktur der Wertetypen

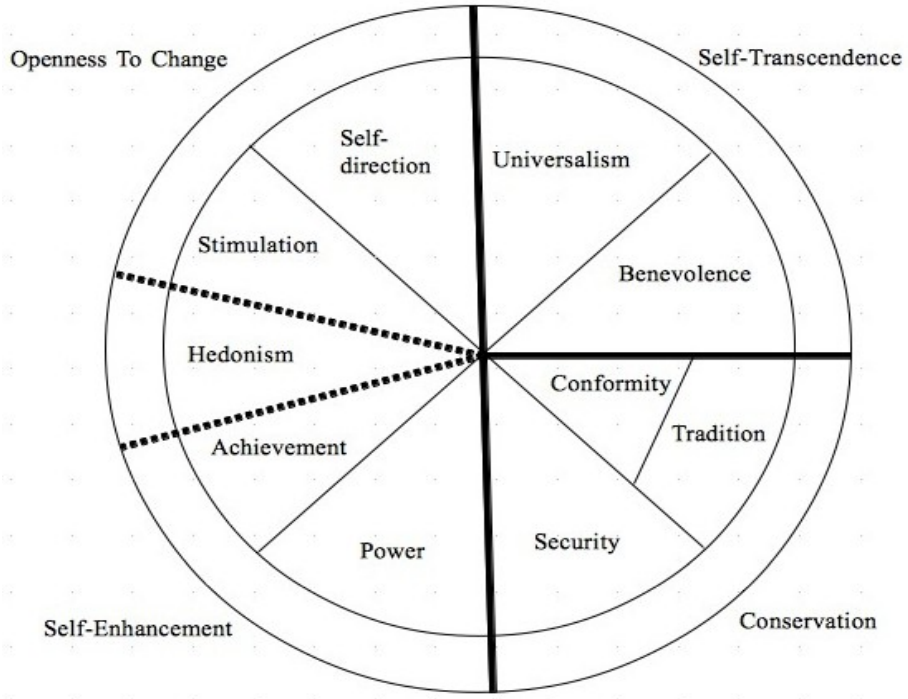


Abbildung 1: Zirkumplex mit Wertetypen und Wertetypen höherer Ordnung (abgeändert von: Schmidt et al., 2007)

Schwartz verknüpfte im Gegensatz zu seinen Vorgängern die Wertetypen in seiner Theorie auf dynamische Weise. Das bedeutet, dass Werte immer in relativer Wichtigkeit zu den übrigen bewertet werden. In einer empirischen Untersuchung verteilen sich daher die motivationalen Typen in einer kreisähnlichen Anordnung (siehe Abbildung 1), wenn sie auf einen zweidimensionalen Raum projiziert werden. Schwartz spricht gemäß der Definition von Guttman (1954) von einem *Zirkumplex*. Dieser ist durch ein bestimmtes Muster von Zusammenhängen, die entlang eines Kreises verlaufen, definiert. Benachbarte Wertetypenpaare sind miteinander kompatibel. Ihre gleichzeitige stark positive Ausprägung ist ohne Probleme möglich. Verfolgt hingegen ein Individuum zwei gegenüberliegende Werte, führt dies zu starken psychologischen oder sozialen Spannungen, weil deren Motive entgegengesetzt sind. Strenggenommen müssten sich für einen Zirkumplex die Elemente in gleicher Distanz zueinander befinden. Diese Voraussetzung wird nicht erfüllt, ebenso wenig die Annahme, dass sich alle

Elemente auf einem Radius befinden. Deshalb bezeichnet Schwartz seine Struktur als modifizierten Quasizirkumplex (Schwartz und Boehnke, 2004: 234, Fußnote).

Aus der kreisähnlichen Struktur lassen sich Hypothesen ableiten, die für die Qualität der Messung als Grundannahmen dienen können:

H1: Je näher zwei Wertetypen liegen, desto größer ist der Teil der Varianz, die sie miteinander teilen (die Korrelationen sind stärker).

H2: Wenn Wertetypen einander gegenüber liegen, müssen ihre Korrelationen negativ sein, um die interne Konsistenz nicht zu gefährden.

H3: Die Zusammenhänge der Items sind so festgelegt, dass jedes Item stärker mit dem Set von Items korreliert, das denselben Wertetypen misst, als mit jenem, das ein anderes misst (Schwartz und Boehnke, 2004: 232).

H4: Nur die Wertetypen hängen in einer kreishaften Weise zusammen. Für einzelne Items gilt das nicht (Schwartz und Boehnke, 2004: 232).

Die Achsen bilden sogenannte *higher-order values* (Werte höherer Ordnung, siehe Abbildung 1). Diese werden als *openness to change* (Offenheit für Wandel) *vs. conservation* (Bewahrung) einerseits und *self-enhancement* (Selbsterhöhung) *vs. self-transcendence* (Selbstüberwindung) andererseits bezeichnet. In dieser Einteilung nimmt Hedonism eine Sonderposition ein. Theoriegeleitet sollte Hedonism in den Bereich *self-enhancement* fallen. Diesem Wertetyp fehlt allerdings die kompetitive Einstellung von Achievement und Power, die ebenso in dieser Dimension zu liegen kommen. Andererseits weist Hedonism einen Kern von Stimulierung und Herausforderung auf, der Hedonism in die Nähe von *openness to change* rückt. Damit liegt dieser Wertetyp zwischen zwei *higher-order values* und vereint beide Dimensionen in sich.

Werte selbst stellen ein motivationales Kontinuum dar. Das soll bedeuten, dass zwei nahe beieinander liegende Werte große Teile eines Ziels teilen, weil sie je nach Nähe zu einem benachbarten Segment bereits von diesem mit eingefärbt sind, dessen Motivation also bereits zum Teil mittragen. Als empirisches Indiz dafür dient der Zusammenhang mit Kovariaten (Geschlecht, Alter etc.): Ihre Korrelation mit den Wertetypen ergibt eine sinusförmige Kurve, das heißt, die Intensität des Zusammenhangs nimmt der Kreisstruktur folgend kontinuierlich zu und ab, was das Bild einer Schwingung ergibt (Schwartz und Boehnke, 2004: 232). Für Forschungszwecke ist es dienlicher, die Wertetypen in diskrete Segmente zu unterteilen (Schwartz, 1992: 46), da diese vorhergesagte Konflikte und Übereinstimmungen von Werten

besser nachvollziehbar machen. An sich wären Werte kontinuierlich. Die Trennung, wonach dieses oder jenes Item in ein Segment aufgenommen wird, basiert rein auf den Überlegungen des Forschers (ebd.: 45).

Ein besonderes Ergebnis der empirischen Untersuchung war, dass zwischen Conformity und Tradition in 80,5 % der Fälle Tradition in dem für Conformity vorgesehenen Segment an der Peripherie zu liegen kam. Die Werte waren nicht zufällig durchmischt, sondern durch die Distanz vom Mittelpunkt signifikant getrennt. Das machte eine Anpassung der Theorie nötig. Dadurch, dass Tradition nicht mehr neben Conformity lag, sondern dahinter, kamen Benevolence und Conformity nebeneinander zum Liegen. Da sich Tradition und Conformity im selben Segment befinden, liegt der Schluss nahe, dass sie ein motivationales Ziel teilen. Den gemeinsamen Nenner bezeichnet Schwartz als „subordination of self in favor of socially imposed expectations“ (Schwartz, 1992: 40), also die Unterwerfung des Selbsts unter die sozialen Erwartungen. Ob eine Aufteilung innerhalb des Segments sinnvoll ist, hängt von den Vergleichsvariablen und davon ab, ob eine diesbezügliche Unterscheidung theoretisch begründet ist.

Abstände vom Mittelpunkt

Die räumliche Distanz eines Wertes zum Mittelpunkt der Kreisstruktur ist eine Größe, die den Zusammenhang mit anderen Werten beeinflusst. Indikatoren, die an der Peripherie liegen, werden schwächer positiv/stärker negativ mit gegenüberliegenden Items und stärker positiv/schwächer negativ mit benachbarten korrelieren. Abstände zum Mittelpunkt geben an, wie stark sich der Konflikt entgegengesetzter motivationaler Ziele ausdrückt, also wie leicht ein Ziel für ein anderes aufgegeben werden kann. Distanzen spielen daher eine Rolle bei der Bestimmung der Wichtigkeit des motivationalen Reizes in Beziehung zu anderen:

> The full pattern of item correlations is relevant in defining a value’s motivational significance, because values form a coherent system of complementary and competing motivations (Schwartz, 2003: 319).

Darüber hinaus zeigen Abstände, wie nahe der Wert einem Menschen geht: „[They] reflect differences between constructs in the degree of their abstractness, closeness to the self, or prevalence in everyday interaction […]“ (Schwartz und Boehnke, 2004: 235). Schwartz und Boehnke beziehen sich auf eine nicht publizierte Doktorarbeit von Melech (2001), wenn sie behaupten, dass eine zentralere Position auch größere normative Wichtigkeit bedeuten kann. Man sieht: Abstände zum Mittelpunkt haben vielfältige Bedeutung und sind schwer konkret

zu interpretieren. Genauere Erkenntnisse zu ihrer Bedeutung und welchen Einfluss sie auf die Messung und Vergleichbarkeit der *human values* von Schwartz haben, sind noch ausständig.

SVS als Erhebungsinstrument

Empirisch wurde der Schwartz Value Survey (SVS) das erste Mal in einer umfassenden Studie in 20 verschiedenen Ländern in 13 unterschiedlichen Sprachen untersucht (zur genaueren Durchführung siehe Seite 22). Die finale Version des Fragebogens enthielt 56 Items, von denen 52 die postulierten zehn Wertetypen und vier Spiritualität erfassten. Eine Hälfte wurde als terminale (Nomina) und die andere Hälfte als instrumentelle (Adjektiva) Werte formuliert und in zwei separaten Listen den Teilnehmern vorgelegt[4]. Die Items wurden so formuliert, dass die Befragten jeden Wert mit „AS A GUIDING PRINCIPLE IN MY LIFE“ bewerten sollten, indem sie auf einer Neunpunkteskala die entsprechende Ausprägungen wählen [*of supreme importance (7), very important (6),* (Kategorien 5 und 4 blieben unbenannt), *important (3),* (Kategorien 1 und 2 blieben unbenannt), *not important (0), opposed to my values (-1)*]. Mit der Bewertung (-1) wollte Schwartz auch „negative“ Werte messen, also solche, die die Menschen in ihren Handlungen und Wahlmöglichkeiten nicht zeigen wollen (Schwartz, 1994a: 26). Items wurden entweder als Nomen oder Adjektiv formuliert, mit einer knappen, beschreibenden Phrase in Klammern.

Befragten wählten zu Beginn dasjenige Item aus, das sie am ehesten repräsentierte. Dasselbe geschah mit dem Item, das sie als am geringsten auf sich zutreffend sahen. Anschließend sollten sie die übrigen Indikatoren bewerten. Damit wurden die Bewertungen „geankert“ (relative Differenzen ergeben sich in Abhängigkeit von dem ähnlichsten und unähnlichsten Wert). Mit dieser Vorgehensweise umgeht Schwartz das sogenannte *ranking*, bei dem Paarvergleiche zwischen den einzelnen Werten vorgenommen werden. Seiner Meinung nach überfordere das die generelle kognitive Leistungsfähigkeit, die Befragten zugemutet werden kann (Schwartz, 1994a: 27). Die Einschätzung der Befragten erfolgt mittels Bewertung (*rating*) und nicht mittels Rangordnung (*ranking*). Methodologisch betrachtet bietet ein *rating* bessere statistische Eigenschaften für quantitative Analysen, da es sich um ein höheres Skalenniveau handelt. Befragten sind nicht gezwungen, zwei für sie gleichwertige Werte in eine hierarchische Ordnung zu bringen. Darüber hinaus bringt ein *rating* den Vorteil mit sich, „negative Werte“ – wie es Schwartz ausdrückt (Schwartz, 1994a: 26) – zu erfassen, also solche, die von den Menschen abgelehnt werden. Werte zu bewerten, anstatt sie zu hierarchisie-

4 Diese Unterscheidung wurde u.a. von Rokeach vertreten. Es handelt sich dabei um den Charakter der Werte: Terminale beschreiben Endzustände, instrumentelle eine Handlungsanweisung. Diese Unterscheidung konnte, wie Schwartz (1992) zeigte, in seiner Studie nicht aufrecht erhalten werden.

ren liegt auch näher an der Art, wie Menschen Werte konzeptionell betrachten, so Schwartz. Oft sind Unterschiede zwischen den Wertetypen für die Befragten nicht immer ganz klar, und eine Prioritätensetzung ist für sie daher nicht immer möglich (Schwartz, 1992: 17).

PVQ – die „light" Version des SVS

Schwartz publizierte mit Kollegen (2001) eine abgewandelte Version des SVS. Es handelt sich um den sogenannten Portrait Value Questionnaire (kurz: PVQ), dessen verkürzte Version in der vorliegenden Arbeit untersucht wird. Wie der Name schon sagt, werden Porträts von Personen beschrieben. Ein Porträt in diesem Sinne ist die prägnante Beschreibung eines Menschen, die ein spezifisches Ziel, ein Streben nach etwas oder einen Wunsch wiedergibt. Ein Beispiel wäre: „Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht." Jedes Item ist ein solches Porträt, das mit einem spezifischen Wert (in diesem Fall *soziale Macht*) gekoppelt ist. Die Befragten geben an, wie ähnlich oder unähnlich sie diesem Porträt sind. Dadurch bewerten sie nicht die Werte an sich. Diese werden implizit aus den Selbsteinschätzungen der Befragten gefolgert (Schwartz, 2003). Schwartz betont, dass es sich bei den Porträts um Personen mit bestimmten Zielen und Bestrebungen handelt. Wären sie nur so beschrieben, dass sie Charakterzüge zeigten, dann wäre es schwierig, sich mit so einer Personen zu vergleichen, wenn einem diese Eigenschaft selber fehlte. Hingegen könne ein Befragter immer noch jemandem gleichen, dem ein Ziel wichtig ist, ohne es selber erreicht zu haben. Diese Annahme kann für die hier geprüfte Fassung des PVQ nicht für alle Items bestätigt werden. Einige Porträts werden sehr stark als Charakterisierung aufgefasst. Diese Items werden im empirischen Teil im Detail besprochen.

Hinter der Neuorientierung der Messung liegen zwei Überlegungen:

> First, it should be more concrete and less cognitively complex than the SVS, hence usable with populations for which the SVS was apparently not suitable. Second, it should differ substantially from the SVS in its format and judgment task to provide an independent test of the theory of value content and structure (Schwartz et al., 2001: 523).

Das Central Coordinating Team (kurz CCT), das innerhalb des European Social Survey (kurz ESS)[5] für die Qualität der Erhebung verantwortlich ist, beanstandete vor der Aufnahme der PVQ-Items in den ESS, dass PVQ die *Ähnlichkeit* und SVS die *Wichtigkeit* messe – es wäre

5 Der ESS ist laut Eigendefinition: „[…] an academically driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations." Seit 2002 im Einsatz mit damals 22, heute mit 30 Teilnehmerstaaten und in seiner sechsten Runde, legt er die strengsten methodischen Maßstäbe an.

denkbar, dass dies ein Nachteil für PVQ sein könnte, weil die Messung nicht mehr die gleiche Schärfe besäße. Schwartz (2003) argumentierte, dass Menschen üblicherweise wenig Zeit darauf verwenden nachzudenken, was wichtig oder weniger wichtig für sie sei. Daher haben sie meist auch gar keine klaren Antworten auf Fragen bzgl. der Wichtigkeit. Hingegen sei der Vergleich anderer Personen mit sich selbst eine weitaus häufiger anzutreffende Strategie: „People constantly assess others and compare them to themselves" (ebd.: 299). Das CCT akzeptierte diese Begründung.

Im Begleittext zu der Skala wird darauf hingewiesen, dass die Befragten die Portraits *mit sich* vergleichen sollen. Es würde, laut Schwartz (2003) einen Unterschied machen, wäre die Aufforderung anders gerichtet, also sich selbst mit dem Portrait zu vergleichen: persönliche Aspekte, die von dem Portrait gar nicht erfasst werden, würden möglicherweise in den Vergleichsprozess Eingang finden und die Antworten verzerren.

Die Befragten antworten auf einer symmetrischen, sechsstufigen Skala, deren Endpunkte „gleicht mir sehr" und „gleich mir ganz und gar nicht" sind. Für diese Form spricht, dass Personengruppen mit niedrigerer formaler Bildung oder ältere Menschen – sofern sie die Anweisung verstehen – ihre Einschätzung oft schwer in Zahlen formulieren können (Schwartz, 2003: 299). Deshalb legt Schwartz auch Wert darauf, die Antwortkategorien zu benennen (was wiederum für die Annahme der gleich großen Intervalle bei Likertskalen Probleme bereiten kann; siehe Kapitel 3.1.2).

In den folgenden Punkten sind die hervorstechendsten Unterschiede zwischen SVS und PVQ zusammengefasst (Schwartz et al., 2001: 524):

- PVQ enthält in der ursprünglichen Fassung 40, der SVS 56 Items.
- Während PVQ Werte indirekt misst, geschieht dies im SVS über eine bewusste Bewertung durch die Befragten.
- Die Messung geschieht im PVQ über die Selbsteinschätzung der Ähnlichkeit einer anderen Person, die durch ein Porträt repräsentiert wird, mit sich selbst. Der SVS erfragt die Wichtigkeit des vorliegenden Items als leitendes Lebensprinzip („*guiding principle in MY life*").
- PVQ setzt einen Reiz durch Personen, die bestimmte Wünsche, Ziele oder Erwartungen haben. Im SVS sind die Reize abstrakte, kontextfreie Werte. Die kognitiven Anforderungen sind völlig verschieden.

- Im PVQ gibt es sechs beschriftete, im SVS neun teilweise beschriftete Antwortkatego-rien, die sowohl positive als auch negative Zahlenzuweisungen beinhalten.
- Im Gegensatz zum PVQ wird die Rangfolge nicht geankert.

Eine weitere Besonderheit des PVQ ist die Form der Porträts: Sie werden jeweils in zwei Sätzen beschrieben, die so einfach formuliert sein sollen, dass selbst 11-jährige sie verstehen (Schwartz, 2003: 274). Der erste Satz hebt die Wichtigkeit eines bestimmten Wertes hervor. Der zweite umfasst eine konkrete Beschreibung eines Zieles oder einer Emotion. Schwartz begründet dieses Vorgehen damit, dass Stimuli mit einem einzigen Satz in vorherigen Versuchen für Befragte teilweise zu wenig Information lieferten, um ein vollständiges Porträt abzugeben. Eine Verbreiterung der Stimuli helfe darüber hinaus auch bei der Übersetzung in andere Sprachen, da mehr Raum gegeben sei, Nuancen in den Bedeutungen hervorzuheben (Schwartz, 2003: 298). Dieses Vorgehen ist riskant: Die Gefahr liegt darin, dass die Sätze durch Drittvariablen beeinflusst werden und dann Unterschiedliches messen können (siehe Kapitel 3.1.3). In diesem Fall spricht man in der Survey-Methodologie von einem doppelten Stimulus (engl.: *double-barreled stimulus*). Schwartz ist sich des Problems bewusst: „People sometimes thought the two sentences contained inconsistent information“ (Schwartz, 2003: 298). Die Items wurden so lange durch kognitive Interviews angepasst, bis der doppelte Stimulus nicht mehr oder nur noch sehr selten nachgewiesen wurde. Schwartz räumt auch ein: „The idiosyncratic theories of individuals regarding how some values are organized inevitably lead to occasional perceptions of inconsistency“ (ebd.: 298). Meine Methodenkombination hingegen zeigt, dass doppelte Stimuli nicht nur zufällig auftreten, sondern gerade jene Items betreffen, die in der quantitativen Untersuchung als problematisch identifiziert wurden.

Bei Tests mit der Reihenfolge und der Gewichtung der einzelnen Sätze innerhalb eines Items verstanden die Versuchspersonen Unterschiedliches. Die Forscher fanden leichte Abweichungen im Inhalt, den die jeweiligen Sätze vermittelten („[…] subtle differences in what the two sentences in each item convey“ (Schwartz, 2003: 301)). Teilweise war einer in seiner Bedeutung stärker und wurde daher als weniger ähnlich wahrgenommen. Dann wiederum gab es Teilsätze, die die Bedeutung des anderen überhaupt veränderten, sie manchmal klarer oder auch konkreter machten. Letztlich wurden jene Kombinationen erwogen, die die Motivation des Wertetypen am besten wiedergaben: „[…] we chose the order that we thought did a better job of capturing the pure motivation we sought for the particular value“ (ebd.: 301). Leider

zeigen die kognitiven Interviews der vorliegenden Arbeit, dass die Wertigkeit der Teilsätze immer noch ein Problem darstellt und zu verzerrenden Antworten führen kann.

Die finale Version, die in den ESS aufgenommen wurde, ist im Anhang (Tabelle 18) zu finden.

Die Unterscheidung kultureller und individueller Werte

Obwohl seine Skalen geeignet sind, Länder zu vergleichen, weis Schwartz ausdrücklich darauf hin, dass Aussagen über Wertedimensionen auf der individuellen Ebene getroffen werden. Schwartz (1994a) fand Hinweise, dass bei Analysen auf der kulturellen Ebene eine etwas andere Struktur auftritt, wenn die Kulturen und nicht die Individuen als Einheit fungieren. Wenn ein Individuum seine eigenen Werte in der Gesellschaft verfolgt, erfährt es dabei psychologische Konflikte und Übereinstimmungen. Auf der anderen Seite stehen die Normen und Strategien der Gesellschaft, um menschliche Aktivitäten zu regulieren. „Seine“ Werte sieht Schwartz als individuelle Werte, die teilweise Produkt der geteilten Kultur und teilweise Produkt individueller Erfahrungen sind (ebd.: 92). Und obwohl kulturelle Werte und individuelle Werte rein statistisch nicht miteinander korrelieren müssen, geht Schwartz aufgrund einer basalen Koppelung der beiden Ebenen davon aus, dass sie sich überlappen. Dafür führt er drei Argumente an: (1) Es gibt konstante psychologische Voraussetzungen in der menschlichen Natur und der Art, wie Interaktionen ablaufen, die die Institutionen der Gesellschaft berücksichtigen müssen, um zu bestehen. (2) Werte werden von Menschen in der Sozialisation aufgenommen und von ihnen gelebt. (3) Es herrschen gesellschaftliche Mechanismen vor, die darüber bestimmen, ob Werte in einer bestimmten Zusammensetzung von den Individuen als gegensätzlich oder übereinstimmend wahrgenommen wird (ebd.: 93)[6].

Ob gesellschaftliche/kulturelle oder individuelle Werte erfasst werden, ist von derAnwendung abhängig:

> The culture-level types are appropriate when one seeks to understand how differences between cultures in their symbol systems, institutions, rates and styles of behavior, and so on are related to cultural value emphases. … The individual-level types should be used when one seeks to understand how differences between individual persons in beliefs, attitudes, or behavior are related to individual differences in value priorities (ebd.: 117).

6 In dem erwähnten Artikel vergleicht Schwartz seinen Value Survey (die Zuordnung zu Wertetypen ist mit Mastery, Hierarchy, Conservatism, Harmony, Egalitarian commitment, Intellectual autonomy und Affective autonomy leicht abgeändert) mit dem von Hofstede, der auf der kulturellen Ebene misst, während der SVS die individuelle Ebene erhebt. Gemäß diesem geglückten Vergleich sollten auch Bedenken bzgl. des Unterschieds der Ebenen beseitigt worden sein. Interessanterweise sind die von Schwartz angeführten Werte sieben an der Zahl, was der Anzahl der Werte in den Davidov-Studien entspricht).

Nachdem die theoretische Grundlage und die Operationalisierung der *human values* nach Schwartz dargelegt wurde, sollen im nächsten Abschnitt durchgeführte Studien die Validität und Reliabilität der Messung beleuchten.

2.3 Kritische Diskussion – Stand der wissenschaftlichen Forschung

Eine Reihe an Publikationen hat sich mit der Messqualität der Messinstrumente von Schwartz aus verschiedenen Perspektiven auseinandergesetzt. Der PVQ ist zwar das untersuchte Instrument der vorliegenden Arbeit, aber da der SVS die Grundlage dafür ist, muss auch seine Messqualität thematisiert werden. Auch die Zusammensetzung der Stichproben wird betrachtet, da hierauf die Schlüsse beruhen, die letztlich für gesamte Populationen gelten sollen.

Der Stand der Diskussion zeigt, welche Themen behandelt wurden, was die Ergebnisse sind und was an Themen noch aussteht. Damit spanne ich den Rahmen auf, in dem sich die vorliegende Arbeit auf die Diskussion bezieht. Die Beiträge werden zunächst thematisch zusammengefasst und dann chronologisch geordnet. Die Themen gliedern sich nach der Form der Analyse (deskriptive und hypothesentestende Verfahren), wichtigen Ergebnissen für die im ESS verwendete PVQ-Version (Davidov-Studien, Messqualität des PVQ) und der Kultur- und Kontextabhängigkeit der Wertemessung.

2.3.1 Deskriptive Erfassung der Wertestruktur

Schwartz (1992) lieferte neben einer detaillierten Beschreibung der Theorie empirische Befunde für die Messqualität des SVS. Die Skala wurde in 20 Ländern – übersetzt in die jeweilige Landessprache – anhand von 40 Stichproben getestet. Sie setzten sich zu einer überwiegenden Mehrheit aus Studierenden oder Lehrenden (33 von 40) zusammen, und in 27 von 40 Stichproben waren über 50 % der Befragten Frauen. Die Stichprobengrößen variierten zwischen N=542 und N=154. Das Durchschnittsalter über alle Stichproben hinweg beträgt rund 30 Jahre. Für 13 Länder wurden mindestens zwei Stichproben gezogen, die eine ähnliche sozioökonomische Herkunft aufweisen. In Hofstedes (1980) Untersuchung, die bereits in 53 Nationen durchgeführt worden war und Werthaltungen in diesen miteinander verglichen hatte, dienten nur die Mitarbeiter ein und derselben multinationalen Firma als Stichprobe. Demgegenüber stützt Schwartz seine Studie auf zwei Samples pro Land. Schwartz kritisiert an

Hofstedes Studie, dass sich die Mitarbeiter einer Firma signifikant von der Restbevölkerung im untersuchten Land unterscheiden könnten. Die Ergebnisse wären somit nicht mehr auf einen größeren Teil der Bevölkerung umzulegen. Er vertritt den Standpunkt, dass Lehrer die größte Repräsentativität aufwiesen. „They play an explicit role in a value socialization, they are presumably key carriers of culture, and they are probably close to the broad value consensus in a society“ (Schwartz, 1994b: 91). Hingegen scheinen ihm Studierende eher Träger von Modernisierungstrends zu sein (ebd.). Können innerhalb eines vorher definierten Raumes diese unterschiedlichen Stichproben ähnliche Ergebnisse erzielen, so kann eine Generalisierung unterstellt werden. Die Kreisstruktur wurde mittels SSA (*smallest space analysis*) gefunden. Erkenntnisse basierten auf einer visuellen Untersuchung und wurden nicht mit hypothesenprüfenden Verfahren getestet. Als Kriterien, um die Existenz eines Wertetypen zu bestätigen, wurde festgelegt, dass mindestens 70 % der Items in dem für sie bestimmten Segment liegen müssen. In zehn Stichproben (25 %) trat das nicht ein. Dennoch sieht Schwartz mit Blick auf die Zahl der verglichenen Länder das Instrument als universal genug an. Wie der nächste Beitrag zeigen wird, revidierte Schwartz dieses Urteil leicht, indem er von einer Auswahl an Werten ausgeht, die universale Bedeutung zu haben scheinen.

Die Datenanalysen lieferten keinen Hinweis darauf, dass noch Werte fehlen könnten. Dies wäre der Fall, wenn Segmente in der visuellen Darstellung leer blieben. Damit wäre ein essentielles motivationales Ziel nicht durch die Werte abgedeckt. Die zehn motivationalen Typen hingegen scheinen die Wertevorstellung von Menschen vollständig abzubilden. Schwartz spricht daher auch von einer *typology of basic value contents* (Schwartz, 1994a: 20). Dass es sich hierbei um eine erschöpfende Typologie handelt, begründet er damit, dass sich praktisch jeder kulturspezifische Wert einem der motivationalen Typen zuordnen lässt. Ein weiteres Indiz wäre, wenn sich nachweisen ließe, dass sich alle Wertetypen in unterschiedlichen Kulturkreisen wiederfinden ließen und sie voneinander unterschieden werden könnten (Schwartz, 1994a: 23). Der Beweis dafür steht noch aus, und wie in dieser Diskussion dargelegt wird, scheint neben positiven Erfolgen auch einiges dagegen zu sprechen.

Schwartz und Sagiv (1995) publizierten Ergebnisse zum SVS, die auf einer Datenbasis von bereits 88 Stichproben aus 40 Ländern stammten. Der Großteil der Stichproben waren wieder Lehrer und Studenten. In sieben von 88 Stichproben waren sie repräsentativ. 26 % aller Stichproben konnten alle zehn Wertetypen voneinander unterscheiden. Der Durchschnitt lag bei acht Wertetypen. Am seltensten in einem eigenen Segment wurden Werte von Conformity (in

60 % der Stichproben), Benevolence (in 64 % d. St.) und Stimulation (in 69 % d. St.) gefunden. In fast allen Fällen wurden sämtliche Werte zumindest in den gemäß Theorie ihnen benachbarten Segmenten aufgefunden. Starke Abweichungen von der postulierten Form könnten – sofern unreliable Messungen ausgeschlossen werden – laut den Autoren als Kulturspezifika interpretiert werden. Auf der Ebene der einzelnen Werte wurden 44 der 56 in 75 % der Stichproben in den für sie vorgesehenen Segmenten aufgefunden. „We conclude that most of these 44 values have a near-universal meaning“ (ebd.: 102). Sollte einer dieser Werte in zukünftigen Studien nicht in dem vorgesehenen Segment oder zumindest in einer benachbarten Region auftauchen, dann wäre das ein starker Hinweis auf eine signifikante Bedeutungsverschiebung in der untersuchten Stichprobe.

Auf der strukturellen Ebene der Werte, die sich in den zwei *higher-order dimensions* zeigt, konnte die Dimension *openness to change vs. conservation* in allen Fällen und *self-enhancement vs. self-transcendence* in 85 Stichproben reproduziert werden. Abweichungen sind abhängig vom größeren kulturellen Raum, aus dem die Stichprobe gezogen wurde. Die Autoren unterscheiden (beginnend mit der geringsten und abschließend mit der größten Zahl an Abweichungen) Kulturräume aus Westeuropa und englischsprachige Länder (als eigene Kategorie), Osteuropa, Lateinamerika, den Fernen Osten und schließlich den Nahen Osten. Es gibt allerdings zu jeder geographischen Region zumindest eine Stichprobe, die den modifizierten Quasizirkumplex gut wiedergibt. Können in einer Stichprobe wenige Werte unterschieden werden, und ist die Abweichung von der Theorie groß, bedeutet das entweder, (a) dass die Theorie nicht in dieser Stichprobe angewendet werden kann, oder (b) dass Probleme auf der Messebene vorliegen: Die Befragten haben die Aufgabe, die Verwendung der Antwortskala oder die Items nicht verstanden. Oder sie haben gar nicht die Motivation und Erfahrung, die Aufgabe zu bewältigen.

Um sicher zu gehen, dass Abweichungen nicht voreilig als kulturelle Differenzen interpretiert werden, stellen die Autoren Cut-off-Kriterien auf, um Kulturspezifika von unreliablen Messungen zu unterscheiden: In einer Test-Retest-Untersuchung wurde festgestellt, dass die Wertetypen relativ stabil gemessen werden, während sich auf der Ebene der einzelnen Werte durchaus Abweichungen ergaben. Die Autoren gehen von ca. 10 % der Werte aus, die nicht in den vorhergesehenen Regionen in Untersuchungen zu finden sein werden. Sie legen Forschern, die die Skala verwenden, nahe, mittels Split-Half-Methode zu überprüfen, ob Abweichungen sich innerhalb einer Stichprobe reproduzieren lassen oder auf Zufall beruhen. Ergeben sich in den Hälften ähnliche Muster, so ist von einer kulturellen Abweichung auszuge-

hen: „On these bases, we estimated that unreliability of measurement accounts for approximately two thirds of the deviations on average, and probably more in the less developed countries“ (ebd.: 113).

Im Kapitel 2.2 wurde erwähnt, dass der SVS für manche Befragte schwer zu beantworten war. Der PVQ ist ein leicht abgeändertes Instrument, das immer noch die selbe Struktur reproduzieren sollte, aber mit weniger Aufwand für die Befragten. In der Studie von 2001 testete die Gruppe um Schwartz in vier Stichproben Ähnlichkeiten der Wertestruktur und der Wertehierarchie zwischen dem neuen und dem alten Instrument. Zwei von vier Stichproben wiesen eine solche auf. Unterschiede führen die Autoren auf die höhere Komplexität des SVS zurück, der von gewissen Befragten nicht verstanden wurde. Die Autoren kommen zu dem Schluss, dass die Theorie der dynamischen Wertestruktur nicht vom Instrument abhängig ist. Erhebungen wurden in Südafrika (repräsentativ), Italien (repräsentativ) und Uganda (13- bis 14-jährige Mädchen) durchgeführt. Die Forscher testeten an israelischen Studierenden auch ein *multi-trait-multi-method*-Design (sowohl für SVS als auch PVQ). Indikatoren für diskriminante und konvergente Validität wurden gefunden. Vorausgesagte Zusammenhänge der motivationalen Typen mit ausgewählten Kovariaten wurden in den vier Samples bestätigt, und Konstruktvalidität wurde vorerst etabliert.

Die Autoren erkennen aber auch, dass die Ergebnisse den Universalitätsanspruch stark angreifen: „It is doubtful that the value content and structure identified by the theory are truly universal, but the theory may be more nearly universal than past findings with SVS have suggested“ (Schwartz et al., 2001: 538).

Schwartz bleibt in seiner Wertekonzeption nicht auf einer deskriptiven Ebene stehen, auf der Wertetypen unterschieden werden, sondern er unterstellt ihnen eine dynamische Organisation gemäß ihrer Vereinbarkeit oder Unvereinbarkeit untereinander. Bilsky und Koch (2000) stellen sich die Frage, wie sehr die *basic value types* tatsächlich universell sind. Es besteht die Möglichkeit, dass ihre spezielle Anordnung ein Artefakt ist, das auf Methodeneffekte zurückzuführen (also abhängig von der Art der Befragung) ist. Die Autoren testeten, ob sich die von Schwartz unterstellte zweidimensionale Struktur mit drei anderen Instrumenten replizieren lässt. Als Vergleichsinstrumente verwendeten sie „Kilmann Insight Test“ (Kilmann, 1975), „Ways to live“ (Morris, 1956) und „Personal Values Questionnaire“ (McClelland, 1991). Die

Werte der Skalen wurden den *higher-order values* von Schwartz zugewiesen, und in einer empirischen Studie mittels multidimensionaler Skalierung (MDS) ausgewertet. Das Ergebnis zeigte, dass es zwar zu leichten Unterschieden kam, diese aber nicht gravierend genug waren, um die zweidimensionale Struktur von Schwartz in Frage zu stellen. Vorausgesetzt, dass die Instrumente ähnliche Motivationen messen, kann die von Schwartz unterstellte Struktur mit anderen Instrumenten repliziert werden, was gegen die Hypothese eines beeinflussenden Methodeneffekts spricht.

2.3.2 Inferenzstatistische Tests

Spini (2003) verwendete in seiner Studie konfirmatorische Faktorenanalyse im Rahmenwerk eines multiplen Gruppenvergleichs (MGCFA). Damit ist diese Studie die erste, die die Theorie der *human values* von Schwartz mittels hypothesenprüfender Verfahren testete. Die Forschungsfrage kreiste *einerseits* um das Thema, ob sich die Wertestruktur auch im internationalen und -kulturellen Vergleich bewähre oder ob die Messungen in unterschiedlichen Populationen schlecht vergleichbare Konstrukte zu Tage bringen würden; und *andererseits*, ob die Zuordnung der Items zu ihren Konstrukten angemessen sei. Die Daten wurden in Studentenpopulationen (N=3.859) in 21 verschiedenen Ländern erhoben. Der SVS wurde in die jeweilige Landessprache übersetzt. In 16 von 21 Samples waren mehr als 50 % der Teilnehmenden Frauen.

Da bereits Schwartz (1992) feststellen musste, dass sich nicht alle Items gleichermaßen in allen untersuchten Ländern in die vorgesehene Ordnung einreihen ließen, besteht hinreichender Grund, weitere Tests mit der Skala durchzuführen. Spini überprüfte seine Hypothesen durch die Testung auf Äquivalenz (oder auch: Invarianz) der Messung[7]. Sie ist für die vergleichende Sozialforschung Voraussetzung, um überhaupt Unterschiede in verschiedenen Populationen sinnvoll interpretieren zu können. Ein angemessenes Vorgehen zur Testung legten Van de Vijver und Leung (1997) fest, auf das sich Spini bezieht. Mittels Strukturgleichungsmodellen mit latenten Variablen lässt sich die Hypothese testen, ob die zehn Wertetypen im internationalen Vergleich unidimensional sind und welche Items die besten Indikatoren für die Wertetypen darstellen. Spini testet in seinem Artikel: (a) konfigurale Invarianz (die Faktorenstruktur in den Ländern ist dieselbe), (b) metrische Invarianz (die Skalen werden in den Ländern gleich verwendet; Unterschiede der Itemscores können sinnvoll interpretiert werden;

7 Horn und MacArdle (1992) definieren Äquivalenz der Messung als „[…] whether or not, under different conditions of observing and studying a phenomenon, measurement operations yield measures of the same attribute […]“ (ebd.: 117).

Äquivalenz auf dieser Ebene erlaubt Vergleiche der strukturellen Beziehungen zwischen Items), (c) skalare Invarianz (Unterschiede der Mittelwerte in verschiedenen Ländern sind interpretierbar), (d) Faktorvarianzinvarianz (Faktorvarianzen werden über die Gruppen hinweg als gleich angenommen) und (e) Fehlervarianzinvarianz (zusätzlich zur Faktorvarianz werden auch die Fehlervarianzen gleichgesetzt; im Fall von Äquivalenz kann man davon ausgehen, dass die Items in den untersuchten Ländern gleich reliabel sind). (a), (b) und (c) sind ineinander geschachtelte Modelle (*nested models*) und daher mit Hilfe eines χ^2-Tests gegeneinander zu testen. Gleiches gilt für (a), (b), (d) und (e).

Die Ergebnisse der Studie zeigten folgendes Bild: Weder metrische noch konfigurale Äquivalenz konnte zwischen den Wertetypen gefunden werden, mit Ausnahme von Hedonism. Darüber hinaus gelten diese Aussagen nur für eine begrenzte Zahl von Items pro Wertetyp. Vergleichende Aussagen über Mittelwerte oder die Güte der Messung sind nur in einer Teilmenge von Ländern möglich. Spini stieß mit seiner konfirmatorischen Analyse auch auf eine andere Auswahl an Items pro Wertetyp als Schwartz (1994a), der in 75 % der 97 Stichproben 44 gleiche Items in den vorhergesagten Regionen vorfand. Dieses Beispiel zeigt, dass die Wahl der Analysemethode einen Einfluss auf das Ergebnis hat.

Schwartz und Boehnke (2004) testen die theoretischen Struktur des SVS anhand empirischer Daten mittels konfirmatorischer Faktorenanalyse. In ihrer Studie untersuchten sie 46 Stichproben aus 27 Ländern mit insgesamt N=10.857. Auch hier machten das Gros (40 Samples) der Teilnehmer Angehörige von Bildungsinstitutionen (Lehrer und Studenten) aus. Die Stichproben wurden per Zufall in zwei gleich große Gruppen aufgeteilt, und mittels konfirmatorischer Faktorenanalyse untersuchten die Autoren Hypothesen, die sich als ungeklärte Fragen aus vorhergehenden Untersuchungen ergeben hatten. Die Ergebnisse wurden durch die Gruppen kreuzvalidiert:

H1: Die zehn Werte liegen entlang eines motivationalen Kontinuums. Orthogonale Faktoren sollten sich schlechter bewähren als oblique, was bestätigt wurde. Die Varianzen frei zu schätzen ist zulässig.

H2: Die Lage von Tradition im selben Segment mit Conformity an einer Position signifikant an der Peripherie („modifizierter Quasizirkumplex“ (Schwartz und Boehnke, 2004: 235)) passt besser zu den Daten als ein reiner Quasizirkumplex mit zehn Werten und zehn Segmenten.

H3: Die Hypothese, Tradition und Conformity als einen gemeinsamen Wert und ihre Unterscheidung als hinfällig zu betrachten, konnte durch die unschlüssigen Testergebnisse weder bestätigt noch verworfen werden. Schwartz und Boehnke plädieren hingegen für eine Trennung, weil sich in zahlreichen Anwendungsstudien oft gezeigt hat, dass beide Wertetypen unterschiedlich mit anderen Konstrukten korrelierten, und man daher von einer diskriminanten Validität ausgehen kann.

H4: Power und Achievement liegen gemäß Daten nicht in einem gemeinsamen Segment. Sie scheinen demnach getrennte Konstrukte zu sein, obwohl beiden das Streben nach sozialem Ansehen zugrunde liegt.

H5: Die Daten lieferten Hinweise, dass eine Vier-Faktoren-Lösung durch die Darstellung mittels *higher-order values*, die prinzipiell sparsamer und daher vorzuziehen wäre, gegenüber einer Zehn-Faktoren-Lösung keine Verbesserung der Theorie darstellt. Daraus lässt sich ableiten, dass es sich bei der Wertestruktur um ein motivationales Kontinuum handelt, da die Einteilung in die von Schwartz vorgesehen *higher-order values* (*openness to change*, *self-transcendence*, *conservation* und *self-enhancement*) durch jede beliebige Zusammenlegung von benachbarten Wertetypen ersetzt werden könnte. Grenzen zwischen den Wertetypen sind rein gedankliche Stützen, um sich die Wertetypen besser vorstellen zu können. „In sum, the support for a motivational continuum of values gives researchers the freedom and flexibility to choose higher-order combinations of adjacent values particularly suited to the topics they study“ (Schwartz und Boehnke, 2004: 252).

H6: Hedonism liegt wie von Schwartz (1992) gefordert zwischen den Blöcken *openness to change* und *self-enhancement,* mit einer leicht stärkeren Korrelation mit Stimulation als mit Achievement.

H7: Universalism umfasst, wie die Datenlage zeigt, die Subdimensionen „Natur“ und „soziale Anliegen“.

H8: Bei Security spielen die Subdimensionen „persönliche Sicherheit“ und „Sicherheit der Gruppe“ im Verständnis der Befragten eine signifikant unterschiedliche Rolle.

Von den Autoren wurden offene Fragen für zukünftige Forschung aufgeworfen:

- Sind die übrigen Wertetypen in ihrem Gehalt tatsächlich eng genug umschrieben, oder gibt es andere wie Security und Universalism, die besonders breit in ihrem Inhalt sind?

- Gibt es generell signifikante Unterschiede in der Lage der Zentralität der Wertetypen zueinander? Wie wären die zu interpretieren, und ließen sie sich mit statistischen Tests überprüfen? Eine solche Unterscheidung könnte eine Bereicherung für die Wertetheorie darstellen.

Steinmetz et al. (k.A.) kritisieren an der Studie, dass die Autoren zwar die zirkuläre Struktur mit einer angemessenen Methode testen, aber gleichzeitig nicht dafür kontrollierten, dass sie ein Artefakt sein könnte: Zum einen wurden 46 Stichproben zusammengelegt, ohne kulturelle Differenzen zu kontrollieren. Andererseits wurde in der Studie auch nicht der mögliche Effekt der Methode des SVS auf die zirkuläre Struktur berücksichtigt.

Die Gruppe um Steinmetz (2009) lieferte einen weiteren Befund für die Einschätzung der Qualität der Werteskala nach Schwartz in der Version des PVQ. Spini (2003) hatte bereits für den SVS multiple Gruppenvergleiche mittels konfirmatorischer Faktorenanalyse zwischen Ländern angewendet und verschiedene Ebenen der Invarianz getestet. Steinmetz et al. wendeten eine ganz ähnliche Überlegung auf den Vergleich von Gruppen innerhalb eines Nationalstaates an. Denn auch diese benötigen für einen reliablen Vergleich Äquivalenz der Messungen (Vandenberg und Lance, 2000). Die Studie überprüfte die Hypothese, ob drei Gruppen, die sich in ihrer formalen Bildung unterscheiden, signifikant andere Anzahlen an validen Antworten erzeugten. Dahinter steht die Überlegung, dass Personen mit einer höheren formalen Bildung eher fähig wären, die Items korrekt zu lösen und ihre Werthaltungen wiederzugeben, weil sie besser an Testsituationen gewöhnt sind und sich mehr Gedanken bei der Beantwortung von Fragen machten (*test-wiseness*). Von der ursprünglich 40 Items umfassenden Skala wurden aus ökonomischen Zwängen 28 Items, die die Wertetypen am besten wiedergaben, ausgewählt. In einer Pilotstudie wurde auch getestet, ob es einen signifikanten Unterschied in der Beantwortung zwischen Items gebe, die im PVQ – wie von Schwartz vorgesehen – jeweils aus zwei Sätzen (n=69) oder aus einem einzigen Satz (n=68) bestehen. Die Unterschiede waren gering. Daher entschieden sich die Autoren für die einfachere Variante mit 28 Einzelsatz-Items. Auch die Antwortskala wurde verändert, indem statt einer Sechs-Punkte-Skala eine Vier-Punkte-Skala verwendet und in der Polung gegenüber der schwartzschen Anwendung umgedreht wurde. Pro Wertetyp wurden drei Items in die Skala aufgenommen. Ausnahmen sind Universalism und Tradition mit nur jeweils zwei Items pro Konstrukt.

Die Analyse zeigte, dass über die Gruppen hinweg volle metrische Invarianz, skalare Invarianz allerdings nur in der Mehrzahl der Wertetypen und dann nur partiell bestand. Fehlervarianzen konnten ebenso nur partiell gleichgesetzt werden. Weil sie aber nur für Security gänzlich unterschiedlich waren, gehen die Autoren davon aus, dass Fehlervarianzen prinzipiell als Reliabilitäten angesehen werden können. Das dynamische Modell von Schwartz wurde auch hier in den Daten wiedergefunden und unterschied sich in seiner Konfiguration auch nicht in den einzelnen Gruppen (konfigurale Invarianz).

Werden die Unterschiede zwischen den Gruppen in den Messfehlern betrachtet, so lässt sich die Hypothese, dass formal weniger gebildete Personen zufälligere Antworten geben als höher gebildete, widerlegen: „These results suggest that individuals with different levels of education differ less in the thought they devote to values than to political beliefs“ (Steinmetz et al., 2009: 612).

Hindernisse der Studie sehen die Autoren in der Beschränkung auf zwei bis drei Items pro Konstrukt, da schnell eine Grenze erreicht ist, wenn partielle Invarianz getestet werden soll.[8] Da bei einem Test auf Invarianz auch nicht ausgeschlossen werden kann, dass eine Drittvariable, die mit der Gruppenvariable korreliert, die Größe der Streuung beeinflusst, schlagen die Autoren vor, in der zukünftigen Forschung MIMIC-Modelle (multiple indicators multiple causes; Muthén, 1989) anzuwenden. Diese können solche Effekte kontrollieren, und wenn sie mit MGCFA kombiniert würden, auch statistische Tests für das Ausmaß der Invarianz liefern.

2.3.3 Davidov-Studien

Davidov, Schmidt und Schwartz (2008) prüften die ESS-Version des PVQ auf die Güte ihrer Invarianz in 20 der teilnehmenden Länder. Wie sich bereits in einer Teilstudie von Schmidt et al. (2007; siehe weiter unten) herausgestellt hatte, korrelierten jeweils drei benachbarte Wertetypenpaare so stark miteinander, dass sie zu einem gemeinsamen Faktor fusioniert werden mussten (UNBE, COTR, POAC). Im multiplen Gruppenvergleich zwischen den teilnehmenden Staaten mussten darüber hinaus Kreuzladungen (vgl. Schmidt et al., 2007) eingeführt werden:

- COTR lädt auf imprich (PO)
- COTR lädt auf iprspot (PO)

8 Für die Identifikation eines Messmodells braucht es mindestens zwei Items, die über die Gruppen hinweg gleichgesetzt sind.

- UNBE lädt auf imprich (PO)
- UNBE lädt auf ipadvnt (ST)
- POAC lädt auf ipmodst (TR)

Die Kreuzladungen ergaben sich zwischen Konstrukten und Items, die nicht direkt benachbart sind. Die Richtung der Ladung (das Vorzeichen) wird so verstanden, dass bei negativen Kreuzladungen der Wertetyp (latentes Konstrukt) nicht die gesamte gegensätzliche Spannung abdeckt. Hingegen zeigen positive Kreuzladungen an, dass die Gegensätzlichkeit durch das Konstrukt überbewertet wird und durch eine positive Ladung abgemildert werden muss. Da die Kreuzladungen nur von kombinierten Faktoren ausgehen, schließen die Autoren daraus, dass die Zusammenlegung eine Verzerrung der Schätzung hervorruft. Sie stellten auch fest, dass die ESS-Version zwar die motivationale Struktur wiedergibt, „[…] but the CFAs do not support all of the fine-tuned distinctions in the theory“ (Davidov et al., 2008: 16).

Zu den Ergebnissen sei gesagt, dass ein zumindest metrisch invariantes Modell in den 20 Ländern durch die Daten unterstützt wird. Volle skalare Invarianz konnte nicht nachgewiesen werden. Mittelwertvergleiche sind nur in speziellen Gruppen von Ländern möglich, in denen zumindest partielle Invarianz in ausgewählten Wertetypen festgestellt wurde.

Die Autoren waren nicht nur an der Möglichkeit interessiert, Mittelwerte zu vergleichen, sondern sie wollten auch wissen, ob die Wertestruktur – also die Stärke der Kovarianzen zwischen den Wertetypen – in den Ländern vergleichbar ist. Sie nennen das „das Muster der Distanzen“ („*pattern of distances*“). Dazu wurde die Annahme getroffen, dass die Kovarianzen zwischen den Wertetypen in allen Ländern gleich seien. Die Analyse ergab, dass sie es nicht sind. „This finding suggests that societal factors may influence the strength of opposition and/or congruence among the different values“ (Davidov et al., 2008: 20).

Im selben Jahr (2008) testete Davidov auf Äquivalenz über Länder und über die Zeit mit den Daten der zweiten Welle des ESS. Bereits auf konfiguraler Ebene zeigten sich erste Hürden: nur 11 von 20 Ländern wiesen dieselbe Konfiguration in beiden Runden auf, „[…] suggesting that the cross-country invariance may not be stable over time“ (Davidov, 2008: 41). In der Gruppe der Länder der zweiten Runde konnte in 14 Ländern metrische Invarianz festgestellt werden. „In spite of cultural differences, people appear to understand the meaning given to

the values by their indicators in a similar manner. Cognitive tests in each country could supplement these findings“ (ebd.: 43).

Im Zeitvergleich konnte sowohl metrische als auch skalare Invarianz in neun Ländern nachgewiesen werden[9]. In acht Ländern mussten bis zu drei benachbarte Wertetypen zusammengeführt werden. Der Autor relativiert das Ergebnis, da auch bei einem negativen Test auf Invarianz die zugrunde liegenden Konstrukte immer noch invariant sein können. Saris und Gallhofer (2007) hatten gezeigt, dass der Test auf metrische Invarianz zu strikt sein kann. Verzerrungen können u.a. durch unterschiedliche Reaktionen auf verwendete Methoden hervorgerufen worden sein. Auch hier resümiert Davidov, dass kognitive Interviews Einblick in die *kognitive Äquivalenz* bringen können.

2010 publizierte Davidov einen Artikel, in dem er die Invarianztestung auf die dritte Welle des ESS ausweitete. Die Messmodelle der vorhergehenden Artikel wurden beibehalten (sieben Wertetypen, fünf Kreuzladungen). Vergleiche mehrerer Länder (MGCFA) zeigten, dass in einigen Ländern noch weitere Wertetypen zusammengelegt werden mussten und die Referenzstruktur nur in acht Ländern repliziert werden konnte (diese unterscheiden sich aber nach der Landessprache). In diesen konnte zwar metrische Invarianz, aber weder volle, geschweige denn partielle skalare Äquivalenz festgestellt werden. Dieser Punkt weist darauf hin, dass nur in einzelnen, speziell vorher getesteten Subgruppen Mittelwertvergleiche durchgeführt werden können.

In 17 Ländern wurde erfolgreich auf konfigurale, metrische und skalare Invarianz getestet. In diesen Ländern können Mittelwerte über die Zeit verglichen werden. Sechs der Modelle kamen ohne Modifikationen aus. In den übrigen elf mussten zusätzliche Wertetypen zusammengelegt, Kreuzladungen eingeführt oder Fehlervarianzen frei geschätzt werden.[10] Aus diesen notwendigen Veränderungen schließt Davidov: „One may argue that it is not clear what the new meaning for each country actually is“ (Davidov, 2010: 185). Konvergente und diskriminante Validität fehlen bei manchen Wertetypen aufgrund von Kreuzladungen[11].

Wie Davidov selber feststellte, ist die statistische Testung zwar eine notwendige, aber nicht hinreichende Bedingung, um festzustellen, ob unterschiedliche Befragte dasselbe verstünden.

9 Das Referenzmodell war das von Davidov (2008): sieben Wertetypen (inkl. UNBE, COTR und POAC) und fünf Kreuzladungen

10 Für die genaue Dokumentation der Anpassungen vgl. den Originalartikel.

11 Signifikante Kreuzladungen weisen darauf hin, dass das Konstrukt nicht mehr nur durch die vorgesehenen Items, sondern auch durch solche gemessen wird, die ganz andere Konstrukte messen sollten.

Auch hier weist Davidov auf die Notwendigkeit von kognitiven Interviews hin, um mehr Sicherheit zu erhalten. Er befürwortet auch die Ergänzung der Skala um weitere Items, da die derzeitige Version – wie gezeigt – nicht in der Lage ist, alle zehn Wertetypen nach Schwartz zu unterscheiden.

2.3.4 Kultur- und Kontextabhängigkeit der Messung

Fontaine, Poortinga, Delbeke und Schwartz (2008) gingen der Frage nach, ob die Wertestruktur in unterschiedlichen kulturellen Gruppen gleich sei und ob die Wertekonstrukte gleich verstanden würden. Dazu grenzten sie Unterschiede, die durch die Stichprobenzusammensetzung entstanden, von jenen ab, die durch strukturelle/kulturelle Variablen hervorgerufen wurden. Sie analysierten den SVS in 38 Ländern mit jeweils zwei Stichproben (Lehrer und Studenten), die von Schwartz (1992; 1994a) erhoben worden waren. Wegen der geringen Stichprobengrößen und mangelnder Normalverteilung der Items, verwendeten sie deskriptive Verfahren (mehrere Varianten der MDS) statt konfirmatorischer Faktorenanalyse, um die Ähnlichkeit der Werteanordnung zu untersuchen. Die Ergebnisse zeigten systematische Unterschiede im Ländervergleich.[12] Kontextvariablen auf der Ebene der Vergleichseinheiten, die die strukturellen Unterschiede in den Ländern kontrollieren sollten, waren Bildung, Wirtschaftsleistung, Ausmaß der Arbeitsteilung, Lebenserwartung, Bevölkerungswachstum und Stand der Demokratisierung und wurden im Societal Development Index (SDI) für weitere Regressionsanalysen zusammengefasst. Mit ihnen sollte die Stärke des Einflusses der Kontextvariablen auf die Ausprägungen bei den Befragten geschätzt werden. Für die Analyse wurde zunächst ein Referenzmodell errechnet, das auf den aggregierten Daten über alle 38 Länder und Stichprobenarten (Studenten und Lehrer) hinweg basierte. Dieses glich im Großen und Ganzen dem der schwartzschen Theorie: zehn Wertetypen, angeordnet in einer kreisähnlichen Struktur entlang zweier Dimensionen (*higher-order dimensions*).

Mit Hilfe von Bootstraps (aus einer vorhandenen Stichprobe werden weitere Stichproben gezogen) konnte der Effekt der Stichprobenfluktuation geschätzt werden: „On average, random sampling fluctuation accounted for only about half of the structural differences between the overall configuration and the sample-specific configurations“ (Fontaine et al., 2008: 359). Das, was an unerklärter Varianz (*sample-specific*) übrig blieb, könnte auf kulturelle Unterschiede zurückzuführen sein. Werden die Items einzeln unter den Gruppen verglichen, so zeigen die Ergebnisse, dass es signifikante Unterschiede zwischen den Stichprobenarten (Stu-

12 Die Autoren nahmen an, dass die verschiedenen linguistischen und soziohistorischen Entwicklungen das Verständnis der Werte prägt.

denten oder Lehrer) und verschiedenen Graden des SDI gibt. D.h. Lehrer und Länder, die einen niedrigeren SDI aufweisen, zeigen größere Unterschiede zur Referenzstruktur. Besonders stark ist der Effekt von SDI in der Gruppe der Lehrer. Das bedeutet, dass Lehrer in Ländern mit niedrigerem SDI mehr Unterschiede zum Referenzmodell aufweisen als Lehrer in Ländern mit höherem SDI oder Studenten im selben Land. Werthaltungen von Studenten und besonders entwickelte Länder sind ihm ähnlicher. In der Diskussion bieten die Autoren fünf Gründe, die diese systematischen Effekte erklären können: (a) Staaten mit einem höheren SDI sind in den ausgewählten 38 Ländern überrepräsentiert. Das Referenzmodell wäre somit durch diese Auswahl verzerrt, und die Unterschiede der Länder mit niedrigerem SDI ergäben sich daraus. (b) Leichte Änderungen durch Übersetzungen und Unterschiede im Verständnis könnten besonders häufig zustande kommen, je stärker sich die Zielpopulationen von einer studentischen unterschiede, die an einer westlichen Universität studiert.[13] (c) Lehrer und Befragte aus Ländern mit niedrigem SDI haben geringere Erfahrung mit der Teilnahme an Erhebungen, was die Autoren als *lack of test-wiseness* bezeichnen. Daher geben sie ihre Werthaltungen ungenauer wieder. (d) Studenten und Befragte aus Ländern mit höherem SDI befinden sich in einem sozialen Kontext, der Autonomie fordert und fördert. Solche Erfahrungen spielen in der Wertetheorie eine entscheidende Rolle bei der Entwicklung eines deutlich erkennbaren Wertesystems mit klaren Abgrenzungen zwischen den Wertetypen und starken Konflikten zwischen ihnen: „The repeated experiences that particular values are congruent or in conflict lead to the crystallization of individuals' value structures […]" (ebd.: 361). (e) „[…], societies and social groups may actually differ somewhat in the underlying conflicts and congruencies that characterize their value systems" (ebd.: 360). Inwieweit jeder genannte Punkt für sich oder im Zusammenspiel mit anderen diese Variation bewirkt, lassen die Autoren offen. Allerdings schließen sie aus der gesamten Analyse und im Speziellen aus dem Ergebnis, dass SDI besonders unter den Lehrern für Varianz sorgt, dass es sinnvoll wäre, in zukünftigen Untersuchungen Stichproben zu verwenden, die sich von Lehrern und Studenten unterscheiden.

Sowohl Methodeneffekte als auch eigentümliche Bedeutungsunterschiede schließen die Autoren als mögliche Erklärungen für die systematischen Positionsunterschiede (in der MDS) der Items aus (ebd.: 361). Sie bevorzugen für ihre Erklärung sozialpsychologische Effekte.

13 Dieses Argument wird nicht weiter erklärt und macht ohne weitere Erläuterungen kaum Sinn. Eine mögliche Interpretation wäre, dass die Items durch Mitwirkung von Studenten entwickelt und an Studenten, die an westlichen Universitäten studiert haben, getestet worden waren. Wie die Autoren erklären, handelt es sich bei den Items um recht abstrakte Begriffe, die mit all ihren Konnotationen schwierig zu übersetzen sind. Je ferner die Zielpopulationen von dem Verständnishorizont der Studenten, desto häufiger die Abweichungen in der Übersetzung oder im Verständnis.

Stärker gemessene Wertekonflikte in Ländern mit höherem SDI und der Gruppe der Studenten werden auf den Grad der Einbettung (*embeddedness*) zurückgeführt. Auf die Messebene umgelegt würde das bedeuten, dass die Distanz der Items im zweidimensionalen Raum auf soziokulturelle Unterschiede zurückzuführen ist.

Davidov und De Beuckelaer (2010) prüften anhand der zweiten und dritten Runde des ESS, welchen Einfluss Übersetzungen in der Etablierung metrischer und skalarer Invarianz von international vergleichenden Studien haben. Tourangeau, Rips und Rasinski (2000) hatten festgestellt, dass Äquivalenz unwahrscheinlicher wäre, wenn Grundgesamtheiten mit unterschiedlichen Sprachen verglichen würden. Peytcheva (2008) fand Hinweise, dass das besonders auf solche Items zuträfe, die Einstellungen, Normen, Meinungen oder auch Werte messen, also einen stark affektiven Charakter aufweisen. Damit sollten auch die *human values* von diesem Effekt betroffen sein.

Die Autoren verglichen jene Länder bzw. Regionen, die innerhalb der Teilnehmerstaaten mit zumindest einer anderen Region eine gemeinsame Sprache teilten. Die Teilgruppen waren: England und Irland; Frankreich, französischsprachiges Belgien und französischsprachige Schweiz; Deutschland, Österreich und deutschsprachige Schweiz; die Niederlande und flämischsprachiges Belgien. Alle diese Länder zusammen ergaben Stichprobengrößen in ESS2: N=16.915 und in ESS3: N=16.992. Über alle Länder hinweg wurden MGCFA durchgeführt. Modifikationen wurden sowohl konventionell (mittels Modifikationsindizes) als auch mittels JRule eingeführt.[14] Die Tests zeigten, dass nach Einführung zusätzlicher Kreuzladungen, Fehlervarianzen und Zusammenführung von Wertetypen die Länder beider Runden metrisch partiell invariant sind (die Auswahl der Länder unterstützte in keiner Runde das von Davidov etablierte Modell mit sieben Faktoren und fünf Kreuzladungen). Weder volle noch partielle skalare Invarianz konnte für das Gesamtmodell entdeckt werden. Im Test des Maßes der Äquivalenz zwischen Gruppen mit denselben Sprachen (die Gruppen wurden paarweise miteinander verglichen) ergab sich hingegen, dass partielle skalare Invarianz sehr wohl gefunden wurde. Allerdings gilt dies nicht für sämtliche Wertetypen und für alle Länderpaare. Englisch- und niederländischsprachige Länder wiesen für alle Wertetypen partielle skalare Äquivalenz auf, französisch- und deutschsprachige fast nie für alle.

14 JRule ist ein Programm, das auch in der vorliegenden Arbeit verwendet wurde. Für eine genauere Erklärung siehe Saris, Satorra und van der Veld (2009) oder Kapitel 4.4.

Die Ergebnisse sprechen dafür, dass Übersetzungen, die für international vergleichende Studien notwendig sind, einen verzerrenden Einfluss haben, auch wenn so strenge Richtlinien wie im ESS gelten. Für kulturelle Einflussgrößen konnte trotz der Ergebnisse nicht vollständig kontrolliert werden. Es gibt kein absolutes Kriterium, das die in den Sprachgemeinschaften gefundenen Ähnlichkeiten vollständig auf die Sprache zurückführen könnte. Die übrigen Einflussfaktoren bezeichnen die Autoren als „*non-language-related aspects of culture*“ (ebd.: 503). Die von mir getestete Skala ist eine deutsche Übersetzung der ursprünglich auf Englisch verfassten. Spezifika, die durch die Analyse der kognitiven Interviews als auch der quantitativen Daten auftreten, können also zu einem Teil auf die Übersetzung der Items zurückzuführen sein.

Meuleman, Davidov, Schmidt und Billiet (2012) widmen sich der Fragestellung, wie sehr Populationen in Europa (Teilnehmerstaaten sind jene des ESS der ersten drei Wellen) in ihrer Positionierung im schwartzschen Motivationskontinuum eher durch landesspezifische Eigenheiten oder durch soziale Determinanten (Geschlecht, Alter, Einkommen und Bildung) bestimmt werden.

Hypothesen zu den einzelnen Hintergrundvariablen wurden aus der Literatur zusammengetragen und in ein Schema gegossen, das kennzeichnet, ob die jeweilige Variable auf den jeweiligen Wertetyp vermindernd oder verstärkend wirkt. Das Schema ist in Tabelle 2 wiedergegeben:

	UN	BE	CO	TR	SE	PO	AC	HE	ST	SD
Geschlecht	+	+	0	0	0	-	-	-	-	-
Alter	+/-	+/-	+	+	+	+/-	+/-	+/-	-	-
Einkommen	+/-	+/-	-	-	-	+/-	+/-	+	+	+
Bildung	+	+	-	-	-	+	+	+	+	+

+: positiver Effekt; -: negativer Effekt; 0: kein Effekt; +/-: widersprüchlicher Effekt

Tabelle 2: Antizipierte Effekte von Kovariaten auf zehn Wertetypen (aus: Meuleman et al. 2012)

Da sich Effekte durch die sozialen Hintergrundvariablen über die Zeit nicht verändern, wurden die Stichproben der einzelnen Länder in den verschiedenen Wellen verschmolzen und jene 25 Staaten ausgewählt, die zumindest an zwei Wellen teilgenommen hatten. Die Anzahl an unterscheidbaren Wertetypen schwankte in den verglichenen Staaten. Letztlich wurde eine Sechs-Faktoren-Lösung (UNBE, COTR, POAC und SDST wurden zusammengeführt) ge-

wählt, da sie in allen Ländern zu finden ist. Analysen zeigten, dass die Messungen zumindest partiell metrisch invariant sind.

Das Bestimmtheitsmaß (R^2) als Maß für den Anteil der durch die Hintergrundvariablen erklärten Varianz schwankte von Wertetyp zu Wertetyp. Die soziale Verortung spielte besonders bei jenen Werten eine Rolle, die zur Achse *openness to change vs. conservation* zählen. STSD und HE können zu knapp 20 %, UNBE und SE nur zu weniger als 10 % durch soziale Variablen erklärt werden. COTR und POAC liegen dazwischen. Die Autoren sehen die meisten Effekte der sozialen Hintergrundvariablen als über Länder hinweg relativ ähnlich an. Das scheint im Speziellen besonders für Alter und Geschlecht zu gelten. Sie folgern daraus: „[…] certain processes of value formation are cross-culturally robust." (ebd.: keine Seitenangabe)

Einige Unterschiede zwischen den Ländern schienen durch sozioökonomische Kontexte moderiert zu werden, so z.B. bei Geschlecht und Bildung. Die konservativere Werthaltung von Frauen in gewissen Ländern korrelierte stark negativ mit der Höhe des Zugangs von Frauen in diesen Ländern zum Arbeitsmarkt. Bei Bildung zeigte sich, dass sie in manchen Ländern einen positiven, in anderen einen negativen Effekt auf Hedonism hat. Die Autoren zeigten, dass der positive Effekt besonders in ärmeren Ländern zu finden sei. In ökonomisch stärker entwickelteren war der Effekt negativ.

Die Autoren fassten die Ergebnisse der in 18 von 25 Ländern gefundenen Effekte von Geschlecht, Alter und Bildung zusammen (siehe Tabelle 3). Die Effekte des Einkommens waren im Ländervergleich nicht konsistent und wurden daher ausgelassen.[15] In Klammer sind die aus Tabelle 2 abgeleiteten Zusammenhänge angeführt:

	UNBE	COTR	SE	POAC	HE	STSD
Geschlecht	+ (+)	n.r.E. (0)	+ (0)	- (-)	- (-)	- (-)
Alter	n.r.E. (+/-)	+ (+)	+ (+)	- (+/-)	- (+/-)	- (-)
Bildung	+ (+)	- (-)	- (-)	+ (+/-)	n.r.E. (+)	+ (+)

+: positiver Effekt; -: negativer Effekt; 0: kein Effekt; +/-: widersprüchlicher Effekt; n.r.E: nicht robuste Effekte

Tabelle 3: Tatsächliche Effekte von Kovariaten auf Wertetypen (aus: Meuleman 2012)

Die Zusammenfassung bietet auf einen Blick Aufschlüsse über Verknüpfungen sozialer Kontextvariablen mit den *human values*. Wie die Autoren aber selbst kritisch bemerken, ist die Frage der Kausalität nicht immer eindeutig geklärt: in sein Geschlecht und sein Alter wird

15 „Income may reflect life circumstances and chances als well als opportunity structures much more than a deeply anchored link to personal values" (ebd.: 21).

man hineingeboren. Bei Einkommen und Bildung (vor allem der tertiären) ist die Determiniertheit nicht mehr unbedingt gegeben. Der positive Effekt von Bildung auf POAC (gehört zum *higher-order value self-enhancement*) kann ebenso dadurch zustande gekommen sein, dass aufgrund einer verstärkten Disposition zu Selbstaufwertung man auch eher geneigt sein kann, höhere Bildung anzustreben.

2.3.5 Messqualität des PVQ

Mohler und Wohn (2005) überprüften die Wertestruktur mittels der ESS-Daten in 19 Ländern. Sie wendeten wie Schwartz ein visuelles Analyseverfahren an (multidimensionale Skalierung, kurz: MDS). Ihr Fazit: Die Skala zeigt Abweichungen von der theoretischen Struktur. Sie sind so groß, dass ihr die Autoren der ESS-Version die Universalität absprechen. Probleme bestehen darin, dass manche Wertetypen in anderen aufgehen oder Wertetypen ihre Plätze innerhalb des Quasizirkumplex vertauschen. Letzteres trifft sowohl auf Wertetypen zu, die in direkter Nachbarschaft zueinander stehen, als auch auf solche, die einander gegenüberliegen. Auf eine detaillierte Auflistung aller Probleme in den verschiedenen Ländern wird an dieser Stelle verzichtet, und es sollen nur die Ergebnisse für Österreich zusammengefasst werden: Stimulation und Hedonism liegen in einem gemeinsamen Segment anstatt in getrennten Segmenten. Die Segmente Security und Tradition/Conformity sind vertauscht (Tradition und Conformity stellen jeweils eigene Segmente dar). Power und Security liegen damit nicht mehr nebeneinander, sondern fast schon einander gegenüber. Dieses Ergebnis wiederum wurde auch in anderen Ländern gefunden und scheint auf eine messtechnische Schwäche der Skala hinzuweisen.

2007 prüften Schmidt et al. eine deutsche Version des PVQ. Die Forscher untersuchten, ob (1) sich die für universelle Werte postulierte Beziehungsstruktur auch mit dem PVQ zeigen lässt, (2) sowohl SVS als auch PVQ diskriminante und konvergente Validität besitzen und (3) Validität und Dimensionalität des PVQ mit der deutschen Stichprobe des ESS bestätigt werden können. Dazu wurden mehrere Studien durchgeführt. *Studie 1* zog ein Sample aus Studierenden einer deutschen Hochschule (N=321). Eine Analyse mittels MDS ergab zehn voneinander getrennte Segmente, und damit konnte die von Schwartz postulierte Struktur wiedergegeben werden. Einzige Ausnahme gegenüber der von Schwartz präferierten Form des modifizierten Quasizirkumplex war, dass die Wertetypen Tradition und Conformity nebeneinander statt hintereinander zum Liegen kamen. Die konfirmatorische Faktorenanalyse ergab eine

bessere Lösung mit neun statt zehn Werten (Tradition und Conformity korrelieren fast mit r=1). Die Skalen wiesen konvergente und diskriminante Validität auf, die unabhängig von der Messmethode (SVS oder PVQ) bestanden. Die *single-trait-multi-method*-Korrelationen[16] sind allerdings relativ niedrig (r_{max}=76, r_{min}=51), was streng betrachtet bedeutet, dass bei einigen Wertetypen nur knapp die Hälfte der Varianz durch die Verwendung der Methode (SVS oder PVQ) erklärt werden kann. Die übrigen 50 % sind „Rauschen". *Studie 2* übernahm die aus Studie 1 abgeleitete dimensionale Struktur und testete ihre Anpassung in einer repräsentativen Stichprobe (erste Welle ESS Deutschland; N=2.913). Dabei stellte sich heraus, dass eine 7-Faktoren-Lösung die beste Anpassung erzielte. Zusammengelegt wurden Universalism mit Benevolence (UNBE), Tradition mit Conformity (COTR) und Power mit Achievement (POAC). Des weiteren mussten Faktorladungen zugelassen werden, die so von der Theorie nicht vorgesehen waren:

- POAC lädt auf ipmodst (TR)
- UNBE lädt auf impdiff (ST)
- UNBE lädt auf ipcrtiv (ST)
- COTR lädt auf iprspot (PO)

Mit Hilfe abgeleiteter Zusammenhänge aus der Theorie zwischen den menschlichen Werten, Autoritarismus und umweltrelevanten Verhaltensweisen konnte Konstruktvalidität für die repräsentative Stichprobe festgestellt werden.

Studie 3 überprüfte die Test-Retest-Reliabilität an einer studentischen Stichprobe in einem Abstand von 43 Tagen (Nt=0=395; Nt=1=157). Sie lag für die einzelnen Skalen pro Wertetyp zwischen r=0,82 und r=0,62 und wurde unter Berücksichtigung der Itemanzahl pro Konstrukt (3 bis 6) als befriedigend von den Autoren bezeichnet.

Empfehlungen der Autoren sind: (a) PVQ und SVS können als „gleichwertige Alternativen angesehen werden" (Schmidt et al., 2007: 259); (b) die Beschränkung im ESS auf zwei bis drei Items pro Konstrukt für den PVQ kann zu Lösungsproblemen der Strukturgleichungen führen, da die Wertetypen sich nahezu gleichen und dadurch Kovarianzmatrizen produziert wurden, die positiv nicht definit sind (was in weiterer Folge zu Lösungsproblemen mittels konfirmatorischer Faktorenanalyse führt). Die Autoren empfehlen daher, mehr Items pro Messmodell zu verwenden, um der Bandbreite der Werteypen gerecht zu werden. Das würde wahrscheinlich benachbarte Wertetypen inhaltlich schärfer voneinander abgrenzen.

16 Campbell und Fiske (1959) definieren diese Diagonale als die *Validitätsdiagonale* (ebd.: 82).

Aufgrund der kumulierenden Hinweise auf problematische Ergebnisse mit der ESS-Version des PVQ vertieften sich Knoppen und Saris (2009a) in das Problem der Auswahl der Items für die Erhebung. Sie bezogen sich dabei auf den Artikel von Davidov et al. (2008), in dem benachbarte Wertetypen kombiniert worden waren (UNBE, POAC und COTR), da die sie messenden Items nicht mehr eigenständige Konstrukte erfassen konnten. Den Nachteil in der simplen Zusammenlegung sehen die Autoren darin, dass einerseits die kombinierten Wertetypen schwer inhaltlich zu interpretieren sind und andererseits Informationen verlorengehen. Nimmt man die Ergebnisse von Mohler und Wohn (2005) hinzu, die zeigen, dass selbst durch eine MDS die Struktur nicht wie theoretisch gefordert wiedergefunden werden kann[17], ist die Skala zu überdenken. Die Autoren sind der Meinung, dass das Problem durch eine andere Auswahl an Items gelöst werden könnte, ohne deren Anzahl zu verändern. Die Items sollten so ausgewählt werden, dass sie ihren Wertetyp homogener repräsentierten. Das hätte zur Folge, dass diese wünschenswerterweise zwar schwächer miteinander korrelieren würden, aber auch die inhaltliche Validität eingeschränkt wäre (genau das, was Schwartz vermeiden wollte; siehe Kapitel 3.1.7, Seite 59). Um Befunde für ihren Standpunkt zu liefern, analysierten sie die Daten von Schmidt et al. (2007), die die 40 Items lange PVQ-Version umfasste. Wertetypen dieser PVQ-Version wiesen viel schwächere Korrelationen als in der ESS-Version auf. Die ausgewählten Wertetypen wurden paarweise mit ihren direkten Nachbarn, so wie sie von der schwartzschen Theorie vorhergesagt wurden, in einem konfirmatorischen Faktorenmodell geprüft. Die Autoren konnten zeigen, dass die hohen Korrelationen zwischen den Wertetypen im ESS auf die spezifische Auswahl der Items zurückzuführen ist. Items, die ähnliche Werte messen sollten, korrelierten stärker mit Items benachbarter Wertetypen, weil die zugrunde liegenden Dimensionen enger verwoben sind. Für die weitere Verwendung der ESS-Skala skizzierten die Autoren zwei Szenarien: Entweder werden die von den Davidov-Studien propagierten Konfigurationen verwendet (vier reine und drei zusammengelegte Wertetypen) oder nur sieben reine Wertetypen, ohne die in der Analyse als problematisch identifizierten (Power, Tradition und Universalism). Die Autoren gelangten auch zu der Überzeugung, dass die Items für Universalism eher „*equality*“ (wurde von Schwartz als Teilmenge von Universalism definiert), und jene für Tradition eher „*humble*“ (Teilmenge von Tradition) messen würden.

17 Die Autoren merken an, dass diese unterschiedlichen Ergebnisse durch drei Varianten zustande gekommen sein könnten: (a) Die ESS-Version enthält substantielle Fehlerquellen, (b) der schwartzsche Quasizirkumplex basiert auf einer unrepräsentativen Stichprobe (Studenten und Lehrer), (c) Mohler und Wohn verwendeten keine zentrierten Daten, denn laut ESS-Angabe (http://ess.nsd.uib.no/ess/doc/ess1_human_values_scale.pdf) ist dies für MDS nicht notwendig. Die Autoren behaupten, Schwartz hätte dies bei seinen Analysen gemacht.

Knoppen und Saris (2009b) weiteten ihre Analysen auf die übrigen Wertetypen aus. An den Daten von Schmidt et al. (2007) testeten sie jeden Wertetyp paarweise mit seinen beiden Nachbarn auf mögliche Kreuzladungen und Subfaktoren (korrelierte Fehlervarianzen). Die durch JRule vorgeschlagenen Verbesserungen wurden auch auf ihre inhaltliche Sinnhaftigkeit hin bewertet.

Bereits eine konfirmatorische Faktorenanalyse nach schwartzscher Spezifikation lieferte sowohl einen signifikanten χ^2-Wert als auch in JRule mehr als doppelt so viele mögliche Verbesserungen, als von einem Modell erwartet werden kann, das richtig formuliert wäre. Das ist ein starker Hinweis darauf, dass die Modellspezifikation nicht durch die Daten gestützt werden kann. Die Wertetypen Tradition, Conformity, Security, Self-direction und Universalism vereinten in sich mehr als eine Dimension. Achievement und Benevolence konnten sowohl mit nur einem als auch mit zwei Faktoren dargestellt werden. Die Wertetypen Power, Hedonism und Stimulation konnten auf Grund beschränkter Itemzahlen nicht getestet werden. Diese – so sehen es die Autoren – werden durch jeweils sehr unterschiedliche Items gemessen, die auf unterschiedliche Dimensionen zurückgreifen. In mehreren Fällen luden die Items höher auf benachbarte Wertetypen als auf die vorgesehenen. Zusammengefasst lässt sich sagen, dass nur vier Wertetypen (Achievement, Hedonism, Stimulation und Benevolence) über Items richtig gemessen werden. In anderen Fällen gibt es hohe Kreuzladungen. Ein schlichtes Summieren der ESS-Bewertungen würde die Werte stark verzerren.

	korrekte Zuordnung	*Kreuzladungen*	*anderer Wertetyp/ alleine für sich*	*in einem PVQ-Subfaktor (ESS-Item korreliert mit einem PVQ-Item; gemeinsam anderer Subfaktor)*
TR	imptrad			ipmodst (being humble)
CO				„not to irritate other“ „obedience“
SE		impsafe (auf CO)		ipstrgv („safe country“)
PO	iprspot	imprich (auf AC)		
AC	ipshabt, ipsuces („showing success“)			
HE	impfun, ipgdtim			
ST	impdiff, ipadvnt			
SD	impfree		ipcrtiv („being creative“)	
UN		ipudrst (auf BE)	ipeqopt („equality“)	impenv („concern for nature“)
BE	iphlppl, iplylfr			

Tabelle 4: Ergebnisse des paarweisen Vergleichs (aus: Knoppen und Saris 2009b)

Die Autoren präsentierten in Tabelle 4 einen Vorschlag für die Änderung der ESS-Version, die 22 statt 21 Items umfasst, die 11 Werte messen (Universalism wurde in *equality* und *environment* aufgespalten). Die Items werden in diesem Vorschlag auch nur mehr durch einen Satz statt durch zwei Sätze wiedergegeben. Wie auch die Gruppe um Steinmetz (2009) feststellte, konnten Saris und Gallhofer (2002) in der Pilotstudie zum ESS für drei Items nachweisen, dass beide Versionen Gleiches messen, und selbst eine Änderung der Angabe, wie sehr der Befragte der porträtierten Person gleiche, konnte durch einfachere Aufgaben („how important a value was“ und „to what extend they liked to …“) ersetzt werden, ohne signifikante Einbußen in der Messung zu erleiden.

Die Forschergruppe um Beierlein (k.A.) nahm Kritikpunkte von Knoppen und Saris auf und testete deren Annahmen in zwei Studien. *Studie 1* untersuchte, ob die Ergebnisse sich unter Beibehaltung der Methode (PVQ mit 40 Items; paarweiser Vergleich der Wertetypen) in einer

repräsentativen Stichprobe in Deutschland replizieren ließen. *Studie 2* erhob unter Studenten Daten mit der neuesten in Entwicklung befindlichen Skala, dem PVQ5X, um ein vollständiges Modell mit 33 Items nach den Anleitungen von Knoppen und Saris mittels CFA zu testen.

Zwei Besonderheiten zeichnen diese Studie aus, die mit dem herkömmlichen Zugang zum Thema bricht: Erstens wurde bei der ersten Studie mit zentrierten Daten gerechnet, obwohl laut Angabe von Schwartz selbst dies für konfirmatorische Faktorenanalyse nicht nötig sei. Das ist meines Wissens nach die erste CFA für die *human values* nach Schwartz, die so durchgeführt wurde. Zweitens beinhaltet die zweite Studie in der CFA anstelle der Zentrierung einen Methodenfaktor, der die Skalenverwendung kontrollieren soll.

Zum Großteil konnten die Autoren mit ihren Ergebnissen in einer repräsentativen Stichprobe in Deutschland jene von Knoppen und Saris unterstützen. Vereinzelt traten Probleme bei der faktoriellen Struktur auf. Für vier Wertetypen wurden Subdimensionen gefunden:

- AC: „showing success“ und „ambition“
- SE: „national security“ und „personal security“
- UN: „protecting the evnironment“ und „societal concern“
- SD: „intellectual openness“ und „independence“

2.4 Zusammenfassung der Diskussion

Ergebnisse sind ambivalent bezogen auf die Messqualität und die Operationalisierung der Theorie. Im Bereich der international vergleichenden Erhebung ist zu beachten, dass die Faktoren SDI (Artefakt, weil Referenzmodell womöglich stark verzerrt ist), Übersetzung (*culture-specifics*) und die Auswahl der Variablen (Davidov; Schmidt et al.; Knoppen und Saris) einen Einfluss auf die Messung haben. Innerhalb einer Population hingegen scheint sich das Instrument zu bewähren (Steinmetz et al. 2008). Die ersten zwei Argumente sprechen gegen eine Universalität der Messung. Abgesehen von kulturellen Eigenheiten spielt womöglich auch der Kontext (vgl. Konty, 2002, Hechter et al., 1999) eine wichtige Bedeutung in der Entwicklung der Werte.

Die Validität und Reliabilität des Instruments besteht in Folgendem:

1. 50 % der Varianz ist „Rauschen“ (Validität).

2. Die Ergebnisse sind besonders auf eine bestimmte Teilbevölkerung hin validiert (*test-wiseness*; *culture-specific embeddedness*).

Bei Anpassungen bestehender Modelle an die Daten ist ein Vorgehen, das bestimmte Kreuzladungen zulässt, arbiträr. Ohne diese scheint das Modell jedoch nicht gut genug zu den empirischen Daten zu passen. Die Frage ist also: Soll die Theorie an die Fakten angepasst werden, oder finden sich genügend Hinweise, dass die Qualität der Messung und weniger die Theorie selbst das Problem darstellen? Es fehlt eine systematische Prüfung der Dimension der kognitiven Prozesse während des Beantwortungsprozesses für die Werte-Items des ESS in Österreich. Sie könnte die Antwort auf die Frage liefern. Darüber hinaus sollte der Anpassung von Modellen mit bisher verwendeten Methoden nachgegangen werden, um anschlussfähige Ergebnisse zu produzieren, die eine Reflexion der Methoden an sich möglich macht.

3 Methode

Im folgenden Kapitel werden zunächst methodische Probleme angesprochen, die auftreten, wenn es um die Messung von Werten geht. Diese Felder ergaben sich durch die Analyse der kognitiven Interviews und der quantitativen Daten.

Da die verwendeten methodischen Ansätze in der vorliegenden Verwendung nicht getrennt, sondern integriert werden, wird auf das spezielle Mixed-Methods-Design eingegangen. In einem weiteren Schritt werden die einzelnen Methoden dargelegt, mit denen der Hauptteil der Analyse durchgeführt wurde: konfirmatorische Faktorenanalyse und kognitive Interviews.

3.1 Methodologische Probleme

3.1.1 Der Zugriff auf Werte – *values as truisms*

Wie kann man sich vorstellen, dass Befragte den wahren Wert eines Wertes wiedergeben? Eine Antwort auf diese Frage liefert den notwendigen Kontext, um auch nachvollziehen zu können, was theoretisch beim Prozess der Messung geschieht. In der Forschung zur Anwendung in der Survey-Methodologie ist bis dato keine vollständige Theorie bekannt, die spezifisch das Antwortverhalten auf Wertefragen und dessen Zustandekommen untersucht.

Zur Erhebung von Einstellungen gibt es hingegen Theorien. Da Einstellungen Werten nahestehen, ließen sich womöglich Parallelen bzgl. ihrer Repräsentanz ziehen. Ein älteres Modell, das die kognitive Struktur von Einstellungen erläutert, ist das sogenannte Aktenschrankmodell (*file-drawer model*) (Tourangeau, et al., 2000). Die Anhänger dieses Konzepts gehen davon aus, dass Befragte auf eine geordnete Sammlung bereits vorhandener Einstellungen zu gewissen Themen zurückgreifen, die – sehr vereinfacht gesprochen – in beschrifteten Laden im Gedächtnis vorhanden sind. Der Stimulus (im Fall einer Befragung ist das das Item) zeigt dem Befragten, in welcher der Laden er oder sie nachsehen soll. Da empirische Befunde gegen diese Theorie sprechen, bevorzugen Tourangeau und Kollegen (2000) hingegen die Vorstellung, dass die Wiedergabe von Einstellungen auf ein Survey-Item gemäß eines *belief-sampling*-Modells erfolgt: Mehrere Quellen, die strategischen Überlegungen unterliegen, speisen die Antwort auf eine Einstellungsfrage. Die Auswahl möglicher Quellen, die die Basis der strategischen Überlegungen darstellen, ist nach diesem Modell nicht immer gleich,

sondern unterliegt einer bestimmten Wahrscheinlichkeit, wie Befragte – abhängig vom Kontext – Erfahrungen aus der Vergangenheit ins Bewusstsein holen. Die jeweilige Erinnerungsstruktur, die dabei zu Tage tritt, entspricht der Einstellung. Sie ist ein Netz aus Verbindungen. Wie gut auf den gesamten Zusammenhang zurückgegriffen werden kann, hängt damit vom Reiz und von der Vordergründigkeit des Netzes ab. Das bestechende an dem Ansatz ist, dass Einstellungen (und womöglich auch Werte) dadurch sowohl relativ konkret vorliegen können, wenn sich Leute diese Einstellung sehr oft bewusst machen, als auch sehr vage sein können. Diese Sichtweise lässt gleichermaßen Raum für Veränderung wie für Verfestigung zu. Denn Werte sind nicht in Stein gemeißelt, sondern werden u.a. auch von Schwartz als veränderbare Gebilde aufgefasst. Wertestrukturen haben sich nicht ein für alle Mal in einem frühen Entwicklungsstadium festgelegt, sondern sind konstanter Veränderung unterworfen: „People evolve value priorities that cope simultaneously with their basic needs and with the opportunities and barriers, with the ideas of what is legitimate or forbidden, in their environment." (Schwartz, 2003: 263) Ändern sich die Umstände, führt das auch zu einer Änderung der Wertehaltungen. Schwartz geht auch davon aus, dass soziale Erlebnisse prägend sind und unterschiedliche Menschen in ihren Werten verbinden, wie das z.B. durch die gleiche Stellung in der Gesellschaft vermittelt wird. Begrenzt wird die Veränderbarkeit von Werten, also das, was an dem Individuum sozialisiert wird und ihm vermittelt werden kann, von zwei intrinsischen Voraussetzungen: Dies sind die individuellen Bedürfnisse und das Temperament.

Die Forschergruppe um Bernard (2003) widmete sich der Frage, was Werte überhaupt darstellen bzw. welche Darstellung sie in unserer Persönlichkeit erfahren. Zunächst halten die Autoren fest, dass Menschen üblicherweise über Werte nicht bewusst nachdenken. Menschen gehen davon aus, dass andere Personen ihre Werte teilen und sehen deshalb wenig Veranlassung, sich Argumente dafür oder dagegen einfallen zu lassen. Die *values-as-truisms*-Hypothese, die die Autoren vertreten, sieht Werte besonders durch ihre affektive Dimension geprägt und durch starke Gefühle und vergangene Erfahrungen in uns verankert. Kognitive Argumente scheinen hingegen wenig zu prägen. *Truisms* sind direkt übersetzt *Binsenweisheiten* oder *Gemeinplätze* – also Begriffe, über die nicht tiefgehend reflektiert wird. Entgegen dieser Theorie wurden Werte besonders in den populäreren Skalen (die Autoren zitieren Inglehart und Schwartz) als bewusste Ziele operationalisiert.

Gestützt auf Ergebnisse aus Studien zu Einstellungsrepräsentationen von der Gruppe um Wilson (1989) hatten Maio und Olson (1998) deren Erkenntnisse auf Werte umgelegt. Die Essenz der Analyse ist, dass sich die Messung von Einstellungen/Werten durch Reflexion der

Gründe dafür oder dagegen ändert, wenn Einstellung hauptsächlich über Affekte in den Köpfen bestehen („[…] if attitudes have strong, amorphous affective component with few supporting cognitions (Bernard et al., 2003: 3). Die Autoren erklären dieses Phänomen damit, dass Personen mit wenig bewussten und reflektierten Gründen zufällige nennen müssen, wenn sie danach gefragt werden. Die spiegeln aber ihre Einstellungen/Werte nicht ganz wider und/oder widersprechen ihnen womöglich. Das bewegt Personen letztlich auch, ihre Einstellung/ihren Wert in der genannten Ausprägung anzupassen. Diese Anpassung ist ein Hinweis darauf, dass es den Einstellungen/Werten an kognitiver Unterstützung (die Autoren sprechen von *cognitive support*) mangelt. Hingegen berichteten die Befragten in Maios und Olsons (1998) Untersuchung, mit den erhobenen Werten starke Affekte zu verbinden. Die motivationale Stärke von Werten erhalten diese also weniger durch kognitive Gründe als durch Affekte. Das macht auch einen wichtigen Unterschied zwischen ihnen und Einstellungen aus. Denn in der motivationalen Hierarchie von Menschen nehmen Werte eine zentralere Position ein als Einstellungen. Wenn ihnen die kognitive Basis fehlt, hat das andere Auswirkungen als bei Einstellungen.

Die Studie der Gruppe um Bernard (2003) ist eine Erweiterung der Studie von Maio und Olson (1998), weil damals nur se*lf-transcendence*-Werte des SVS mit der Hypothese der v*alues-as-truisms* getestet wurden und die neue Studie den gesamten SVS und die Materialismus-Postmaterialismus-Skala von Inglehart miteinbezieht. Die Hypothese bestätigte sich sowohl bei der Skala von Schwartz als auch bei der Skala von Inglehart.

Die Autoren leiten aus der Studie die Einsicht ab, die aus meiner Sicht relevant für die Erhebung von menschlichen Werten ist: Eine Überprüfung der eigenen Ursachen für Werte ändert deren Beurteilung; allerdings sind jene neuen Urteile von minderer Qualität und sagen nur schlecht daraus abgeleitetes Verhalten voraus (Bernard et al., 2003: 21). Die Autoren führen das darauf zurück, dass sich Personen üblicherweise nicht für die Bewertung von Werten rechtfertigen müssen. Im Gegenteil: Eine Rechtfertigung bewirkt, dass sich die Befragten von der eigentlichen Basis – nämlich den Affekten – für Werte trennen und sich in der Beurteilung auf relativ gesehen irrelevante Aspekte konzentrieren. Wenn wir diesen Gedanken weiterspinnen, dann müssten Messungen, die sich stärker auf die kognitive, bewusste Ebene beziehen, schlechtere Vorhersagen für Verhalten sein, als jene, die stärker den affektiven Aspekt berücksichtigen. Weiters bedeutet das auch für die Verwendung von kognitiven Interviews, dass sie möglicherweise kein besonders gutes Mittel darstellen, die Validität von Wertemessungen zu prüfen, sofern sie nicht auch die affektive Komponente berücksichtigen. Ansonsten

bilden sie lediglich den relativ kleinen Bereich ab, der üblicherweise nicht in die Beurteilung einfließt. Die Bewertungen, die während eines solchen Interviews abgegeben werden, sind somit auch stark verzerrt und nicht unbedingt aussagekräftig.

3.1.2 Skalierung von Werten

Werte sind wie die meisten sozialwissenschaftlichen Konstrukte so komplex und mehrdimensional, dass sie nicht durch einen einzelnen Indikator alleine erfasst werden können. Um ein Messinstrument zu entwickeln, in dem alle Indikatoren dasselbe Konstrukt messen, verwenden Forscher sogenannte Skalierungsverfahren, deren Endprodukt eine Skala ist. Diese besteht laut Holm (1970: 356) aus einer Menge an Items, die entlang einer Dimension messen. Items werden bei Schnell et al. (2008) folgendermaßen definiert: „Items sind in der Regel Fragen bzw. Aussagen (Statements), denen die Befragten zustimmen oder die die Befragten ablehnen sollen“[18] (ebd.: 179).

Ob dem Item eher zugestimmt oder ob es eher abgelehnt wird, ist durch die Schwierigkeit der Items festgelegt. In ihrer ursprünglichen Verwendung versteht man darunter den Anteil der richtig gelösten Items. Die Bezeichnung stammt aus der Entwicklung von psychologischen Tests und wurde für die Erstellung von Einstellungsfragen übernommen. In der Survey-Forschung werden solche Items als schwierig definiert, die von dem oder der Befragten abgelehnt werden (also nicht gelöst werden). „Ein schweres Item wird also von weniger Personen richtig gelöst als ein leichtes Item“ (Schnell et al., 2008: 183). Eine Person wird einer Frage zustimmen, wenn für sie die Aussage zutrifft (oder anders formuliert: wenn die im Item formulierte Aussage ihrer tatsächlichen Einstellung entspricht). Im Fall von Skalen sind die Aussagen die Einstellungen zu den latenten Konstrukten. Die Theorie besagt, dass der Grad der Zustimmung zu diesen Konstrukten der Ausprägung der Einstellung entsprechen sollte. Daraus ergibt sich pro Item eine eigene Verteilung, die sogenannte Itemcharakteristik. Sie beschreibt in Abhängigkeit von der Ausprägung in der latenten Variable die Wahrscheinlichkeit der Zustimmung. Typisch für gebräuchliche Items in den Sozialwissenschaften ist eine S-Form mit monotonem Anstieg. In einer speziellen Form entspricht sie der Verteilung einer normalverteilten Variable: Der Anteil der Zustimmungen oder Ablehnungen an den äußersten Ausprägungen ist selten (die Kurve daher flacher) und nimmt in den mittleren Kategorien zu (die Kurve ist steiler).[19]

18 Im vorliegenden Fall der Skala von Schwartz braucht es eine kleine Adaption der Definition, da sie zwar mehrere Dimensionen, wohl aber nur ein Kontinuum der Motivation abbildet.

19 Die genauen Spezifizierungen des Grades der Steigung bei normalverteilten Variablen wird an dieser Stelle ausgelassen. Das Beispiel soll lediglich anschaulichen Zwecken dienen.

Shalom Schwartz entschied sich bei der Konstruktion seiner Skala likertskalierte Items zu verwenden. Eine Likertskala wird auch als „*Methode der summierten Ratings*“ bezeichnet, weil die Konstruktion prinzipiell erlauben sollte, die Bewertungen der einzelnen Items zu einem Index zusammenzufassen, der die darunterliegende latente Variable wiedergibt. Daher sollten die Items auch nur eine Dimension messen.

Rensis Likert, der Namensgeber und Entwickler der Methode, bevorzugte ein fünfteiliges Antwortschema:

Stimme stark zu	Stimme zu	Teils, teils	Lehne ab	Lehne stark ab
1	2	3	4	5

Die Antwortmöglichkeiten der *human values* in der verwendeten Version sparen hingegen bewusst die mittlere Kategorie aus und zwingen somit die Befragten, sich zu positionieren. Die mittlere Kategorie ist Streitpunkt in der Survey-Forschung, weil zum einen manche Befragte sich bewusst einer Mittelposition zuschreiben wollen, aber andererseits einige Forschungsergebnisse darauf hinweisen, dass sie ebenso als „weiß nicht“-Kategorie verwendet wird (Porst, 2011: 81). Auch messen die Items inhärent mehr als nur eine eindeutige, latente Variable aufgrund des motivationalen Kontinuums, das Werte abbilden. Wie wir sehen werden, ist ihre Itemcharakteristik auch so beschaffen, dass sie eine überdurschnittlich hohe Zustimmung aufweisen, also relativ leicht[20] zu sein scheinen und damit linksschief verteilt sind. Das liegt daran, dass latente Konstrukte gemessen werden, die *per definitionem* wünschenswert sind. Üblicherweise werden für Skalenkonstruktionen nur Items ausgewählt, die gut zwischen den Gruppen unterscheiden können, also eine relativ hohe Trennschärfe besitzen. Schwartz empfiehlt, die Schiefe seiner Items zu korrigieren, indem die Itemwerte zentriert werden (für Details siehe Kapitel 4.2, Seite 77).

Eine weitere Eigenart likertskalierter Items ist, dass ihr Messniveau in der angewandten Forschung üblicherweise als intervallskaliert angesehen und verwendet wird. Es kursieren diesbezüglich im World Wide Web und in der *scientific community* Gerüchte über ungeschriebene Regeln, dass eine solche Skala ab fünf Kategorien intervallskalenähnliche Eigenschaften aufweist. Der Grund für diese Daumenregeln liegt auf der Hand: viele leistungsfähige statistische Auswertungsverfahren benötigen dieses Niveau, um Tests überhaupt durchzuführen. Intervallskalen sind so definiert, dass die Abstände zwischen den Intervallen gleich groß sind. Diese Bedingung ist allerdings nicht automatisch gegeben. Es kann durchaus sein, dass bei einer verbalisierten Skala (das bedeutet, dass die einzelnen Kategorien neben ihnen zugewie-

20 Im Englischen wird dafür der Begriff „ipsatizing“ verwendet. Er beschreibt die Bewertung eines Objekts aufgrund der bisherigen Leistungen desselben.

senen Zahlen auch ausformuliert sind), wie es die ESS-Version des PVQ ist, die Abstände uneinheitlich von Respondent zu Respondent interpretiert werden. Damit soll gesagt sein, dass die Annahme der gleichgroßen Intervalle nicht automatisch zutrifft, auch wenn die augenscheinliche Betrachtung eine solche suggeriert.

Bisherige Analysen der PVQ-Items der ESS-Version rechneten mit Verfahren (z.B. ML bei CFA), die an sich ein Intervallskalenniveau und eine Normalverteilung in den Daten voraussetzen würden. Simulationsstudien hatten gezeigt, dass unter gewissen Annahmen bei likertskalierten Items Analysen generell nicht zu falschen Schlüssen führen. Zumindest scheint das Verfahren für die Testung von Invarianzen über Gruppen hinweg angemessen zu sein (siehe Welkenhuysen-Gybels, 2004). Hingegen haben Lubke und Muthén (2004) diese Annahme kritisiert.

Davidov et al. (2011) nahmen die Kritik auf, dass bisherige Untersuchungen mit den PVQ-Items des ESS aus den genannten Gründen zu falschen Ergebnissen hätten führen können. Sie testeten die Skala auf Invarianz mittels Schätzverfahren, die für ordinalskalierte Daten angemessen sind (hierzu haben Muthén et al. (1997) geeignete Methoden entwickelt)[21]. Es stellte sich heraus, dass die Ergebnisse keine Änderungen der Modellannahmen bringen. Aus diesem Grund wird für die vorliegende Arbeit die herkömmliche Schätzung mittels MLR (das ist die gegen Verletzungen der Normalverteilungsannahme robuste Version des ML-Schätzers; siehe Kapitel 3.3, Seite 64) verwendet.

3.1.3 Doppelte Stimuli (*double-barreledness*)

Enthält ein Item mehr als einen Stimulus, aber nur eine Möglichkeit darauf zu antworten, führt das notwendigerweise zu einem Konflikt. Die Wahrscheinlichkeit, dass zwei Stimuli vermittelt werden, ist hoch, wenn das Item aus zwei Sätzen besteht. Die Items der ESS-Version des PVQ bestehen jeweils aus einem Satz, der die Wichtigkeit eines Wertes betont, und aus einem, der aussagt, was eine Person gerne tut. Theoretisch sollten diese beiden Teile dasselbe messen (siehe Abbildung 2). Das CCT stellte diese Annahme in Frage (Schwartz, 2003) und führte eine Pilotstudie durch. Das Ziel war, herauszufinden, ob am Beispiel von drei Items substantielle Unterschiede zwischen den Teilsätzen bestehen. Dazu wurde ein *multiple-trait-multiple-method-* (MTMM-) Design angewendet. Das CCT kam zu dem Schluss, dass bei den drei ausgewählten Items die zwei Sätze dasselbe messen. Unter dieser Voraussetzung spricht rein messtechnisch nichts dagegen, einen Einzelsatz oder eine Kombination zu ver-

21 Mplus bietet Schätzverfahren wie WLS, WLSM, WLSMV an, die diese Kriterien erfüllen.

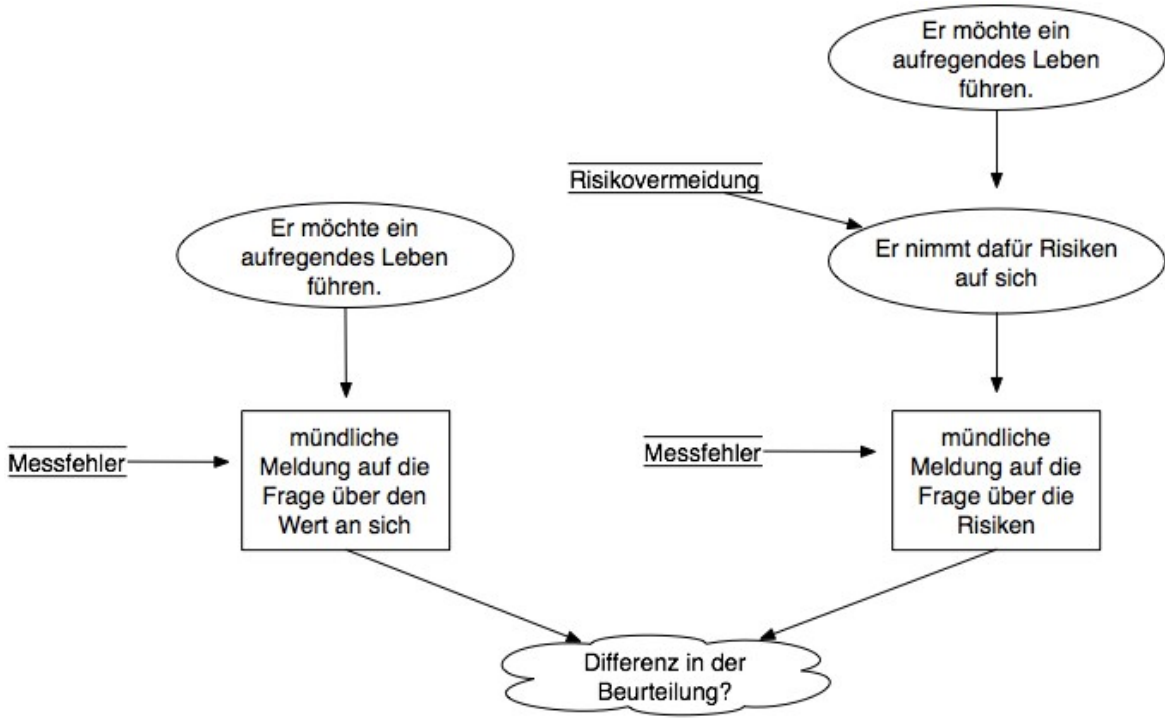


Abbildung 2: Doppelter Stimulus (aus: Schwartz, 2003)

wenden. Schwartz (ebd.: 298) bevorzugt die kombinierte Variante, weil sich in Studien gezeigt hatte, dass die Befragten die Porträts in Form von Einzelsätzen als zu wenig informativ kritisiert hatten. Es wurden auch nur Formulierungen ausgewählt, die kognitiv vorgetestet worden waren. Wie die kognitiven Interviews der vorliegenden Arbeit allerdings zeigen, ist trotz der Studie des CCT das Problem der doppelten Stimuli immer noch aktuell.

3.1.4 Werte: normative Ideale oder persönliche Prioritäten

Geben die Befragten ihre eigene Meinung wieder oder reproduzieren sie Prioritäten, die sie von sozialen Normen gesteuert als wünschenswert erleben? Schwartz selbst gibt zu, dass Werte kulturelle Ideale repräsentieren und Befragte daher geneigt sein könnten, eine wünschenswerte Hierarchie dementsprechend zu formulieren: „Nonetheless, because values represent cultural ideals, people may be inclined to provide the value priorities they perceive as desirable“ (Schwartz, 1992: 50).

Schwartz gibt zwei empirische Hinweise darauf, dass es sich bei den erhobenen Daten um persönliche Werte handelt:

1. Eine ausgiebige Anzahl an Befragten antwortete außerhalb der Norm (also des Durchschnitts aller Befragten). Die meisten Items wiesen Ausprägungen über das gesamte Antwortspektrum aus. Kein Wert wurde mit weniger als vier Kategorien beantwortet.

2. Wären die Antworten von kulturellen Normen geprägt und die individuellen Antworten Ausreißer (also individuelle Missinterpretationen der Normen), dann dürften Moderatoren (also die persönlichen, nicht durch kulturelle Normen geprägten Determinanten wie Alter und Geschlecht) keine Aussagekraft haben. Die Untersuchung der Daten wies allerdings auf das genaue Gegenteil hin.

Daraus lässt sich für Schwartz auch ableiten, dass von individuellen Einstellungen gegenüber Werten nicht darauf geschlossen werden kann, wie die in der Gesellschaft herrschenden Erwartungen diesbezüglich aussehen (Schwartz, 1992: 51).

Wie noch im empirischen Teil gezeigt wird, kann es in der Messung solcher Werthaltungen aufgrund der dualen Definition von Werten (nämlich der persönlichen und der als in einer Gesellschaft vorherrschenden Meinung) zu Interpretationsspielräumen kommen, was die Messung sehr unpräzise werden lässt.

3.1.5 Messung von Kontext?

Neben der Frage, wessen Werte abgebildet werden, gibt es das Problem des Zugriffs auf Werte. „*In reality, people may not always know what their values are […]*“ (Hitlin und Piliavin, 2004: 367). Hitlin und Piliavin beziehen sich dabei auf Hechter (1993). Im öffentlichen Diskurs kommt das Bedürfnis nach Werten nicht zu kurz, doch nur selten wissen die Verfechter oder ihre Anhänger, was sie konkret darunter verstehen. Diese Themen treten aber üblicherweise in Kontexten auf, die sehr konkret sind. Forscher hingegen hegen ein starkes Interesse an solchen Werten, die besonders abstrakt sind und wenig Bezug zum Leben ihrer Forschungsobjekte haben. Der Mangel an Kontext bereitet in der Folge Schwierigkeiten bei der Messung. Schwartz (1992) stellte fest, dass die Messung von Werten unterschiedlich ausfällt, wenn Kontext in die Betrachtung einfließt. Er ist der Meinung, dass „seine“ Werte transsituational seien und als solche auch in den Befragten vorlägen und erfassbar wären. Es fehlt allerdings eine schlüssige Theorie, wie sich ganz generell Werte formieren und im Speziellen, wie sich transsituationale von kontextualen unterscheiden, um diese Behauptung aufrechtzuerhalten. Seligman und Katz (1996) fanden Hinweise darauf, dass die Messung von Werten stark von der Situation abhängt, in der sich die Befragten mental befinden. Am Beispiel von zwei Situationen, die den Wert menschlichen Lebens thematisierten – Abtreibung und Todesstrafe – stellten die Forscher mittels der Rokeach-Studie einen Unterschied in der Bewertung fest. Und das, obwohl in beiden Fällen Wert menschlichen Lebens zu bewerten war. „[…] within

an individual, values may be organized differently for the two issues […]“ (Seligman und Katz, 1996: 54). Sie schreiben weiters:

> […] that individuals may have encapsulated, modular or multiple value-systems associated with different issues. Our basic idea is that individuals may construct value-systems in the context of some specific issues and not simply apply some general value system to issues as they arise. This perspective challenges the traditional view of the value system as a single ordered set of values that is important to the self-concept and helps guide thought and action. For the traditional value system to serve as an effective guide, it is assumed that it transcends situations. Our view sees value systems as dynamic rather than static and creatively applied in situations rather than rule-bound (ebd.: 55).

Alleine über Werte zu reflektieren bedeutet bei manchen Befragten bereits, dass sich die Werteausprägung verändern kann:

> Values are often simply truisms (Maio und Olson 1998), endorsed but lacking argumentative support; those with stronger cognitive reasons supporting their values report less change when forced to introspect about those values (Hitlin und Piliavin 2004: 379).

Einige Forscher bevorzugen deshalb eine Messung, die kontextsensibel ist (vgl. Konty, 2002). Schwartz geht in seiner Konzeption davon aus, dass seine Messung transsituational ist. Die kognitiven Interviews fördern diesbezüglich ambivalente Ergebnisse zu Tage.

3.1.6 Soziale Erwünschtheit

Human values – also Werte – sind in ihrer Operationalisierung einem offensichtlichen Vorwurf ausgesetzt: Sie sind als etwas Erstrebenswertes formuliert. Der Schluss liegt nahe, dass erwünschtes Antwortverhalten einen starken Einfluss auf die Beantwortung von Werte-Items haben könnte. Paulhus (2002) definiert sozial erwünschte Beantwortung (engl. *social desirable responding*, kurz: SDR) als „the tendency to give positive self-descriptions“ (ebd.: 49). Das heißt, Befragte versuchen sich eher in einem guten Licht darzustellen als eine ehrliche Antwort auf eine Frage zu geben. Wie die Gruppe um Schwartz (Schwartz, Verkasalo, Antonovsky und Sagiv, 1997) feststellten, scheint es sich bei Werten um Konzepte zu handeln, die von SDR relativ verschont bleiben. Warum sie der Meinung sind, soll dieses Kapitel beleuchten.

Es gibt verschiedene Ansätze, wie das Ausmaß von SDR erfasst werden kann. In der theoretischen Debatte erscheinen drei größere Strömungen der Operationalisierungen (nach Paulhus, 2002):

- *Minimalist constructs*: Sie umfassen die Messungen der *self-deception scales* von Edwards (1957) (kurz: SD-scale) und Wiggins (1959) (kurz: Sd-scale).
- *Elaborate constructs*: darunter fallen die MMPI Lie Scale von Hathaway und McKinley (1951), die SDR-Skala von Crowne und Marlowe (1964; kurz: MC) und *personality inventory* von Eysenck und Eysenck (1964), Sackheim und Gur's SDR-scale (1978).
- *Accuracy constructs*: Ihre Proponenten gehen von der Richtigkeit der Aussagen von Befragten aus. Diese geben keine übermäßig positive Darstellung ab, sondern weisen die Persönlichkeitsmerkmale, die sie über sich selbst angeben, auch tatsächlich auf. Ohne ins Detail zu gehen, gibt es mehrere Hinweise darauf, dass diese Posi-tion nur schwer haltbar ist (vgl. Paulhus, 2002).

In dieser Auflistung soll hervorgehoben werden, dass die untersuchten psychologischen Konstrukte immer Persönlichkeitsmerkmale, also *traits*, waren. Das sollte bei der weiteren Analyse immer im Hinterkopf behalten werden. Ergebnisse und Schlüsse müssen für Werte immer erst angepasst werden.

Es bestehen zwei konkurrierende Ansichten darüber, was SDR theoretisch darstellt und wie Messungen davon verzerrt werden können: *Einerseits* wird SDR als Antwortstil und *andererseits* als Persönlichkeitsmerkmal gesehen. In der derzeitigen wissenschaftlichen Debatte ist noch kein definitives Urteil gefallen, um welches von beiden es sich tatsächlich handelt. Vielmehr stehen beide Ansichten als konkurrierende Thesen zueinander. Tan und Grace (2008) geben einen schlüssigen Überblick über die vorherrschenden theoretischen Strömungen zu dem Thema. Sie sind der Meinung, dass besonders die Hypothese, bei Werten handle es sich um Persönlichkeitsmerkmale, aufgrund stichhaltigerer Hinweise zu bevorzugen ist. Das bedeutet, dass die Tendenz, sozial erwünscht zu antworten, in der Persönlichkeit verankert liegt.

Paulhus hat mit seiner Arbeit zu SDR einen erheblichen Einfluss auf die Debatte. Auch er bevorzugt (Paulhus, 2002) die Hypothese eines Persönlichkeitsmerkmals und führt dafür den Beweis, indem er seine SDR-Skala (*Balanced Inventory of Desirable Responding*, kurz: BIDR) mit den Big 5 (Costa und McCrae, 1989) korreliert (seine *self-deception*-Faktoren korrelieren zumindest mit $r > 0,3$; *impression management* zumindest mit $r>0,1$). Die Frage, warum SDR sich nicht als eigener Faktor bei Persönlichkeitsuntersuchungen herauskristallisiert, erklärt er damit, dass „[…] BIDR response styles are simply disguised measures of nor-

mal personality.*"* (Paulhus, 2002: 58). Dennoch ist eine Unterscheidung zu treffen: zwischen einem reinen Persönlichkeitsmerkmal und der Neigung, Antworten zu verzerren. Den dafür treffenden Begriff sieht Paulhus in *departure from reality*, also dem Teil der Abweichung von Selbsteinschätzungen zu einem Thema, der durch glaubwürdige Persönlichkeitsmessungen herausgerechnet werden konnte. Alles, was darüber an Einschätzung hinaus geht, ist die Überschätzung des Befragten. Das ist die Antwortverzerrung (engl. *response bias*), die Paulhus als „[…] any systematic tendency to answer questionnaire items on some basis that interferes with accurate self-reports" (ebd.: 49) definiert. *Departure from reality* ist für Paulhus das Kriterium schlechthin: es schafft Klarheit darüber, ob – wie das die *accuracy concepts* feststellen – Personen mit einem hohen SDR-Index diejenigen sind, die diese Merkmale auch tatsächlich aufweisen. *Departure from reality* unterscheidet *response bias* von tatsächlichen Persönlichkeitsmerkmalen.

Aufbauend auf Wiggins (1991) Zwei-Faktoren-Modell erstellt Paulhus eine neue Theorie über die Dimensionen, welche SDR umfasst. Dabei stützt er sich sowohl auf die Arbeiten von Damarin und Messick (1965) und Sackheim und Gur (1978) zur grundlegenden Aufspaltung von SDR als teilweise Selbst- und teilweise Fremdtäuschung als auch auf Wiggins zwei Verzerrungsfaktoren „Alpha" und „Gamma". Alpha entspricht Wiggins' Wert *agency,* der auch p*ower, mastery* und *assertion* beinhaltet. Gamma korreliert mit dem Wert *communion* mit seinen Unterkategorien *intimacy, union* und *solidarity*. Paulhus definiert diese beiden Faktoren als „*constellation of biases and traits"* (Paulhus, 2002: 63), die durch die Werte *agency* und *communion* begründet werden. Das scheint besonders erwähnenswert, weil diese Übereinstimmung von Werten mit SDR weitreichende Auswirkungen auf die Messung von Werten haben könnte. Das ist womöglich auch der Grund, warum Schwartz et al. (1997; siehe Seite 56) einen Zusammenhang zwischen SVS und SDR fanden. Leider scheint Paulhus diesen Artikel nicht zu kennen, und er nimmt keinen Bezug darauf.

Zusammen mit der Idee, dass SDR sowohl eine Komponente Selbsttäuschung als auch eine Komponente Fremdtäuschung umfasst und diese Komponenten auch zu bestimmten Teilen mehr oder weniger bewusst eingesetzt werden, entwickelt Paulhus (2002) sein zweistufiges Konzept:

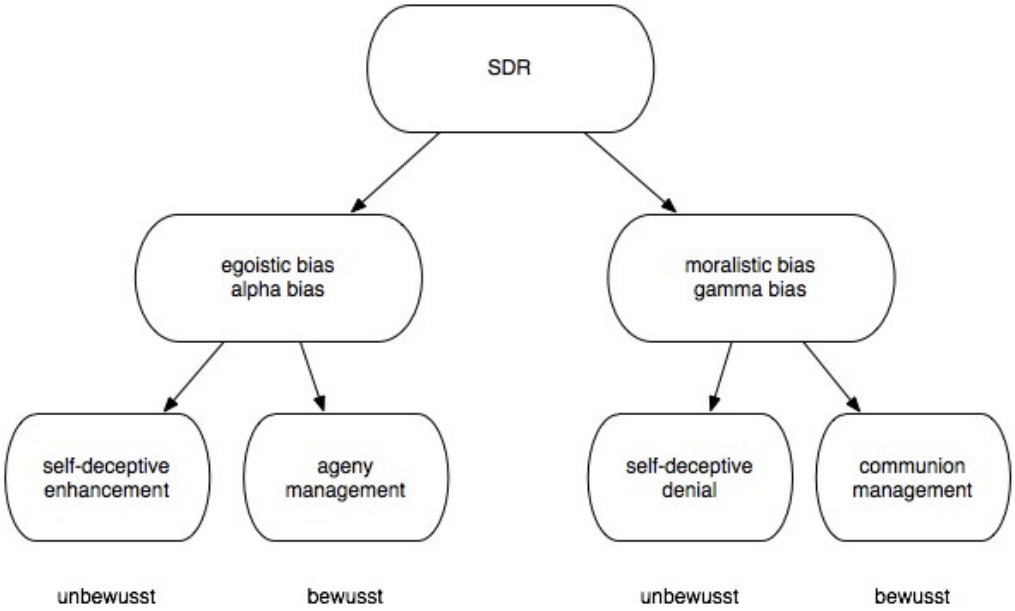


Abbildung 3: SDR nach Paulhus (aus: Paulhus, 2002)

Sowohl *egoistic bias* als auch *moralistic bias* lassen sich zu einem Teil durch Selbsttäuschung und zum anderen Teil durch Fremdtäuschung erklären. Weil die *impression-management*-Faktoren (*communion management, agency management*), also die kontrollierte Wirkung nach außen, besonders vielen Einflussgrößen unterliegen, die von Situation zu Situation verschieden sind, hält Paulhus diese eher für *response sets*, also Gruppierungen von Antwortarten, die im Gegensatz zu *response styles* sehr leicht veränderbar und situationsabhängig sind. Diese Nomenklatur ist leider etwas verwirrend, weil der Begriff „*style*“ bei Paulhus mehrdeutig verwendet wird: Einmal ist es der Antwortstil, der für jedes Individuum mehr oder weniger einzigartig ist; dann wieder ist *style* bei ihm auch die Unterscheidung in *self-deceptive* und *impression management*, also die Art, wie sich eine Verzerrung ausdrückt, und wer dabei getäuscht wird. Die Gruppe um Schwartz (1997) verwenden ihn wiederum in einer eigenen Weise.

Schwartz et al. (1997) haben sich dem Problem SDR beim SVS gewidmet. In ihrem Artikel beziehen sie ihre theoretischen Überlegungen auf den „alten Paulhus“ (D. L. Paulhus, 1991), der noch ein Modell des BIDR mit einer einzigen Ebene und drei Faktoren vertritt (*self-deceptive enhancement, self-deceptive denial, impression management*). Ausgespart war damals noch der Zusammenhang von SDR mit Wiggins’ Werten und die Zerlegung von *impression management* in zwei eigenständige Faktoren. Schwartz et al. beziehen sich in ihrer Arbeit nur auf den Faktor *impression management*, weil er den Teil der Verzerrung misst, der zielgerichtet angewandt wird und ihnen am wichtigsten erscheint. Das entsprechende SDR-Messinstrument ist MC (Marlowe-Crown-Skala).

Sie nehmen die von Tan und Grace später formulierte Gegenüberstellung von SDR sowohl als kontrollierbaren Faktor als auch als Teil der Persönlichkeit auf und testen die Hypothesen gegeneinander anhand des SVS. Schwartz et al. hatten in vorherigen Forschungsarbeiten anderer Forscher festgestellt, dass die bisher davon ausgegangen sind, dass es sich bei SDR um ein kontrollierbares Verzerrungsphänomen handle, was Schwartz et al. als *style hypothesis* bezeichnen[22]. Feather (1973), Kelly et al. (1972) und Goldsmith et al. (1987) testeten den Rokeach Value Survey mittels unterschiedlicher – damals zur Verfügung stehender – SDR-Messungen. Keine der drei Forschergruppen fand signifikante Zusammenhänge zwischen den beiden Konstrukten. Der Rokeach Value Survey schien immun gegenüber SDR zu sein. Was sie nicht berücksichtigt hatten, so Schwartz et al., war die Möglichkeit, dass SDR Teil der Persönlichkeit sei (Schwartz et al. bezeichnen diesen Zusammenhang als *substantive hypothesis*). Mit der dynamischen Konfiguration des SVS von einander ent- und widersprechenden Werten sollte es möglich sein, diese Hypothese zu testen. Die Struktur sagt Konflikte von Werten vorher, die überprüft werden können. Schwartz et al. verwendeten dazu die MC-Skala und den SVS. Wie bereits weiter oben erwähnt, weisen weder ihre Vorgänger noch Schwartz et al. (1997) auf den Zusammenhang zwischen Wiggins' Faktoren Alpha und Gamma und den daraus notwendigerweise resultierenden Korrelationen mit Wertetypen hin.

Paulhus (2002) verortet die MC-Skala zum größten Teil auf dem Gamma-Faktor (*moralistic bias*). D.h. die Arbeit von Schwartz und seinen Kollegen deckt zwar sowohl *self-deception* zu einem kleinen Teil als auch *impression management* und sowohl bewusste als auch unbewusste Vorgänge ab (die Autoren selbst sprechen von der Einschränkung auf die Messebene von *impression management*), wobei aber offen bleibt, wie gut MC die Dimension der Selbstüberhöhung erfasst. Der Gamma-Faktor (entsprechender Wert ist *communion)* beschreibt die Leugnung von sozial abweichendem Verhalten. Bei Werten handelt es sich aber um Konzepte, die *per definitionem* erwünscht sind[23]. Damit müssten eigentlich Skalen angewendet werden, die vor allem *egoistic bias* messen (laut. Paulhus wären das die Skalen *narcisisstic personality inventory* für *self-deception enhancement* und *agency management scale* für *agency management*). Es fehlt daher eine wichtige Dimension, um die Auswirkung von SDR auf Werte vollständig zu beschreiben. Tan und Grace (2008) stellten in ihrer Metaanalyse fest, dass nach neuerer Forschung SDR mindestens durch drei Faktoren abgebildet werden muss, um ein vollständiges Bild zu erhalten.

22 Wie erwähnt führt die Benützung des *style*-Begriffs zu Verwirrung.

23 In bestimmten Kulturen gelten manche Werte (z.B. Power oder Tradition) als nicht besonders wünschenswert (Schwartz et al., 1997: 10, Fussnote Nr. 2).

Als Datenbasis befragten Schwartz et al. zwei Stichproben:

a) N=207, israelische Erwachsene. Die Zusammensetzung ist nach Erwerbsarbeit, Geschlecht, Alter und Schulbildung relativ heterogen zusammengesetzt.

b) N=131, finnische Sozialarbeit-Studenten. Diese Stichprobe war relativ homogen.

Die Studie ergab:

- *Stylistic hypothesis*: Korrelationen zwischen der mittleren Wichtigkeit zu allen Items des SVS und MC betrugen im Durchschnitt etwa r=0,22. Damit sind sie positiv und signifikant, wenn auch nicht sehr groß. Das bedeutet, dass Befragte mit hohem SDR-Wert generell alle Werte höher bewerteten. Die Autoren testeten auch, ob Befragte besonders jene Werte höher einstuften, von denen sie annehmen konnten, dass sie in ihrem kulturellen Kreis als besonders erstrebenswert angesehen werden. Dazu wurde überprüft, ob jene Werte, die im jeweiligen Sample als die relativ wichtigsten eingestuft wurden, auch die höchsten Korrelationen aufwiesen. Diese Hypothese konnte in keiner Stichprobe bestätigt werden. Ein Argument gegen die Richtigkeit des Ergebnisses wäre, dass die Befragten nicht unbedingt ähnliche Vorstellungen vom wichtigsten Wert ihrer primären Vergleichsgruppe haben müssen. Das mag auf die heterogene israelische Stichprobe zutreffen, bei der finnischen ist es aber nicht so wahrscheinlich. Da dieser Effekt allerdings nicht explizit getestet wurde, ist die Hypothese als noch nicht ganz verworfen anzusehen.

Der Einfluss vom verzerrenden Effekt von SDR auf die Bewertung von Werten geben die Autoren mit 3 bis 7 % an. Das ist vernachlässigbar.

- *Substantive hypothesis:* Theoretisch sagten die Autoren die Reihenfolge für jene Werte vorher, die von besonders positiv bis besonders wenig positiv (also evtl. negativ) mit SDR korrelieren würden. Diese war: Conformity, Tradition, Security, Benevolence, Universalism, Power, Achievement, Self-direction, Stimulation, Hedonism. In der überwiegenden Mehrheit der Fälle war die Reihenfolge wie vorhergesagt und signifikant. Es gab eine Abweichung in der finnischen Stichprobe: Die Korrelation zwischen Achievement und MC war um einen Platz verschoben.

Mängel der Studie: Wie bereits erwähnt, basieren die Ergebnisse nicht auf dem letzten Stand der Forschung. Die Marlowe-Crown-Skala erfasst nur einen Teil von dem, was SDR alles

umfasst. Ein anderer Punkt ist, dass die Studie zwar anonym durchgeführt wurde, d.h. die Teilnehmer sahen keine Veranlassung, sich in einem besonders guten Licht zu präsentieren. Das könnte sich aber schlagartig ändern, wenn die Motivation, sich vorteilhaft darzustellen, wächst, etwa bei einem Einstellungsgespräch. Die Autoren erkennen diese Schwachstelle und wollen an ihr durch weitere Forschung arbeiten. Bis dato konnte jedoch keine diesbezügliche neue Studie ausfindig gemacht werden. Die Überlegung wirft allerdings einen kritischen Punkt auf, der eine grundsätzliche Basis der schwartzschen Wertekonzeption berührt: Wenn die Autoren nicht sicher sind, ob verschiedene Situationen zu systematisch unterschiedlichem Antwortverhalten führen, dann ist die Behauptung, dass es sich bei Werten um transsituationale Konzepte handelt, nicht sehr solide.

Es scheint auch ein logisches Problem dabei zu geben, den Anteil von SDR an der Beantwortung von Werten zu erfassen, wenn die Werte selber Teil der Konzeption von SDR sind. Es wäre eine tautologische Erklärung. Diese Studie erweitert zwar die konvergente Validität sowohl des SVS als auch von MC, allerdings fehlt die Feststellung, von welchen Konzepten sich die Instrumente unterscheiden lassen, wenn sie schon so angelegt sind, dass sie dasselbe messen.

3.1.7 Niedrige Reliabilitäten

Ein anderes Problem betrifft die niedrige Reliabilität der verwendeten Items – sowohl in der Vollversion als auch in der ESS-Version. Schwartz ist sich dieses Mangels bewusst und begründet ihn damit, dass die Wertetypen ein breites Spektrum an Inhalten abdecken: „It is impossible to attain high reliability for all ten values, because we are trying to cover a range of content of the full motivational continuum of values with only ten types of values […]“ (Schwartz, 2003: 303). Bei der Auswahl der Items wurden solche aufgenommen, die die größtmögliche Breite der Wertetypen widerspiegeln. Dadurch korrelieren die Items eines Wertetyps notwendigerweise schwächer miteinenander, wenn sie verschiedene Dimensionen des Konstrukts abbilden (Schwartz, 2003: 277). Es gab einen Zielkonflikt zwischen Reliabilität und Inhaltsvalidität. Die Entscheidung fiel zugunsten der Inhaltsvalidität aus. Schwartz empfiehlt daher die Feststellung der Reliabilität mittels der *higher-order values*. Sie werden durch mehr Items repräsentiert. Zu ihrer generellen Verwendung hingegen meint er: „However, working only with them rather than the ten values when relating values to other variables loses a great deal of information […]“ (Schwartz, 2003: 303).

Dieser Überblick über methodologische Probleme mit vielen Unwegbarkeiten soll eine Begründung dafür sein, warum es für die vorliegende Arbeit sinnvoll ist, sich dem Thema mit mehreren Methoden und über gängige Grenzen zwischen methodischen Paradigmen hinweg zu nähern. Im folgenden Kapitel gehe ich darauf ein.

3.2 Methodenmix

Ich kombinierte die Methoden so, dass zuerst die qualitativen Interviews geführt und transkribiert wurden. Dann analysierte ich die quantitativen Daten und unterzog parallel dazu die Interviewprotokolle einer ersten Systematisierung. Die quantitative Analyse bestimmte jene Problemfelder, die anschließend sowohl durch die Interviewprotokolle als auch durch statistische Maße in Kombination erklärt wurden.

3.2.1 Warum überhaupt MM?

Laut Teddlie und Tashakkori (2009) sprechen generell drei Gründe für ein Mixed-Methods-Design:

- Die gezogenen Schlüsse sind bedeutungsvoller.
- Gleichzeitig können sowohl exploratorische als auch konfirmatorische Schlüsse gezogen werden.
- Eine größere Breite voneinander abweichender Perspektiven entsteht, die neuerlicher Ausgangspunkt für weitere Forschung sein könnten.

Darüber hinaus gibt es andere Gründe, die spezifisch die vorliegende Arbeit betreffen. Zum einen ist der Entstehungszusammenhang zu nennen: Diese Arbeit wurde zur Erreichung eines Diploms der Soziologie verfasst. Das Erkenntnisinteresse war dadurch immanent durch einen soziologischen Kontext geprägt. Warum ein Mixed-Methods-Design gewählt wurde, lässt sich anschaulich durch ein Zitat Max Webers zusammenfassen: Er definierte die Soziologie als „[…] eine Wissenschaft, welche soziales Handeln deutend verstehen und dadurch in seinem Ablauf und seinen Wirkungen ursächlich erklären will“ (Weber, 1980: 1). Auf die Integration konfirmatorischer Faktorenanalyse und kognitiver Interviews umgelegt bedeutet dies, von quantitativen Ergebnissen als (zwar nicht unbedingt soziales aber durchaus) sinnhaftes Handeln auszugehen und den dahinterliegenden Sinn in seinem Ablauf und seiner Wirkung mit Hilfe der kognitiven Protokolle ursächlich zu erklären. Den subjektiven Sinn zu verstehen passt in das Bild, das gewisse Forschergruppen der Surveyforschung vom Zweck kognitiver

Interviews haben. Sie argumentieren, dass die reichhaltigen Protokolle den Sinn der erhobenen Antworten (in Form von Zahlen) verstehen helfen. Darin seien sie effektiver als darin, einen Beitrag zur Verbesserung der Validität von Items zu leisten (Yan et al., 2010; Willis, 2005: 109-113).

Eine Brücke zwischen Validität, kognitiven Interviews und der Bedeutung von Skalenwerten[24] hat hingegen Messick (1995) mit seinem Konzept der *comprehensiveness of construct validity* geschlagen. Darin geht er davon aus, dass Konstruktvalidität besonders unter zwei Einflussfaktoren zu leiden hat: *construct underrepresentation* und *construct-irrelvenat variance* (ebd.: 742). Damit bringt er zum Ausdruck, dass Messungen einerseits wichtige Aspekte zur vollständigen Aufklärung des zugrundeliegenden Phänomens fehlen können und dass sie andererseits zu breit angelegt sein und redundante, dafür aber reliable Information erfasst können. Um diesen Problemen zu begegnen, erweitert er Konstruktvalidität auf sechs Dimensionen. Zwei davon haben für die vorliegende Arbeit besondere Bedeutung: *Zum einen* die inhaltliche Vollständigkeit (die üblicherweise als eigenständige Dimension der Validität angesehen wird) und *zum anderen* das, was Messick als „substantive theories, process models, and process engagement“ (ebd.) definiert. Unter *substantiveness* fasst er Theorien und Anweisungen zusammen, die darüber Auskunft geben, wie der Prozess für die Beantwortung eines Items aussieht. Messick erläutert zwei Punkte, die dabei ins Spiel kommen:

> One is the need for tasks providing appropriate sampling of domain processes in addition to traditional coverage of domain content; the other is the need to move beyond traditional professional judgement of content to accrue empiricial evidence that the ostensibly sampled processes are actually engaged by respondents in task performance (ebd.: 745).

Wichtig ist also nicht nur, den Inhalt zu definieren, sondern auch zu überprüfen, ob dieser durch die vordefinierten Prozesse erfasst wird. Und er schreibt weiter:

> Thus, the substantive aspect adds to the content aspect of construct validity the need for empiricial evidence of response consistencies or performance regularities reflective of domain processes […]. Such evidence may derive from a variety of sources, for example, from “think aloud” protocols […]. (ebd.: 745)

Messick bezeichnet damit kognitive Interviews als Instrumente, die den substantiellen Teil der Konstruktvalidität erfassen können. Der Prozess, wie Befragte zu ihren Bewertungen kommen, wird mit der inhaltlichen und der konvergenten/divergenten Validität auf dieselbe Stufe gestellt.

24 Messick forscht auf dem Gebiet der Bildungswissenschaft und Psychologie, und seine Ergebnisse bezieht er auf Forschungen zu Leistungs- und Wissensbeurteilungen im pädagogischen Kontext. Seine Überlegungen scheinen aber durchaus generalisierbar und auf die Einstellungs-/Wertemessung umlegbar zu sein.

Die Methoden zu kombinieren, kann auch dadurch gerechtfertigt werden, dass ein nicht unerheblicher Teil an Problemen, die in kognitiven Interviews entdeckt werden, erst durch den Erhebungs- und Analyseprozess an sich zu solchen werden (vgl. Willis, 2005). Die Wahrscheinlichkeit, dass andere Forscher zu den gleichen Schlüssen kommen, ist höher, wenn die Problemfelder durch eine Vorauswahl mittels einer weiteren Methode eingeschränkt werden (vgl. hierzu die Diskussion um „*inferences*“ bei Teddlie und Tashakkori, 2009).

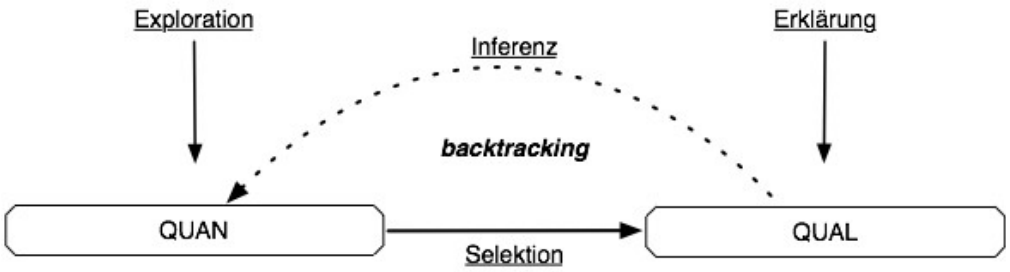
Ein Ziel dieser Arbeit ist es, sich kritisch mit den Methoden an sich auseinanderzusetzen. Es soll deshalb erwähnt werden, dass *einerseits* ein stetig wachsender Korpus an Forschungsergebnissen der Methode der kognitiven Interviews kritisch gegenüber steht und ihre Leistung gegenüber anderen qualitativen Methoden (Experteninterviews und *appraisal method*) relativiert (Presser und Blair, 1994; Willis et al., 2000; Rothgeb et al., 2001). *Andererseits* lieferte die ausgiebige Diskussion über die Messung von Werten mit unterschiedlichen quantitativen Methoden, von der ein Ausschnitt bereits präsentiert wurde, Hinweise, dass auch im quantitativen Paradigma ein Bedarf zur Klärung von Stärken und Schwächen der Methoden besteht (vgl. hierzu vor allem Steinmetz et al., k.A.). Yan, Kreuter und Tourangeau (2010) haben eine Studie durchgeführt, in der sie verschiedene Methoden der Beurteilung der Qualität von Survey-Indikatoren sowohl quantitativen als auch auch qualitativen Ursprungs (Expertenbefragung, kognitive Interviews, quantitative Maße der Validität und Reliabiltät und Fehlerraten von latenten Klassenanalysen) gegeneinander testeten. Sie stellten sich die Frage, ob und wie stark die Ergebnisse der Methoden konvergierten. Die Forscher stellten fest, dass kognitive Interviews und die quantitative Validitätsanalyse die höchste Übereinstimmung bei der Identifikation von Problemen fanden (dieses Ergebnis trifft damit für die in der vorliegenden Arbeit angewendeten Methoden ebenso zu). Die weiterhin bestehende erhebliche Diskrepanz zwischen den Methoden erklären sie damit, dass diese mit unterschiedlichen Annahmen operieren, die wenig bis gar keine Überschneidung aufweisen. Und das, obwohl alle dasselbe Ziel verfolgen.

3.2.2 Wie lässt sich das Design klassifizieren?

Das Design in der vorliegenden Anwendung ist prinzipiell keine Neuheit: Die Entwicklung eines Messinstruments ist ein länger andauernder, interaktiver Forschungsprozess, in dem unterschiedliche Methoden alternierend angewendet werden. Im Idealfall wird das Instrument nach der Entwicklung Pretests unterzogen, geht in die eigentliche Erhebungsphase, wird ausgewertet und dabei evaluiert und mündet schließlich in einer Form der qualitativen Untersu-

chung. Sehr oft sind das kognitive Interviews, Experteninterviews oder Fokusgruppen (vgl. Moosbrugger und Kelava, 2012). Dennoch weist dieses spezielle Design Besonderheiten auf, die eine herkömmliche Klassifizierung schwieriger machen. Es ist in mehreren Komponenten vermischt: *Einerseits* werden zwei unterschiedliche Datenquellen zum selben Gegenstand einmal quantitativ und einmal qualitativ analysiert. Die quantitativen Daten wurden vom ESS übernommen, und ihre Analyse erfolgte sekundär. Die qualitativen Interviews wurden hingegen selbstständig geführt und ausgewertet. *Andererseits* wurden mehrere qantitative Methoden teilweise ergänzend und teilweise kontrastierend eingesetzt. Die quantitative Analyse wurde insofern im Gesamtprozess stärker gewichtet. Diese Gewichtung lässt sich einerseits durch meine unterschiedliche methodische Kompetenz im jeweiligen Paradigma erklären (für die quantitative Analyse wurden weitaus raffiniertere Methoden angewendet als für die qualitativen Protokolle) als auch durch das Erkenntnisinteresse und Objekt der Forschung, nämlich die Evaluierung einer quantitativen Skala. Ein weiterer Grund ist durch das Design bedingt: Die quantitative Analyse ist die Ausgangslage für die Einflechtung des qualitativen Materials. Letzteres gäbe noch bei weitem mehr Informationen her, als in diesem Forschungsprojekt verwirklicht wurden.

Dem Design liegt eine kausale Struktur zugrunde, die den üblichen Forschungsprozess rückwärts geht (siehe Abbildung 4): Kognitive Pretests werden üblicherweise angewendet, um Probleme, die von den Entwicklern nicht bedacht wurden, in einem bestehenden Erhebungsinstrument zu entdecken, bevor sich problematische Messungen ergeben. Hier wird der Prozess umgedreht. Die quantitativen Ergebnisse identifizieren problematische Items, um sie mit den kognitiven Interviews ursächlich zu verstehen. Für diesen Gesamtprozess wurde der Begriff *Backtracking* eingeführt.



Die Definition des Triangulationsdesigns (Creswell und Plano Clark, 2007) kommt dem vorliegenden Design am nächsten. Dessen Ziel besteht darin, „to obtain different but comple-

mentary data on the same topic“ (Morse, 1991, in ebd.: 62). Es werden die Stärken der quantitativen (große Stichproben, Verläufe, Generalisierungen) und qualitativen (überschaubare Stichproben, Finessen, Gründlichkeit) Methoden gebündelt ohne dabei gleichzeitig auch ihre Schwächen zu verstärken. Üblicherweise stehen dann beide Methoden im Design gleichwertig nebeneinander, was hier – wie beschrieben – nicht geschah. In der Variante des Konvergenzmodells werden die Daten und Ergebnisse durch Vergleich oder Kontrast während der Interpretation zusammengeführt. „The purpose of this model is to end up with valid and well-substantiated conclusions about a single phenomenon“ (ebd.: 64).

Nach der Klassifizierung von Teddlie und Tashakkori (2009) wäre das Design am ehesten als *parallel-mixed-methods-multistrand*-Design einzuordnen. Die *strands* sind die aufeinanderfolgende Schritte, die zu einer Studie gehören (beginnend bei der Konzeptualisierung, über Datengenerierung/-analyse bis hin zu den Schlüssen, die daraus gezogen werden). Die vorliegende Forschung ist nicht *monostrand,* weil nicht derselbe Datenpool sowohl quantitativ als auch qualitativ analysiert wird (Teddlie und Tashakkori, 2009: 144ff.). Da die quantitativen Daten bereits vor den qualitativen vorhanden waren, hätte der Forschungsverlauf theoretisch sequentiell ablaufen können: Ergebnisse aus der statistischen Datenanalyse hätten in ein darauf ausgerichtetes qualitatives Design münden können. Die primäre Entscheidung, die Stränge parallel laufen zu lassen – mit der qualitativen Erhebung vor der quantitativen Analyse zu beginnen – beruht auf der Überlegung, dass zu viel Vorwissen des Interviewers den Befragungsprozess stark beeinflussen könnte. Dadurch wird das Forschungsfeld aus einer etischen Sicht vorstrukturiert und die möglichen Ergebnisse eingeschränkt. Um eine stärker emische Perspektive zu erhalten und damit den entdeckenden Charakter qualitativer Methoden (Tashakkori und Teddlie, 1998) in den Vordergrund zu rücken, wurde der parallele Weg gewählt.

3.3 Konfirmatorische Faktorenanalyse (CFA)

Wie ein Index gebildet wird, wurde bereits erwähnt. Laut Bollen (1989) kann eine Messung auch anders definiert werden: „Measurement is the process by which a concept is linked to one or more latent variables, and these are linked to observed variables“ (ebd.: 180). Damit befinden wir uns im Rahmenwerk des Strukturgleichungsansatzes (engl. *structural equation modeling,* kurz: SEM). Im SEM-Ansatz kennt man kausale (Regressionen; in Mplus als „by“-Kommando definiert) und nicht kausale (Korrelationen oder Kovarianzen; in Mplus als „with“-Kommando definiert) Beziehungen, die in Pfaddiagrammen mittels Pfeilen oder in ei-

nem Gleichungssystem dargestellt werden, indem Pfadregeln befolgt werden. Die Pfaddiagramme sind die graphische Darstellung der Gleichungen, die aus Bekannten (Messwerten) und Unbekannten (Regressionsgewichten, Korrelationen, Varianzen, Mittelwerten) bestehen. Mit ihnen lassen sich Zusammenhänge zwischen Indikatoren und latenten Variablen einerseits und latenten Variablen mit anderen latenten andererseits in einem Modell darstellen. Die dafür nötigen Parameter werden durch das Programm geschätzt.

Die Beziehungen zwischen mehreren gemessenen (manifesten) und einer latenten Variable bilden gemeinsam ein Messmodell. Darin bestehen zwischen ihnen Beziehungen (sogenannte Ladungen, auch λ), die das zu erwartende Verhältnis zwischen einem Indikator und dem Konstrukt ausdrücken: eine Änderung um eine Einheit im latenten Konstrukt bewirkt eine entsprechende Änderung in der Manifesten (in diesem Fall wird die Beziehung so definiert, dass – wie bei der Itemcharakteristik – eine Ausprägung in der latenten Variable zu einer bestimmten Ausprägung in der manifesten führt). Um feststellen zu können, was eine Änderung einer Einheit in der latenten Variable bedeutet, benötigt sie eine Skalierung (Brown, 2006). Da wir in den Sozialwissenschaften üblicherweise nicht wissen, wie die Skalen unserer Konstrukte aussehen, muss diese Skalierung über ein externes Kriterium definiert werden. Eine unter mehreren Möglichkeiten ist eine der Ladungen auf „1“ zu setzen. In diesem Fall ist die Beziehung zwischen Latenter und Indikator, der auf die Latente lädt, so definiert, dass sich eine Änderung in einer Komponente im Verhältnis 1:1 in der anderen niederschlägt. Dadurch übernimmt die Latente die Skalierung der Manifesten.

Über die standardisierte Lösung kann das Ausmaß des Zusammenhangs festgestellt werden. Er kann perfekt ($\lambda=+/-1$) sein. Es gibt auch den Fall, dass er überhaupt nicht vorhanden ist ($\lambda=0$) oder sich als Anteil ausdrückt ($-1<\lambda<1$). Ist der Zusammenhang nicht perfekt ($\lambda\neq+/-1$), dann deutet dies auf einen Messfehler (δ) hin (Messfehler der verschiedenen Indikatoren sind *per definitionem* miteinander und mit dem Konstrukt unkorreliert). Mit anderen Worten: Beim SEM-Ansatz wird angenommen, dass sämtliche Varianz durch den geteilten Einfluss der latenten Variable erzeugt wird und der Messfehler zufälliger Natur ist (Bollen, 1989). Wenn auch noch genügend Informationen vorhanden sind, um das Gleichungssystem lösen zu können, ist das Messmodell vollständig identifiziert, und die gewünschten Parameter können geschätzt werden. Gibt es im Gleichungssystem mehr Bekannte als Unbekannte, gibt es auch mehrere mögliche Lösungen für die Parameter. Der Anwender benötigt allerdings nur eine, und zwar jene, die am besten zu der Datenstruktur passt. Das Maximum-Likelihood-Schätzverfahren (ML) ist ein in den Sozialwissenschaften häufig verwendetes. Es hilft, die-

se optimale Lösung zu finden, indem es jene Parameter auswählt, bei denen am wahrscheinlichsten ist, dass sie wieder in einer Zufallsstichprobe auftauchen würden. Oder, wie Brown (2006: 73) ML definiert: „[…] ML aims to find the parameter values that make the observed data most likely (or conversely, maximize the likelihood of the parameters given the data)". Gibt es mehrere Lösungen, wird mittels eines iterativen Prozesses, der sogenannte *starting values* voraussetzt, eine optimale Lösung geschätzt. Diese werden üblicherweise von Auswertungsprogrammen vorgegeben. ML ist ein praktisches Werkzeug, das aber nur unter gewissen Bedingungen effektiv arbeitet:

a) Die Stichprobe muss „groß" sein.

b) Die Daten müssen zumindest intervallskaliert sein (Voraussetzung sind eine ordinale Struktur und gleich große Intervalle).

c) Die Indikatoren müssen multivariat normalverteilt sein.

Bedingung (a) ist im Fall des hier verwendeten ESS mit N>2000 hinreichend erfüllt. Bedingung (b) wurde bereits im Kapitel 3.1.2 besprochen. Bedingung (c) ist bei den vorliegenden Daten nicht erfüllt. Mit dem Statistikprogramm SPSS kann ein Test auf multivariate Normalverteilung nicht berechnet werden. Der nächstbeste Test ist der auf Normalverteilung der einzelnen Variablen, weil sie eine Bedingung für die multivariate Normalverteilung ist. Diese wird für keine der Variablen erfüllt (siehe Kapitel 4.2, Seite 76). Das ist aber für Auswertungsprogramme wie Mplus kein Problem, da hier eine Mehrzahl an Schätzverfahren zur Verfügung stehen, die gegenüber der Verletzung von (c) robust sind. Maximum-Likelihood Robust (MLR) liefert sowohl unter Normalverteilung als auch unter nicht normalverteilten Daten robuste Schätzungen für Parameter und Standardfehler (engl. *standard error*, kurz: S.E.).

Zur Ermittlung der Parameter werden die Matrizen der empirischen und vom Modell geschätzten Varianzen und Kovarianzen miteinander verglichen, um statistisch signifikante Abweichungen (Residuen) aufzuspüren. Sinn dieses Vergleichs ist es, festzustellen, wie gut das Modell den Zusammenhang in den Daten wiedergibt. Das ist die Grundlage der Goodness-of-Fit-Indizes, die den globalen Zustand des Modells beurteilen. Diese werden vom Programm, das für die Analyse verwendet wird, bereitgestellt. Mplus gibt die folgenden aus (in Klammer stehen die gebräuchlichen Abkürzungen samt minimalen Kriterien, die ein Index erreichen muss, um adäquaten Fit zu signalisieren; vgl. hierzu Brown, 2006; Hu und Bentler, 1999): *p-value* des ^{2}-Wertes (abhängig von Freiheitsgraden, kurz: df), *Comparative Fit Index* (CFI größer oder gleich 0,95), *Tucker-Lewis Index* (TLI größer oder gleich 0,95), *Root Mean*

Square Error of Approximation (RMSEA kleiner oder gleich 0,06) und *Standardized Root Mean Square Residual* (SRMR kleiner oder gleich 0,08). Hu und Bentler (1999) verwenden bei den numerischen Angaben die Phrase „close to“, weil die Kriterien Schwankungen unterliegen. Daneben bieten Regressionsgewichte (Ladungen) und das Bestimmtheitsmaß (R^2) Möglichkeiten, die Qualität der einzelnen Komponenten des Modells zu beurteilen. Sie sind auch im SEM-Ansatz die Maßzahlen für die Güte von Reliabilität und formaler Validität (vgl. Bollen, 1989) und haben gegenüber herkömmlichen Beurteilungskriterien, die rein auf Varianz-/Kovarianzanalysen basieren, Vorteile. Diese liegen darin begründet, dass CFA Messfehler kontrolliert, d.h. Kovarianzen werden ausschließlich über gemeinsame Faktoren oder deren Korrelationen untereinander dargestellt (vgl. ebd.).

Auf diese Weise können Theorien, die Zusammenhänge postulieren, getestet werden und auf ihre Güte in einem bestimmten Datensatz hin überprüft werden. Im Fall der Werteskala nach Schwartz, in der es keine kausalen Zusammenhänge, sondern nur Kovarianzen zwischen den Faktoren (Wertetypen) gibt, ist eine dafür geeignete Analysemethode die konfirmatorische Faktorenanalyse (CFA). Sie unterscheidet sich von der exploratorischen Faktorenanalyse (EFA) dadurch, dass sie weit sparsamer ist. Während EFA Ladungen von allen Faktoren auf alle Items zulässt und durch Rotation Ladungen pro Faktor maximiert oder minimiert, bestimmt der Analyst anhand einer Theorie bei CFA, welche Faktorladungen und Korrelationen geschätzt und welche von vornherein unterbunden werden, indem sie im Gleichungssystem auf Null gesetzt werden (vgl. Brown, 2006).

Üblicherweise passen Daten nicht perfekt zu Theorien, die die Zusammenhänge vorgeben. Das passiert, wenn das Modell schlecht spezifiziert wurde (falsche Anzahl an latenten Faktoren, falsche Zuweisung der Indikatoren zu den Faktoren), die Annahme der unkorrelierten Fehler verletzt wurde (ebd.: 159), oder wenn die Theorie die empirische Wirklichkeit nicht adäquat abbildet, etc.. In jedem Fall bedeutet das eine Handlungsanweisung für die Forscher. Latcheva (2009) fasst die drei Optionen, die dabei zur Wahl stehen, zusammen:

1. Die Modelle solange anpassen, bis sie den empirischen Daten entsprechen. Das ist eine praktikable Lösung, birgt aber die Gefahr, dass unter den Anpassungen die Inhaltsvalidität des Konstrukts leidet (weil z.B. Indikatoren, die zwar schlechte Ladungen haben, aber wichtige Dimensionen abdecken, aus dem Modell ausgeschlossen werden). Außerdem stehen oft mehrere Modifikationsvorschläge in Konkurrenz nebeneinander, und eine Vielzahl an angepassten Modellen geben die Daten mehr oder

weniger gleich gut wieder. Dadurch geht allerdings der konfirmatorische Charakter der CFA verloren.

2. Die zugrunde liegende Theorie anpassen. Diese Lösung wäre ganz im Sinne einer positivistischen Sozialwissenschaft. Allerdings stehen die Forschenden vor dem Problem, dass sie für die Evaluierung einer Theorie valide Messungen brauchen. Aus der Diskrepanz zwischen Theorie und Empirie lässt sich nicht logisch ableiten, auf welcher Seite das Problem liegt.
3. die Messung hinterfragen „[…] with regard to their [i.e. the measures] content validity and to systematically improve the measurement through an integration of quantitative and qualitative appproaches to validity“ (ebd.: 1193).

Wie man sieht, macht es einen Unterschied, welche Option gewählt wird. Modifikationen einfach anzunehmen kann bedenkliche Folgen nach sich ziehen. Der Ansatz der vorliegenden Arbeit liegt darin, der dritten Option zu folgen und signifikante Kreuzladungen und Fehlerkovarianzen, die sich aus dem Modell ergeben, als Hinweis für erklärungsbedürftige Missspezifikationen des Modells zu verstehen. Um das Vorgehen im empirischen Teil besser nachvollziehen zu können, wird im Folgenden detaillierter auf die Ursachen von Modifikationen eingegagen.

Kreuzladungen weisen darauf hin, dass der Zusammenhang zwischen den Items eines Konstrukts mit denen eines anderen nicht vollständig durch die Messmodelle und die Kovarianzen der latenten Faktoren erklärt werden können. Das Programm, das keine Ahnung von der zugrunde liegenden Theorie und ihren Implikationen hat, schlägt eine Ladung eines Faktors auf ein „fremdes“ Item dann vor, wenn ausreichend signfikante Abweichung für das Item vorhanden ist. Wurde der Zusammenhang tendenziell überschätzt, ist die Kreuzladung negativ. Wurde er tendenziell unterschätzt, ist sie positiv. Hingegen weisen Fehlerkovarianzen auf eine spezifische Dimension hin, die nur von bestimmten Items geteilt wird. Wie oben beschrieben, ist die Fehlervarianz das Residuum, also der Anteil der Varianz/Kovarianz, der durch Ladungen und Varianz der latenten Variable im Indikator nicht erklärt werden kann (deshalb spricht man im Englischen auch von *unique variance*). Es wird angenommen, dass Messfehler prinzipiell nicht miteinander korrelieren. Der Grund dafür wurde bereits erwähnt: der beobachtete Zusammenhang zwischen zwei Indikatoren wird vollständig über die latenten Faktoren erklärt. Wird diese Annahme verletzt, hat zusätzlich eine diesen beiden Items eigene Dimension einen Einfluss auf den Zusammenhang. Dies kann zutreffen, wenn seine Ursache methodisch begründet ist, wenn z.B. die Items eine spezifische Bewertungsmethode teilen

und beide ähnlich formuliert sind, oder wenn sie von anderen Indikatoren unterschiedliche Anfälligkeiten für Einflüsse wie Antwortmöglichkeiten, Anforderungsmerkmale, Antworttendenzen, Leseschwierigkeiten oder soziale Erwünschtheit teilen (Brown, 2006: 46). Diese Aufzählung an Bedingungen ist nicht immer ganz leicht aus dem reinen Wortlaut der Indikatoren zu entnehmen. In den tatsächlichen Antwortprozess hineinzublicken kann helfen, tieferliegende Gründe zu finden.

Man sieht, Anpassungen inhaltlich zu begründen ist eine Herausforderung. Das spezielle Design der vorliegenden Arbeit soll es ermöglichen, dazu Stellung zu nehmen. Im folgenden Abschnitt geht es um eben jene Methode, die Einblick in den Antwortprozess bringen kann.

3.4 Kognitive Interviews

Meine Interviews führte ich im Sinne der kognitiven Pretesttechniken durch, wie sie von Prüfer und Rexroth (2000) beschrieben werden: Diese Techniken stammen ursprünglich aus der Kognitionspsychologie und fanden ihre Anwendung in den Sozialwissenschaften, um die Beziehung zwischen Frage und Antwort zu erforschen. In der Durchführung gibt es prinzipiell zwei Anordnungen, die sich im Grad der Kontrolle über den Befragungsprozess unterscheiden (ebd.):

- Traditionellerweise werden kognitive Pretests sehr offenen gestaltet. Der Versuchsleiter, der mit dem theoretischen Hintergrund der Fragen vertraut ist, fragt nach oder vertieft einen Inhalt. Die Vorteile liegen auf der Hand: Der „Experte“ weiß aus Erfahrung, wo er weiterfragen muss. Das Spektrum an möglichen Erklärungen erweitert sich. Problematisch ist die mangelnde Vergleichbarkeit der einzelnen Befragtenaussagen, da immer unterschiedliche Stimuli gesetzt wurden.
- Die standardisierte Variante sieht einen sogenannten Evaluationsfragebogen vor, in den die unterschiedlichen Techniken eingearbeitet wurden. Das Instrument gibt die Vorgehensweise vor. Dadurch mangelt es ihm auch an Flexibilität. Vorteilhaft wirkt sich aus, dass die Interviewer nicht theoretisch unterwiesen werden müssen. Die Auswertung ist einfacher, und durch die Standardisierung können auch mehrere Interviews durchgeführt werden. Die Erstellung des Evaluationsfragebogens ist allerdings eine aufwändige Angelegenheit. Vorherige Zielsetzung und mögliche Probleme müssen bekannt sein.

Den theoretischen Hintergrund zu dieser Form der Fragestellung (sich mit den kognitiven Prozessen im Ablauf eines Frage-Antwort-Prozesses zu beschäftigen) bildet eine Bewegung aus den 1980er Jahren mit dem Namen „Cognitive Aspects of Survey Methodology“ (kurz: CASM). Tourangeau, ein Proponent derselben, beschreibt, welchen Paradigmenwechsel sie in den Sozialwissenschaften hervorgerufen hat:

> During the past 20 years, survey methodology has undergone a paradigm shift. The old paradigm was based on a statistical model that focused on the *effects* of survey errors on the estimates derived from survey data. The new paradigm is based on a social scientific model that focuses on the *causes* of survey errors (Tourangeau, 2003: 3).

Die Erkenntnis seiner Forschungsarbeit, die unter dem Banner von CASM durchgeführt wurde, fasst er in einem Satz zusammen: „[…] reporting errors in surveys arise from problems in the underlying cognitive processes through which respondents generate their answers to survey questions“ (ebd.: 5). Ein wichtiger Beitrag von CASM war unter anderem auch, zu einer Systematisierung dessen beizutragen, was im Frage-Antwort-Prozess während einer empirischen Erhebung vor sich geht. Befragte durchlaufen demgemäß vier Prozesse:

- Verständnis der Frage
- Abruf relevanter Informationen bzw. Ereignisse aus dem Gedächtnis
- Treffen von Entscheidungen darüber, wie geantwortet wird
- Zuordnung der „intern“ ermittelten Antwort zu formalen Antwortkategorien

Um diesen Ablauf beobachten zu können, haben sich in der Forschungstradition fünf Techniken mit je unterschiedlichen Anwendungsmöglichkeiten bewährt (Prüfer und Rexroth, 2005):

1. Confidence Rating (Bewertung der Verlässlichkeit der Antwort)
2. Sorting (Sortiertechnik)
3. Paraphrasing (Paraphrasieren)
4. Think Aloud (Methode des lauten Denkens)
    a. Concurrent Think Aloud: Befragte sagen, was ihnen gerade durch den Kopf geht.
    b. Retrospective Think Aloud: Befragte sagen im Nachhinein, was ihnen durch den Kopf ging.
5. Probing (Hinterfragen von Antworten)

Für die Interviews der vorliegenden Arbeit wurde zwar ein Evaluationsfragebogen gestaltet, der allerdings nur als Leitfaden diente. Die Interviews wurden sehr frei geführt. Es war mir

wichtig, den Befragten und mir möglichst viel Freiraum im Prozess zu geben, damit das Feld der möglichen Antworten breit angelegt werden konnte und genügend Platz für freie Assoziationen blieb. Aufgrund der Definition von Schwartz, bei Werten handle es sich um affektive Urteile und weniger um kognitive, und da mein Erkenntnisinteresse auch dadurch geleitet war zu erfahren, wie Befragte mit der Beantwortung von Werten an sich und mit den Porträts im Speziellen zurecht kamen, schien mir eine breite Herangehensweise sinnvoller. In der Auswahl der Techniken beschränkte ich mich auf *Paraphrasing, Think Aloud* und vor allem *Probing*. Think-Alouds sind eine nicht unproblematische Methode, wie Prüfer und Rexroth (2000) erwähnen. Das konnte ich durch eigene Erfahrungen bestätigen: Es könnte die perfekte Methode sein, vor allem das Concurrent Think Aloud, weil es direkten Zugriff auf – zumindest – die verbalisierten Vorgänge der Gedankenprozesse des Interviewten liefert. Der Nachteil ist jedoch, dass es hohe kognitive Anforderung an die Befragten stellt. Prüfer und Rexroth (2000) empfehlen daher auch, diese Technik nicht ohne vorherige Aufwärmphase für den Befragten durchzuführen. Dazu wird der Interviewte gebeten, sich gedanklich an einen anderen Ort zu begeben und dort eine kognitive Aufgabe zu lösen. Währenddessen soll er sagen, was ihm durch den Kopf geht. Ich übernahm diese Übung und bat meine Befragten, sich gedanklich für ca. zehn Minuten in ihre eigene Wohnung zu versetzen. Während sie durch die Räume gingen, ließ ich sie mir von ihnen beschreiben. Das habe ich solange gemacht, bis mir vier Befragte im Nachhinein sagten, dass diese Übung für die tatsächliche Herausforderung als nicht zielführend wahrgenommen wurde. Assoziationen zu konstruierten Porträts liegen zu fern von konkreten alltäglichen Situationen, um wechselseitig Lernerfolge übertragen zu können. Sehr bald ging ich daher auch zum Retrospective Think Aloud über, das den Befragten weitaus leichter fiel, wenn auch klar war, dass so manche Erklärung dazu diente, die eigene Bewertung eines Items u.a. zu rechtfertigen. Dies ist womöglich die Situation, die Maio und Olson (1998) mit ihrer *values-as-truisms*-Hypothese beschreiben.

Probing und Paraphrasing erwiesen sich als die effektiveren Mittel. Bei Probing werden Befragte nach der Beantwortung des Items mit anderen Fragen zum Frageverständnis, der Informationsbeschaffung und der Wahl der Antwortkategorien konfrontiert. Sehr oft waren die tatsächlich angewendeten einzelnen Fragetechniken aber nicht klar einer einzigen Kategorie von *probes* zuzuordnen, was aber auch bei einem offenen Design nicht so ins Gewicht fällt. Hingegen schien mir umso wichtiger zu sein, mich an Hinweise von Willis zu halten, wie man Fallgruben beim Nachfragen vermeidet:

1. Avoid bias and other classic questionnaire design problems. 2. Be especially careful when using paraphrasing. 3. Retrench when our probes “don't work”. 4. Limit speculative probing. 5. Be conservative in the number of probes asked. 6. Don't view the subject as a questionnaire design critique. 7. Avoid inventing problems by accepting that our expectations may be wrong.

Tabelle 5: „Avoiding pitfalls in asking probes“ (aus Willis, 2005: 141)

Sampling

Vor der Durchführung der Studie wurde ein Probeinterview geführt, um den Evaluationsbogen und das Setting zu testen. Das tatsächliche Sample besteht aus elf Interviewpartnern, die nach einem vorher festgelegten Quotenplan nach Alter, Geschlecht und Bildung ausgewählt wurden. Vier Respondenten haben einen Migrationshintergrund. Davon haben drei ihre Wurzeln im ehemaligen Jugoslawien und eine Befragte in Ungarn. Alle verfügten über ausreichende Deutschkenntnisse, um die kognitiven Interviews zu bewältigen. Für das Alter erstellte ich drei Katego-rien mit den Einteilungen 16-30 Jahre, 31-60 Jahre und älter als 61 Jahre. Ich zog eine Altersuntergrenze bei 16, weil hier auch das Alterslimit für die Teilnahme am ESS liegt. Für Bildung wurden die Kategorien so gewählt, dass es eine Gruppe ohne Matura, eine nur mit Matura und eine mit Hochschulabschluss gab. Rekrutiert wurden Bekannte, die sich unentgeltlich zur Verfügung stellten. Wegen meiner beschränkten zeitlichen Ressourcen musste ich die Zahl der Interviews auf elf, sowie den Wohnort auf Wien beschränken. Die Anzahl der durchgeführten Interviews ist durchaus entscheidend dafür, wie viele und welche Probleme im Erhebungsinstrument gefunden werden. Blair et al. (2006) untersuchten diesen Zusammenhang: Wird die Anzahl der charakteristischen Probleme betrachtet, dann gibt es einen Effekt des steigenden, aber abnehmenden Grenzwerts. Das heißt, die Anzahl der entdeckten Probleme steigt beginnend bei fünf bis etwa 20 sprunghaft an, um dann mit zusätzlicher Interviewzahl zwar weiter zu steigen, allerdings nicht mehr im selben Umfang. Ähnliches geschieht, wenn die Schwere der Fehler beobachtet wird: Die Unterscheidung der Schwere liefert keine Unterschiede zwischen den Gruppen. Bereits eine kleine Stichprobe entdeckt schwere Probleme. Generell nimmt die Effektivität ab etwa 20 bis 30 Interviews ab. Die Wahrscheinlichkeit, ein gewisses Problem zu entdecken, nimmt mit steigender Stichprobe zu. Aber selbst bei 50 Interviews wird nur die Hälfte aller Probleme mit einer Wahrscheinlichkeit von 76 bis 99 % entdeckt (ebd.: 4).

Die geführten Interviews waren durchschnittlich eine Stunde lang. Für die Transkription benötigte ich etwa das sieben- bis achtfache der Zeit. Urspünglich verwendete ich zur Auf-

nahme der Interviews mein eigenes Mobiltelefon, das eine Applikation zur Aufnahme und ein sehr sensibles Mikrofon besitzt. Leider ist es unzuverlässig, sodasss sich bei zwei Interviews

	Alter (in Jahren)			**Geschlecht**		**Ausbildung**		
Respondent	*16 – 30*	*31 – 60*	*> 60*	*männlich*	*weiblich*	*<Matura*	*Matura*	*Hoch-schule*
1		x		x			x	
2			x	x				x
3	x				x		x	
4			x		x			x
5	x			x		x		
6		x		x			x	
7		x			x		x	
8	x			x		x		
9	x			x				x
10	x				x	x		
11		x			x			x
Spaltensumme:	5	4	2	6	5	3	4	4

Tabelle 6: Quotenplan kognitive Interviews

herausstellte, dass in einem Fall das Format fehlerhaft war und im anderen die Aufnahme überhaupt nicht stattgefunden hatte. Die Interviews wurden nicht mit denselben Personen nachgeholt, da ein *priming effect* eintreten kann, der die Interviews nicht mehr vergleichbar werden lässt: Die Befragten sind zu einem gewissen Grad an die Form des Frage-Antwort-Spiels gewöhnt und wissen womöglich zu gut, was von ihnen erwartet wird. Sie stellen vielleicht nicht mehr ihre Fähigkeiten als Befragte zur Verfügung, sondern befinden sich dann in der Rolle des *questionnaire design critiques*. Dieser Effekt fiel mir auch während ich die Interviews führte, auf: Etwa nach dem ersten Drittel der Items erläuterten die Befragten von sich aus, wie sie zu ihren Antworten kamen, also ohne mein Zutun. Ein Lerneffekt war nicht zu leugnen. Es fehlt mir allerdings ein Kriterium zu beurteilen, ob sich dieser negativ oder positiv auf die Qualität der Ergebnisse auswirkte. Ich schaltete mich nur dann ein, wenn ich spezifische Informationen suchte, wenn mir Inkonsistenzen in den Antworten auffielen oder die Befragten sich in ihren eigenen Antworten verfingen. Bei den fehlerhaften Interviews kommt zum *priming effect* hinzu, dass es eine Zumutung für den Befragten ist, noch einmal unentgeltlich die Zeit zu investieren. Darüber hinaus ist das Frustrationspotential nach Abbruch auf Grund eines technischen Fehlers, der das Interview nutzlos machte, für alle Beteiligten sehr hoch. Ab dann verwendete ich ein professionelles Gerät und ließ das Telefon als Backup mitlaufen. Die beiden genannten Interviews, die ich aus technischen Gründen nicht auswerten konnte, sind in der Stichprobe von 11 nicht enthalten.

Alle Interviews bis auf eine Ausnahme fanden in einer Wohnung statt, entweder bei den Befragten oder bei mir. Sie liefen zum Großteil störungsfrei und fokussiert ab. Nur in einigen Fällen kamen leider Mitbewohner hinzu, die das Antwortverhalten stark beeinflussten und die Konzentration beider Teilnehmer senkten.

4 Empirie

In diesem Abschnitt wird zunächst dargelegt, wie der Datensatz ausgewählt wurde. Probleme, die auftauchen, wenn unterschiedliche Samples für Analysen herangezogen werden, werden angesprochen. In einem nächsten Schritt werden die Ergebnisse der Analyseschritte *deskriptive Statistik,* e*xploratorische Faktorenanalyse, konfirmatorische Faktorenanalyse* und *Analyse der fehlenden Werte* besprochen und kommentiert. Die Vielzahl an hier verwendeten Methoden soll den derzeitigen Stand der Forschung widerspiegeln. Widersprüchliche Ergebnisse aus den quantitativen Methoden regen dazu an, über die Testung der Qualität von Messungen nachzudenken und neue Wege zu gehen. Darüber hinaus sind wiederholte Diskrepanzen oder Überschneidungen von Problemen in den einzelnen Analyseschritten für die Triangulation wichtig. Während stufenweise strengere Tests mit den Daten durchgeführt werden, werden entlang der einzelnen Testschritte Indizien für Probleme in der Messung gesammelt.

An die beschreibende und exploratorische Phase rückt die essentielle Frage, die in der Diskussion rund um die Messqualität oft übersehen wurde: Was kommt letztendlich bei den Befragten an unterschiedlichen Stimuli an? Sind es auch diejenigen, die von den Entwicklern bezweckt wurden? Denn neben formalen und externen Kriterien zur Beurteilung der Messqualität, die allesamt auf demselben „Werkzeugkasten" der quantitativen Methoden beruhen, geht diese notwendige Dimension verloren. Im Kapitel 4.4.1 gebe ich daher einen Einblick, wie durch eine Kombination aus quantitativen und qualitativen Methoden ein Erkenntnisgewinn gegenüber einem einzelnen Verfahren errungen werden kann. Besonders geeignet für eine schrittweise quantitative Testung scheint mir das Vorgehen von Knoppen und Saris (2009a; b) zu sein. Nach ihrem Vorbild werden die Messmodelle zuerst einzeln und dann in ihrer Nachbarschaft zueinander getestet. Einschränkend in der hier verwendeten Variante des PVQ fällt der Mangel an Items pro Wertetyp aus, sodass nicht die von Schwartz geforderten zehn Wertetypen, sondern nur die vier *higher-order values* verglichen werden können. Detaillierte Ausdrucke und eine Zusammenfassung der Itembezeichnungen und -formulierungen[25] sind im Anhang (Kapitel 7) zu finden.

25 Die Items werden, sofern sie im Text erwähnt werden, mit der Variablenbezeichnung des ESS bezeichnet. Das Kürzel ermöglicht eine schnellere Zuordnung zum Inhalt. Es ist ein Suffix beigefügt, das den zugehörigen Wertetyp kennzeichnet.

4.1 Welche Version? Warum diese?

Der PVQ im ESS bietet die Möglichkeit, mittels der „dynamischen" Eigenschaften der Werte Vorhersagen über Konflikte und Übereinstimmungen von Wertetypen zu prüfen und anhand dieser Hypothesen andere Konstrukte zu testen, zu validieren oder theoretische Zusammenhänge zu prüfen.

Als die Analyse durchgeführt wurde, gab es für Österreich drei Wellen des ESS. Alle drei einer eingehenden Analyse zu unterziehen, sprengt den Rahmen der Arbeit. Aufgrund der methodischen Genauigkeit, die Durchführorganisationen des ESS kolportieren, sollten Ergebnisse eines Datensatzes auf die übrigen Wellen umlegbar sein. Bei der exploratorischen Beschäftigung mit den Daten fiel aber auf, dass sich die Wellen in Bezug auf die *human values* in ihrer Faktorenstruktur unterscheiden. Um Anschluss an die aktuelle Debatte zu erreichen, wurde derjenige Datensatz ausgewählt, der die von Schwartz postulierten Faktorenstruktur am besten reproduziert. Parallele Rechnung in Mplus (dieses Programm ist das Hauptanalyseinstrument der Arbeit) wurden mit den drei Wellen durchgeführt und die Gütemaße der Modellbeurteilung verglichen. In der zweiten Welle lassen sich sowohl die *higher-order value structure* als auch die von Davidov vorgeschlagene Zusammenlegung von Wertetypen (sieben Faktoren und fünf Kreuzladungen) am besten replizieren[26]. Daher fiel die Entscheidung zugunsten dieser Welle aus.

4.2 Beschreibung der Daten

Da 21 einzelne Items hier nicht im Detail deskriptiv analysiert werden können, sollen lediglich Kennzahlen wie Anzahl der fehlenden Werte, Schiefe und Wölbung angeführt werden (die statistischen Kennzahlen sind dem Anhang zu entnehmen).

Die Voraussetzung für die Annahme der Normalverteilung – diese ist eine notwendige Voraussetzung für viele komplexere Analysemethoden – wird von keinem Item erfüllt. Das zeigen die in SPSS durchgeführten Kolmogorov-Smirnov-Tests auf Normalverteilung (Backhaus et al. 2011). Der überwiegende Teil ist rechtsschief verteilt, lediglich die Items imprichPO, ipfruleCO und ipadvntST weisen eine leicht negative Schiefe (> -0,2) auf und sind damit linksschief.

26 Für die Reihung verließ ich mich auf die Informationskriterien AIC und BIC (mit ihnen können auch nicht geschachtelte Modelle miteinander verglichen werden; geschachtelt ist ein Modell dann, wenn es dieselbe Konfiguration, aber mehr Freiheitsgrade besitzt). Je tiefer der Wert, umso besser gibt das Modell die Daten wieder. Aber auch die übrigen Fitmaße wie χ^2, RMSEA, CFI, TLI und SRMR zeigen dasselbe Bild.

Antworttendenzen

Die Schiefe der Daten bedarf einer Erklärung: Sowohl im SVS als auch im PVQ sind Antworttendenzen gegeben. Manche Teilnehmer neigen vermehrt dazu, sämtliche Items hoch zu bewerten, einige geben generell nur mittlere Bewertungen und manche geben an, überhaupt sehr selten hohe Ähnlichkeiten mit den Porträts aufzuweisen. Dennoch ist auch hier eine Varianz gegeben, die für wahr erachtet wird. Für bestimmte Formen der Analyse empfiehlt Schwartz (vgl. 2003: 275) daher die mittlere Antworttendenz eines Befragten über alle seine Antworten entweder als Kovariate zu behandeln oder jede Antwort mit der mittleren Antworttendenz zu zentrieren (englische Begriffe dafür sind *ipsatizing* und *centering*). Das geschieht, indem er die Itemwerte zentriert, d.h. die Mittelwerte der Items werden für jeden Fall vom Mittelwert des Befragten über sämtliche Items (also die durchschnittliche Verwendung der Skala) abgezogen.[27] Es werden nicht mehr absolute, sondern relative Differenzen berechnet. Strukturgleichungsmodelle – und damit auch die konfirmatorische Faktorenanalyse – korrigieren unterschiedliche Skalenverwendung automatisch und brauchen diese Umrechnung daher nicht (vgl. ebd.).

Es könnte auch sein, dass die Schiefe durch soziale Erwünschtheit zustandekam. Die wird hingegen generell als unproblematisch bei den vorhandenen Wertemessungen betrachtet: „There is evidence from Finland and Israel, moreover, that socially desirable responding does not confound self-reported values“ (Schwartz, 2003: 271). Zwar scheint soziale Erwünschtheit kein Problem zu sein. Dem Thema mit all den noch offenen Kritikpunkten wurde ein eigenes Kapitel gewidmet. Dennoch ist bemerkenswert, dass die Rechtsschiefe einiger Items besonders stark bei gewissen Itemgruppen auftritt: Sowohl Universalism- als auch Benevolence-Items sind neben den Security-Items impsafe und ipstrgv, ipcrtivSD und ipgdtimHE überdurchschnittlich schief (die durchschnittliche Schiefe der 18 Items mit positiven Koeffizienten beträgt 0,65). Mit anderen Worten bedeutet das, dass für diese genannten Items eine hohe Zustimmung unter den Befragten herrscht. Der Grund dafür kann in der generellen Wertestruktur liegen: Diese Werte haben für den Großteil der Schichtungen des Samples denselben hohen Stellenwert. Oder aber die Items sind so formuliert, dass Befragte sich auf Grund der Fragestellung geneigt fühlen, dem Item zuzustimmen. Ob es sich hierbei um den Einfluss von sozialer Erwünschtheit handelt, kann nicht geklärt werden. Klarer hingegen wird das Bild bei den linksschiefen Items: Wie die kognitiven Interviews zeigen, liegt es durchaus an ihrer Formulierung, dass die Antworttendenz zwischen den Kategorien 4 und 6 liegt.

27 Siehe „Computing Scores for the 10 Human Values“: http://ess.nsd.uib.no/ess/doc/ess1_human_values_scale.pdf.

Die Wölbung scheint bei den meisten Indikatoren innerhalb erwartbarer Grenzen. Eine Ausnahme bildet das Item iplylfrBE. Hier beträgt die Wölbung mit 3,047 ein Vielfaches des Durchschnitts. Eine positive Wölbung (leptokurtisch) sagt aus, dass die Verteilung bei dem Item nicht mehr gewölbt ist, sondern spitz zuläuft. Das heißt, die Mehrzahl der Befragten hat sich durchgehend entweder mit „gleicht mir sehr“ oder „gleicht mir“ (siehe Mittelwert) eingestuft, mit kaum Varianz. Dieses Item trennt besonders schwach. Einer der Interviewten nennt einen Grund, warum kaum Varianz bestehen könnte:

> [Das Item ist] vielleicht etwas banal, weil es eine Tautologie ist. Weil … wie soll man denn Freunde oder Angehörige sonst definieren als jemand, um den man sich sorgt, und der sich um einen sorgt … ähm … für den man da ist. Dem man treu ist. (Kogn. Int. Nr. 9, iplylfrBE)

Wie aus Tabelle 14 (Seite 108) erkennbar, besitzt das Item auch keine ausgesprochen guten Messeigenschaften. Seine Eignung für diese Skala ist fraglich.

Fehlende Werte (Missings)

Im Bereich der fehlenden Werte spannt sich das Spektrum zwischen missings=24 (ipgdtimHE; 1,06%) und missings=80 (ipstrgvSE; 3,55%) auf. Damit liegen die fehlenden Werte pro Item alle unter 5%. Dieser Wert gilt als Schwellenwert, wenn weitere Analysen mit komplexen Methoden mit fehlenden Werten gemacht werden (vgl. Schafer und Graham, 2002). Üblich für sozialwissenschaftliche Untersuchungen sind 1 bis 10 % (Schnell et al., 2008: 468). Der vorliegende Datensatz liegt damit innerhalb dieses Referenzrahmens. Mehrere Optionen (*listwise deletion* bis zu multipler Imputation) stehen offen, wenn Analysen mit fehlenden Werten durchgeführt werden. In der vorliegenden Arbeit wird einheitlich die Methode des listenweisen Fallausschlusses für sämtliche Methoden angewendet. Detaillierter werden die fehlenden Werte im Kapitel 4.3.3 analysiert.

4.3 Exploratorische Voruntersuchungen

4.3.1 Exploratorische Faktorenanalyse (EFA)

Die Theorie von Shalom Schwartz gibt eine eindeutige Anweisung vor, in welcher Konfiguration Items auf ihre Wertetypen zu laden haben und wie die Wertetypen miteinander in Beziehung stehen:

> The circular arrangement of the values represents a motivational continuum. The closer any two values in either direction around the circle, the more similar their

> underlying motivations; and the more distant any two values, the more antagonistic their underlying motivations. The motivational significance of items that operatio-nalize adjacent values overlaps in part but differs sharply from that of items that operationalize distant values. This structure is a circumplex, except for the placement of tradition outside conformity (Schwartz und Boehnke, 2004).

Seine Ergebnisse hat Schwartz mittels SSA auf exploratorischem Weg gefunden, die letztendlich auch in einem konfirmatorischem Rahmen getestet wurden. Dennoch scheint aus dem Kapitel 2.3 („Kritische Diskussion“) hervorzugehen, dass die Ergebnisse nicht von jeder Stichprobe gleich gut repliziert werden können. Wie eingangs erwähnt, ist daher zunächst zu testen, ob sich die erwartete Konfiguration auch in den verwendeten Daten reproduzieren lässt.

Um zu sehen, wie sich die Items in der zweiten Welle spontan anordnen würden, wurde mittels SPSS 19 eine Hauptachsenanalyse mit korrelierenden Faktoren[28] berechnet. In einem ESS-Appendix[29] wird beschrieben, wie die Faktoren der EFA zu interpretieren sind: Der erste Faktor der unrotierten Lösung stellt einen Methodenfaktor dar. Er spiegelt die Verwendung der Skala wider (vgl. Verkasalo et al., 2009). Eine detaillierte Beschreibung für seine Interpretation ist nicht vorhanden. Daher wird er nicht weiter thematisiert. Die Ladungen des zweiten und dritten Faktors der unrotierten Lösung hingegen können in einem zweidimensionalen Raum geplottet werden und stellen eine graphische Annäherung an den Quasizirkumplex dar (siehe Abbildung 5). In dem Plot wurden Segmente händisch eingezeichnet. Die vier *higher-order values* lassen sich gut herauslesen. Eine Darstellung mit mehreren Segmenten ergibt sich – wie noch gezeigt wird – mittels multidimensionaler Skalierung (siehe Kapitel 4.3.2, Seite 82). Die rotierte Lösungsmatrix ist dem Anhang (Tabelle 25) zu entnehmen.

28 Die Rotation geschieht nicht wie bei *varimax* mit Faktoren, die normal, sondern mit solchen, die in spitzen Winkeln zueinander stehen können (oblimin-Rotation). Die Theorie von Schwartz lässt ausdrücklich eine Korrelation der Faktoren zu.

29 „Computing Scores for the 10 Human Values“: http://ess.nsd.uib.no/ess/doc/ess1_human_values_scale.pdf

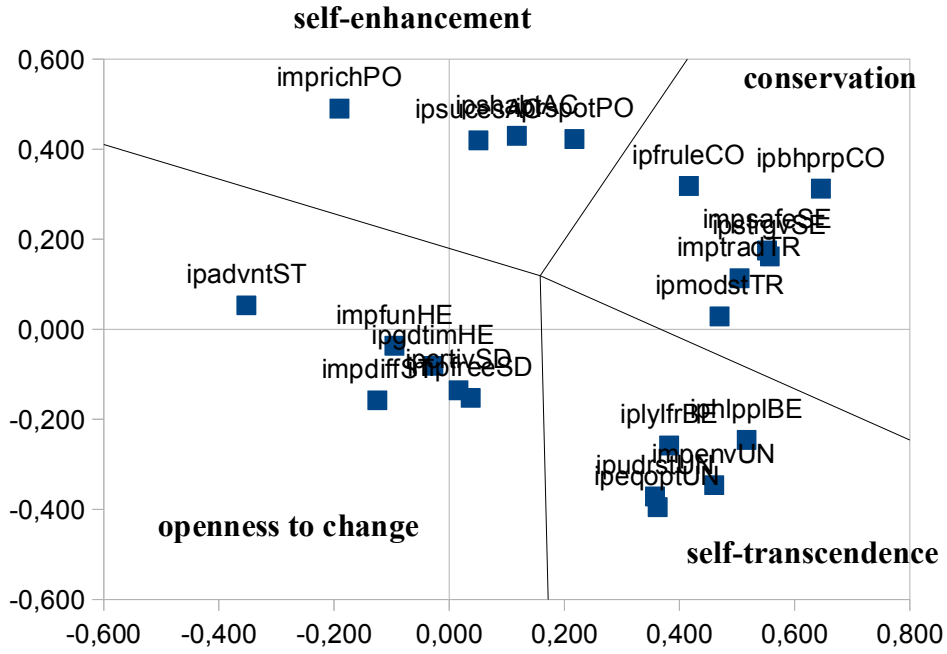


Abbildung 5: EFA geplottet (x-Achse=Faktor 2; y-Achse=Faktor 3)

Um ein objektiveres Bild der Faktorenaufteilung zu erhalten, wurden die Ergebnisse mit jenen verglichen, die das Programm Mplus mit dem ihm eigenen Algorithmus errechnet (Fit-Statistiken für die zweite Welle des ESS für Österreich siehe Tabelle 7). Ohne rechnerische Probleme konnte eine Lösung mit sechs Faktoren ermittelt werden. Für jene mit sieben wurden zwar Faktorladungen ermittelt, allerdings ohne Berechnung der Standardfehler. Berechnungen mit mehr als sieben Faktoren konvergieren nicht. Da die Modelle mit unterschiedlicher Faktorenanzahl nicht ineinander geschachtelt sind, können sie theoretisch nur über ein Informationskriterium miteinander verglichen werden (*sample-size adjusted BIC*). Aber auch die übrigen Fit-Statistiken zeigen ein einheitliches Bild: Mit steigender Faktorenzahl verbessert sich der globale Modell-Fit (CFI, TLI, RMSEA, SRMR).

Der Grundstruktur von Schwartz in der zweiten Welle des ESS entspricht am ehesten eine Lösung der exploratorischen Faktorenanalyse mit 7 Faktoren. Das ist auch die Zahl an Faktoren, die Schmidt et al. (2007) kolportierten und mit denen Davidov in seinen Studien arbeitete. Allerdings ist die Ladungsstruktur der EFA nicht hilfreich für die weitere Testung der Qualität der Items. Werden die Items gemäß ihrer höchsten Ladung den Konstrukten zugeordnet, so ergeben sich Lösungsprobleme für das Programm, denn es werden nicht positiv definite Faktorkorrelationsmatrizen errechnet, die auf problematische Spezifikationen im Modell hinweisen.

	Wertetyp/ Faktorenzahl	Chi2	df	CFI	TLI	RMSEA	SRMR	Sample-size adjusted BIC
Welle ess2at	3	1429,313	150	0,886	0,840	0,065	0,034	125149,226
	4	830,234	132	0,938	0,901	0,051	0,025	124629,973
	5	507,593	115	0,965	0,936	0,041	0,019	124382,724
	6	373,576	99	0,975	0,948	0,037	0,016	124319,662
Kommentar: keine SE berechnet	7	275,096	84	0,983	0,957	-	0,016	124319,662
	8	KEINE KONVERGENZ. ANZAHL DER WIEDERHOLUNGEN WURDE ÜBERSCHRITTEN.						

Tabelle 7: EFA mittels Mplus (geomin-rotiert [oblique 0,5])

Das Beispiel unterschiedlicher Lösungen exploratorischer Faktorenanalysen mit unterschiedlichen Programmen wirft grundsätzlich die Frage auf, ob eine Dimensionsreduktion mit dieser Methode im Fall der *human values* nach Schwartz sinnvoll ist: Der in SPSS gebräuchliche Algorithmus extrahiert den ersten Faktor mit einem Maximum an Varianz. Was an Varianz durch diesen Faktor nicht erklärt wird, wird in einem weiteren Schritt durch den zweiten Faktor maximal erklärt etc. Dieses Vorgehen entspricht nicht völlig der Vorstellung von der dynamischen Struktur der Werte. Die Anzahl an Faktoren, die dadurch extrahiert werden, ist wahrscheinlich verzerrt. Mplus hingegen scheint mit seinem Algorithmus den Kern der Theorie besser wiederzugeben (es können bis zu sieben Faktoren extrahiert werden – ein Ergebnis, das durch die Analysen in den Davidov-Studien untermauert wird, auch wenn die darin angewandte konfirmatorische Faktorenanalyse mit dem SEM-Programm AMOS durchgeführt wurde). Ladungen der Mustermatrix[30] des SPSS-Ausdrucks stimmen mit der mittels Mplus errechneten Vier-Faktoren-Lösung im Großen überein. Das bedeutet aber auch, dass prinzipiell nur die Anzahl der extrahierten Faktoren vom Algorithmus betroffen ist. In der Vier-Faktoren-Lösung bei beiden Programmen stimmen die Größenordnungen der Faktorkovarianzen überein (siehe Tabelle 8). Unterschiede ergeben sich in der Polung der Zusammenhänge. Die Pearsonschen Korrelationskoeffizienten sind gegengleich bis auf die Ausnahmen *change–self-transcendence* und *self-enhancement–conservation*. Laut der Theorie von Schwartz sollten Konstrukte, die einander im motivationalen Kontinuum gegenüberliegen, dies auch durch relativ hohe negativen Korrelationen zeigen. Da es sich allerdings um eine exploratorische Analyse und nicht um eine konfirmatorische handelt und daher auch Annahmen nicht getestet werden, wird dieser Punkt an dieser Stelle nicht weiter behandelt. Dies

30 Das ist die Ladungsmatrix der rotierten Lösung.

wird im Teil „konfirmatorische Faktorenanalyse“ nachgeholt. Für Ladungsmatrizen der Mplus- und SPSS-Lösung siehe Anhang.

Mplus-Matrix	*change*	*self-enhancement*	*conservation*	*self-transcendence*
change	1,000			
self-enhancement	-0,340	1,000		
conservation	0,198	0,064	1,000	
self-transcendence	-0,184	0,014	0,179	1,000

SPSS-Matrix	*change*	*self-enhancement*	*conservation*	*self-transcendence*
change	1,000			
self-enhancement	0,442	1,000		
conservation	-0,266	0,073	1,000	
self-transcendence	-0,255	-0,032	-0,209	1,000

Tabelle 8: Faktorkorrelationsmatrix Mplus – SPSS

Die Faktorkorrelationen der EFA sind im Sinne einer räumlichen Darstellung, wie sie von der schwartzsche Theorie eingefordert wird, nicht sinnvoll interpretierbar.

4.3.2 Multidimensionale Skalierung (MDS)

Schwartz erzielte gute Ergebnisse, was die räumliche Darstellung seiner Theorie betrifft, mit SSA (smallest space analysis; korrespondierende Funktion in SPSS: multidimensionale Skalierung, kurz: MDS) , auch wenn Steinmetz et al. (k.A.) für eine objektive Einschätzung des zirkulären Zusammenhangs problematische Vorannahmen dieser Methode kritisieren.

Dennoch ist MDS eine der Methoden, die in der bisherigen Analyse der Daten häufig Verwendung fand (vgl. Schwartz, bis 2003; Schmidt, 2007; Mohler und Wohn, 2005). Schwartz erklärt sein Vorgehen mit MDS wie folgt:

> My criteria are based on considering the whole continuum of values by performing multidimensional scaling and examining the configuration of items in a two-dimensional space (2 dimensions are usually sufficient to obtain a good representation of the associations, though 3 may rarely be required) […]. The SSA programs usually yield the clearest structures. (Schwartz, 2003: 286)

Sowohl Saris und Knoppen (2009 a; b) als auch Beierlein et al. (2011) unterließen in ihren Analysen eine vorherige Untersuchung der „dynamischen“ Struktur des Werteuniversums in ihren Stichproben, als sie die Wertetypen mit ihren angrenzenden Nachbarn verglichen. Die Annahme, dass sich die Wertestruktur wie postuliert ergibt, hat sich – wie ich zeigen konnte – bisher nicht eindeutig in jeder Stichprobe reproduzieren lassen können. Der Vergleich zwischen Nachbarn, die gar keine sind, kann zu verzerrenden Schlüssen führen.

Daher wird in einem nächsten Schritt ein solches Modell mit den Daten gerechnet und die Ergebnisse mit jenen von Mohler und Wohn (2005) verglichen. Im Detail wird auf die Analyse nicht eingegangen, da sie einzig dem Zweck dient, eine graphische Darstellung des Wertekontinuums zu erzeugen. Die Güte der Modelle wird daher nicht interpretiert.

MDS wurde jeweils einmal für zentrierte (siehe Anhang, Abbildung 8) und einmal für nicht zentrierte (siehe Anhang Abbildung 7) Daten gerechnet, um den Unterschied im Ergebnis zu demonstrieren. In beiden Fällen verwendete ich die im Statistikprogramm SPSS enthaltene Prozedur ALSCAL.[31]

ALSCAL liefert keine Einteilung in Segmente. Diese muss wie auch bei dem EFA-Plot händisch eingefügt werden. Die Ziehung der Linien erfolgte gemäß der Anweisungen von Schwartz.

> A possible approach to defining a region is that a line joining two items for the same value should not cross the line joining the items from any other value. If the items do indeed form a region, reflecting their own inter-correlation and their similar pattern of correlations with other items, I would treat them as an acceptable index for a value (Schwartz, 2003: 286).

Zwischen zentrierter und unzentrierter Version fällt auf, dass letztere nur sieben, die erste hingegen acht Segmente unterscheidet. Stimulation geht in der unzentrierten Version in den Segmenten von Achievement und Hedonism auf. Dies lässt sich durch die folgenden Analysen mittels CFA nicht nachvollziehen. Ansonsten ist die strukturelle Anordnung der von Schwartz postulierten relativ ähnlich. Wie schon in der Diskussion um die Skala erwähnt, sind die Wertetypen Universalism und Benevolence genauso wie Tradition und Conformity nicht voneinander unterscheidbar. Sie erscheinen auch in der graphischen Darstellung als jeweils kombinierte Segmente. Power und Achievement sind gemäß Grafik separate Wertetypen. Ein ähnliches Ergebnis liefert die CFA. Allerdings haben gerade diese beiden Wertety-

31 Schmidt et al. (2007) verwenden ebenso ALSCAL von SPSS. Im Gegensatz zu jener Studie verwendete ich statt Produktmomentkorrelationen Rohdaten.

pen gewissermaßen einen Sonderstatus, weil sie genau an der Grenze zwischen eigenständigen Konstrukten liegen, die aber eine sehr große Ähnlichkeit vereint.

Weitere Abweichungen von der Theorie sind:

- SE, das in der Grafik zwischen Universalism-Benevolence statt zwischen Conformity-Tradition und Power liegt. Dieses Ergebnis wurde auch von Mohler und Wohn (2005) für Österreich und andere Länder gefunden.
- In der zentrierten Lösung haben die Wertetypen Stimulation und Hedonism im Vergleich zur Theorie ihre Plätze getauscht, und damit grenzt laut MDS Stimulation an Achievement und nicht an Hedonism.

4.3.3 Missing-Analyse

„Das Hauptproblem bei der Analyse von Datensätzen mit fehlenden Werten besteht darin, dass die Ursachen für das Fehlen mit inhaltlichen Aspekten der Untersuchung in systematischem Zusammenhang stehen, die Daten also meist nicht ‚zufällig' fehlen" (Schnell et al., 2008: 468). Es wurde bereits erwähnt, dass soziale Erwünschtheit keine merkbare Erklärung in der Varianz leistet. Es ist nicht auszuschließen, dass dennoch verzerrende Effekte durch den Ausfall von Befragten eintreten könnten. Generell wird in der Literatur zu fehlenden Werten zwischen drei Formen von *missings* unterschieden (Schnell et al., 2008; Schafer und Graham, 2002):

- MCAR: *missing completely at random.* Das Fehlen kann durch keine bekannte Variable erklärt werden. Das Muster der fehlenden Werte ist völlig zufällig.
- MAR: *missing at random.* Es gibt bekannte Variablen aus dem Datensatz, die einen Einfluss auf das Fehlen haben.
- MNAR: *missing not at random*. Der Ausfall wird ausschließlich durch die Variable selbst erklärt.

Um MAR auszuschließen, wurden die fehlenden Fälle mit den Kontextvariablen Geschlecht, Alter und Höhe der formalen Bildung in Kreuztabellen miteinander verglichen. Da die Fallzahlen so gering sind (maximal 80), wurden höhere Methoden wie Regressionen ausgeschlossen. Einzig die Variablen ipsucesAC und iprspotPO produzierten mit der Kontextvariable Bildung signifikante χ^2-Statistiken. Formale Bildung wirkt sich somit auf die Anzahl der ungültigen Antworten in den genannten Variablen aus. Das Zusammenhangsmaß *Cramers V*, das

verwendet wird, wenn Variablen mit nominalem Skalenniveau verglichen werden und zumindest eine Variable mehr als zwei Ausprägungen hat (Akremi et al., 2011), liefert allerdings sehr bescheidene Maße für die Ausprägung des Zusammenhangs. Für beide Porträts gilt CramersV<0,1.[32]

Um festzustellen, in welchem Ausmaß die Variable selbst einen Einfluss auf die Anzahl der Missings hat, wurde die Anzahl der fehlenden Werte jedes Items mit dem Durchschnitt aller Missings über alle Items verglichen (siehe Anhang, Tabelle 22). Überdurchschnittlich hohe fehlende Werte produzieren die Items iprspotPO, ipudrstUN, impsafeSE, imptradTR, iplylfrBE, ipfruleCO und ipbhprpCO. Aufgrund der geringen absoluten Anzahl lassen sich – wie gesagt – keine inferenzstatistischen Tests berechnen. Um dennoch einen Schritt Richtung Objektivität zu erreichen, wurden die Ergebnisse mit ähnlichen Grundgesamtheiten[33] verglichen. Der ESS umfasst zwei weitere deutschsprachige Länder: Deutschland und die Schweiz (hier wurde nur der deutschsprachige Teil der Stichprobe für die Analyse verwendet). Der Tabelle 9 sind Überschneidungen zwischen den Stichproben zu entnehmen.

überdurchschnittlich in 3 Stichproben	*überdurchschnittlich in 2 Stichproben (eine davon Österreich)*	*überdurchschnittlich nur in Österreich*
ipfrule (CO)	ipudrst (UN)	imptrad (TR)
ipstrgv (SE)	ipbhprp (CO)	
iprspot (PO)	iplylrf (BE)	

Tabelle 9: Überschneidung in drei Stichproben (AT, CH und DE) von Items mit überdurchschnittlich vielen fehlenden Werten

4.4 Konvergierende Methodentriangulation

Die nachfolgende Analyse basiert auf dem Vorgehen von Saris und Knoppen (2009a; b). Sie analyiserten die Vollversion des PVQ mittels CFA. Geprüft wurden sowohl die Messmodelle für die einzelnen Wertetypen als auch die Zusammenhänge mit unmittelbar benachbarten Wertetypen. Hinweise für Missspezifikationen liefern die SEM-Programme basierend auf den Modifikationsindizes. Es wurde allerdings bemängelt (Saris et al., 2009), dass diese χ^2-basierten Maße nicht die Sensitivität des Tests berücksichtigen. Dazu fehlt die Berechnung der

32 Da dieses Ergebnis keinen zusätzlichen Informationsgehalt beinhaltet, wird es nicht im Anhang erwähnt.
33 Ich beschränke mich in diesem Fall auf die Homogenität der offiziellen Landessprache.

Power. Knoppen und Saris (2009b) verwenden deshalb das Programm JRule, das den Ausdruck des SEM-Programms einliest und mit Hilfe von Parametern, die durch den Anwender vorgegeben werden, die Power berechnet. Auf Grund dieser Information liefert das Programm Vorschläge für die Interpretation der Missspezifikation (die Alternativen sind: missspezifiziert, nicht missspezifiziert und unbestimmt). Daher kommt auch der Name: Judgment Rule. Die Anwendung von JRule behielt ich für die Analyse bei. Bei den vorliegenden Daten kann aber schon jetzt gesagt werden, dass nur in den seltensten Fällen diese Zusatzinformation eine Änderung in der Interpretation der vorgeschlagenen Modifikation mit sich brachte. Das heißt, dass die von Mplus vorgeschlagenen Modifikationen fast immer auf signifikante Missspezifikationen hinwiesen, auch unter Berücksichtigung der Power des Tests.

Knoppen und Saris (2009b) gingen in ihrer Analyse schrittweise vor:

1. Sie errechneten den Fit eines Modells für alle 40 Items mit den zehn postulierten Faktoren.
2. Sie prüften jedes Messmodell (also jeden Wertetyp mit seinen Items).
3. Sie prüften jede Kombination aus Zwei-Faktoren-Modellen, die sich durch die Nachbarschaft in dem Zirkumplex ergeben können: Jedes Konstrukt wurde mit dem unmittelbar vorhergehenden und nachfolgenden separat getestet.

Dieses Vorgehen ermöglichte ihnen auf den verschiedenen Ebenen Probleme in der Messung zu erkennen (z.B. unvorhergesehene Subdimensionen oder Kreuzladungen auf „fremde“ Wertetypen). Schwartz hatte seine Theorie auf die Anwendung durch CFA konkretisiert:

> Note that the theory postulates a circular arrangement of the 10 values, not of the items. For items, it postulates that each item correlates more highly with the set of items that measure the same value than with the set of items that measure a different value. Thus, in technical terms, the theory assumes that the items in the value survey form 10 latent factors and only the factors relate to one another in a circular manner (Schwartz und Boehnke, 2004).

Der ersten Schritt eines globalen Modells lieferte – wie bereits durch andere Autoren vorgeführt (Schmidt et al., 2007) – Hinweise, dass das Modell nicht zu den Daten passt. Die Folgeschritte ermöglichen ins Detail der Analyse vorzustoßen.

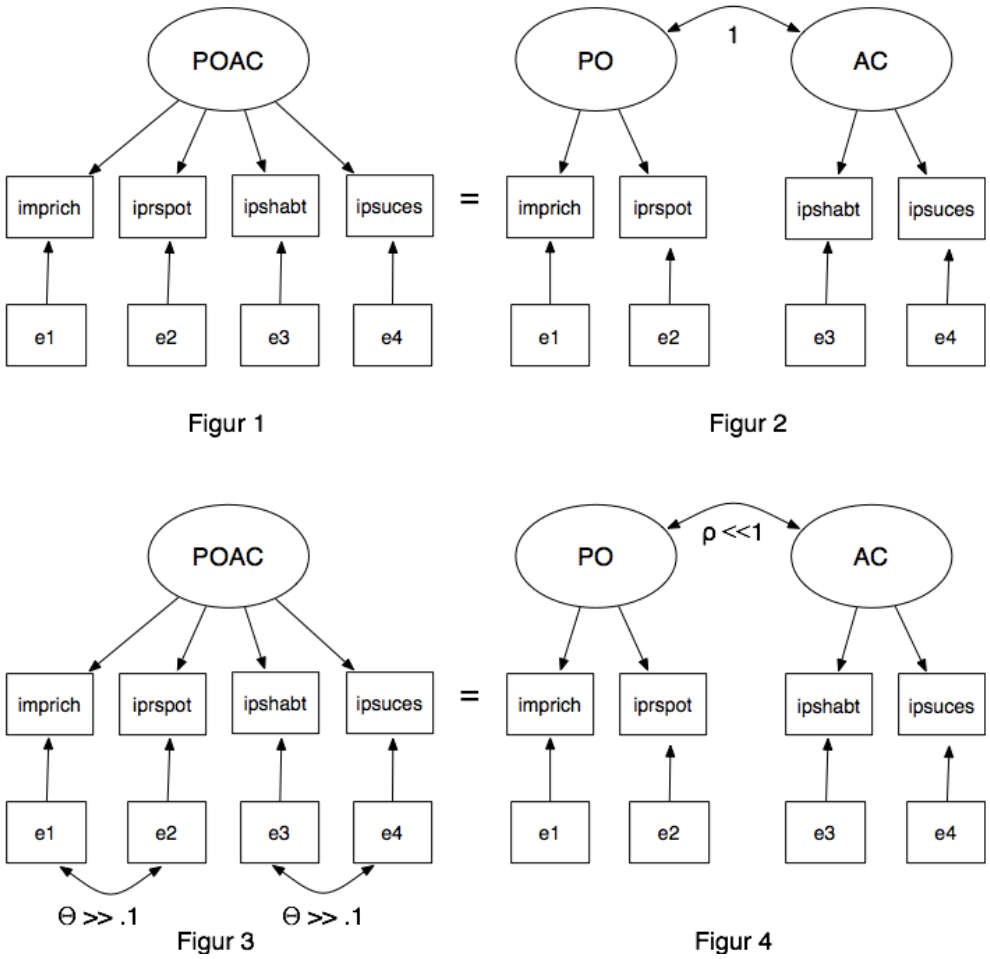


Abbildung 6: Äquivalente Repräsentationen (aus: Knoppen und Saris, 2009b)

Im zweiten Schritt lassen sich die Messmodelle daraufhin prüfen, ob das als einheitlich definierte Konstrukt mitunter aus mehreren Subdimensionen besteht. Knoppen und Saris konnten damit prüfen, ob es sich bei den zehn postulierten Wertetypen auch um ebenso viele handelte, oder ob sich nicht durch die Items selbst neue Dimensionen ergaben, die erfasst wurden, aber vorher in der Operationalisierung nicht definiert worden waren. Diese neuen Dimensionen lassen sich aus den Modifikationen herauslesen (Knoppen und Saris, 2009b: 6): Wenn die Korrelation zwischen den Fehlervarianzen zweier Items eine signifikante Verbesserung bringt, dann deutet das auf eine Dimension hin, die nur diese zwei Items miteinander verbindet. Es können zwei neue Messmodelle definiert werden, die miteinander korrelieren (siehe Figur 3 und 4 in Abbildung 5). Trifft es dann auch zu, dass die Faktorkorrelation deutlich von 1 abweicht, kann in diesem Fall von mindestens zwei Dimensionen ausgegangen werden. Damit sind beide Versionen (Figur 3 und 4) äquivalente Repräsentationen voneinander (Knoppen und Saris, 2009b: 7).

Der dritte Schritt prüft über die Modifikationsindizes (kurz: MI), ob sich signifikante Kreuzladungen ergeben, wenn die Modelle paarweise mit ihren jeweiligen Nachbarn getestet werden. Aus dem Zitat von Schwartz wird deutlich, dass Kreuzladungen gegen die theoretische Annahmen sprechen, denn theoretisch sollten die Items nur über die Wertetypen miteinander in Verbindung stehen. Die direkte Ladung eines Items auf einen benachbarten Faktor ist eine Verknüpfung, die nicht vorgesehen ist.

Für die vorliegende Analyse werden die drei Schritte von Knoppen und Saris unter Verwendung von JRule verwendet. Aufgrund der geringen Anzahl an Items (21) können für den zweiten Schritt nur die vier *higher-order values* als Konstrukte dienen. Damit sind auch die speziellen Ergebnisse der MDS für den analysierten Datensatz in der Auswertung nicht relevant, weil sich die aufgefundenen Unregelmäßigkeiten innerhalb der *higher-order values* auftauchten. Die Beurteilung der Messmodelle wird auch keine „neuen“ Dimensionen finden können. Stattdessen können Annahmen über die Zusammengehörigkeit von Wertetypen in der Zusammensetzung des ESS und damit die Annahmen von Schmidt und Kollegen (2007) und der Davidov-Studien überprüft werden. Die vier Messmodelle sind wie folgt definiert:

openness to change SDSTHE	self-enhancement POAC	conservation SECOTR	self-transcendence UNBE
gemessen durch	*gemessen durch*	*gemessen durch*	*gemessen durch*
ipcrtiv	ipshabt	impsafe	ipeqopt
impfree	ipsuces	ipstrgv	ipudrst
ipadvnt	imprich	ipfrule	impenv
impdiff	ipsrpot	ipbhprp	iphlppl
ipgdtim		ipmodst	iplylfr
impfun		imptrad	

4.4.1 Beurteilung der Messmodelle

Da die Grundannahmen der Normalverteilung für die vorliegende Stichprobe nicht gegeben sind, wurden die Parameter mittels MLR geschätzt. Um die ausgewiesene χ^2-Statistik richtig interpretieren zu können, muss sie mit dem *scaling factor* multipliziert werden. Die angegebenen χ^2-Werte sind deshalb korrigiert. Zur Beurteilung dienen jene globalen Fit-Statistiken, die Mplus in der Standardeinstellung ausgibt. Diese sind CFI, TLI, RMSEA und SRMR. Das sind in der Standardliteratur gebräuchliche Werte.

Ein Modell mit den sieben Faktoren ergibt den folgenden Fit:

N	korrigierter χ^2	df	CFI	TLI	RMSEA	SRMR
2022	1535,963	168	0,877	0,846	0,057	0,060

Der RMSEA-Wert ist überraschend gut, im Gegensatz zu den übrigen Fit-Maßen. Es genügt allerdings nur ein Blick auf die Liste der vorgeschlagenen Modifikationen, und es wird klar, dass grobe Anpassungen gemacht werden müssten, um das Modell an die Daten anzupassen. Solche Spezfikationen nachträglich einzuführen ist prinzipiell eine Möglichkeit. Sie ist allerdings datengetrieben und nicht durch weitere Überlegungen fundiert. Analytisch richtiger wäre es, die Messmodelle einzeln zu betrachten.

Eine zusammenfassende Gesamtbeurteilung aller Einzelmodelle zeigt, dass mit der zweiten Welle des ESS der PVQ-Items sieben Wertetypen von einander unterschieden werden können. Das entspricht den bisherigen Annahmen, ebenso die Art, wie die vier *higher-order factors* in Subdimensionen aufgespalten werden können (Zahl in Klammer), wobei wir sehen werden, dass das nicht immer so präzise möglich ist:

- *openness to change* (3)
- *conservation* (2)
- *self-transcendence* (1)
- *self-enhancement* (1)

Das Ergebnis ist keine Überraschung und beinhaltet auch keine Neuigkeit. Dennoch liefert eine detailliertere Betrachtung der Messmodelle und deren Abweichungen von der Theorie Hinweise, an welchen Stellen konkret Probleme auftreten. Daher nun zu den Ergebnissen der Messmodelle im Detail.

Openness to change

Das Ausgangsmodell weist folgende Charakteristik auf:

N	*korrigierter Chi2-Wert*	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
2179	314,551	9	0,916	0,860	0,107	0,044

Den individuellen Anteil am korrigierten χ^2-Wert und damit an Abweichungen leisten Hedonism mit χ^2=75,4, Self-direction mit χ^2=140,1 und Stimulation mit χ^2=99,1. Um diese Werte würde sich jeweils der korrigierte χ^2-Wert verbessern, würden die Items der Wertetypen aus dem Modell ausgeschlossen. Damit scheinen für das Messmodell *openness to change* die Self-direction-Items am problematischsten zu sein, da sie für die relativ höchste Abweichung verantwortlich sind. Sie haben auch die geringsten Ladungen mit $\lambda \approx 0{,}5$. Wie sich noch zeigen wird, stellt das Messmodell *openness to change* besondere Anforderungen an die Theorie.

Das Messmodell für den *higher-order value* passt nicht gut mit den Daten zusammen. Das zeigen alle Fit-Statistiken. Mplus schlägt deshalb Modifikationen vor, um das Modell besser an die Daten anzupassen. JRule weist nur „unbestimmte" Modifikationen aus (die Power liegt über dem standardmäßig vorgesehenen festgesetzten Maß von 0,8). Es kann somit nicht mit ausreichender Wahrscheinlichkeit ausgeschlossen werden, dass es sich bei der Modifikation um eine fälschlich signifikante handelt. JRule sieht als Entscheidungsgrundlage für diesen Fall vor, die erwartete Änderung des Parameters (engl. *expected parameter change;* kurz EPC), wenn dieser Parameter frei geschätzt wird, als Entscheidungskriterium dafür heranzuziehen, ob die Modifikation signifikant ist oder nicht. Alle EPC liegen über dem Limit von 0,1 (dieser Wert wurde von mir vorgegeben, da eine Änderung im standardisierten Parameter ab 0,1 als signifikant angenommen wird).[34] Damit sind sämtliche Modifikationen (in diesem Fall handelt es sich aufgrund eines einfachen Messmodells mit einer latenten Variable nur um Fehlerkovarianzen) laut JRule tatsächliche Missspezifikationen. Dieses Muster zieht sich im Folgenden durch die gesamte Analyse: „unbestimmte" Modifikationen werden trotz ihrer hohen Sensitivität fast immer als tatsächliche definiert, sofern ihr standardisierter EPC größer als 0,1 ist. Im Folgenden wird darauf verzichtet, die Ausnahmen zu erwähnen. Alle genannten „unbestimmten" Entscheidungen haben immer einen standardisierten EPC > 0,1. Werden Modifikationen von JRule wegen ihrer niedrigen Power als tatsächliche eingestuft, werden diese auch so definiert.

Welche der elf vorgeschlagenen Modifikationen nun als erste zugelassen werden soll, muss der Anwender selber entscheiden. Die höchste Veränderung im EPC und χ^2-Wert bringt die freie Schätzung der Fehlerkovarianz zwischen ipgdtimHE und impfunHE. Die Korrelation dieser beiden Items ist (siehe Anhang, Tabelle 20) überdurchschnittlich hoch und alle Messmodelle zeigen überdurchschnittlich hohe und ähnlich starke Ladungen für sie. Mit ih-

34 Das Programm „JRule for Mplus" beinhaltet leider keine Angabe der standardisierten geschätzten Parameter. Daher muss die Entscheidung immer mit dem Mplus-Ausdruck abgeglichen werden, der diesen Parameter beinhaltet.

nen zu beginnen hat daher methodische Gründe. Das nun angepasste Modell weist noch immer mangelhafte Fit-Indizes auf. JRule erkennt die restringierte Fehlerkovarianz zwischen ipadvntST und impdiffST als einzig klar missspezifizierten Parameter. Wird diese zugelassen, so verbessert sich der Fit, allerdings erreicht das Gesamtmodell noch immer nicht zufriedenstellende Fit-Maße.

Wird auch noch die Fehlerkovarianz zwischen den Self-direction-Items (vorgeschlagene Modifikation) zugelassen, dann hat das *higher-order value* Modell noch immer keinen ausreichenden Fit erreicht (CFI=0,964; TLI=0,911; RMSEA=0,085). Es werden von JRule noch 12 tatsächliche Missspezifikationen gefunden, manche mit einem standardisierten EPC von über 0,9. Es handelt sich dabei ausschließlich um Kreuzladungen jedes Faktors mit sämtlichen „fremden" Items. Dieser Befund stimmt mit dem mangelhaften globalen Fit überein.

Wird das Modell in seine äquivalente Repräsentation überführt, bleibt der Fit derselbe. Hinzu kommt allerdings die Information der Faktorkorrelationen, die für SD mit ST ρ=0,757, SD mit HE ρ=0,738 und HE mit ST ρ=0,803 beträgt. Damit liegen sie signifikant unter 1. Die drei Wertetypen sind somit – unter der falschen Annahme, dass das Modell an sich korrekt wäre – stark korrelierende Subdimensionen des Faktors *openness to change.* Der globale Fit deutet an, dass das Modell mit drei stark korrelierenden Faktoren besser wäre als ein Ein-Faktor-Modell, aber immer noch nicht gut genug, um für sich stehen zu können. Es ist generell problematisch, Faktoren nur durch zwei Indikatoren zu messen. Aus diesem Grund wird für die Analyse der paarweisen Überprüfung im anschließenden Kapitel bevorzugt mit den *higher-order values* gearbeitet, wobei berücksichtigt wird, dass es sich dabei um eine Missspezifikation des Modells handelt.

Faktor	*Item*	*standardisierte Ladung*
ST	impdiff	0,751
	ipadvnt	0,672
SD	ipcrtiv	0,612
	impfree	0,609
HE	ipgdtim	0,785
	impfun	0,789

Self-transcendence

N	*korrigierter Chi2-Wert*	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
2155	17	5	0,994	0,989	0,025	0,013

Dieses Messmodell liefert die eindeutigsten Ergebnisse: Es scheint fast perfekt zu den Daten zu passen. Mplus weist keine MIs aus. Das bedeutet, dass Universalism- und Benevolence-Items auf der Ebene der Messung dasselbe messen. Versucht man dennoch die Konstrukte ent-

weder durch eine frei geschätzte Fehlerkovarianz zwischen den jeweils zusammengehörenden Items zu unterscheiden, so können wichtige Parameter des Modells (robuster χ^2 und Standardfehler der Schätzungen, engl. *standard error*, kurz: SE) nicht geschätzt werden. Werden nur partiell Fehlerkovarianzen zugelassen, so sind diese so niedrig (std. $\theta<<0{,}1$; p-value>0,05), dass man auch gar nicht von einer spezifischen Kovarianz sprechen kann.

Die höchste Ladung besitzt das Item impenvUN. Knoppen und Saris (2009b) beschreiben in ihrer Analyse, dass gerade jenes Item die niedrigste Reliabilität hat. Darüber hinaus würden alle drei Universalism-Items so unterschiedliche Dimensionen messen, dass sie theoretisch nicht durch einen gemeinsamen Faktor repräsentiert werden könnten. Zusammen mit den Benevolence-Items lässt sich aber sehr gut der *higher-order value* darstellen.

Faktor	*Item*	*standardisierte Ladung*
UNBE	ipudrst	0,576
	impenv	0,624
	iplylfr	0,567
	iphlppl	0,608
	ipeqopt	0,584

Conservation

N	*korrigierter Chi2-Wert*	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
2098	142,604	9	0,943	0,905	0,074	0,034

Das Modell passt annähernd an die Daten. Der JRule-Output liefert widersprüchliche Entscheidungsgrundlagen: Zwar gibt es eine als solche definierte Missspezifikation, allerdings wäre eine andere „unbestimmte“ prinzipiell vorzuziehen, weil sie einen weitaus größeren Beitrag zur Verringerung der Abweichungen leistet. Dabei handelt es sich um die Fehlerkovarianz zwischen impsafeSE und ipstrgvSE. Wird diese zugelassen, zeigen die Fit-Statistiken eine hinreichend gute Anpassung des Modells an die Daten an. Das äquivalente Modell mit einem Faktor Security und einem kombinierten Faktor COTR kann ohne Probleme berechnet werden. Die Korrelation der Konstrukte ist mit $\rho=0{,}777$ ausreichend unähnlich, um die Konstrukte im Modell getrennt wahrzunehmen.

Faktor	*Item*	*standardisierte Ladung*
COTR	ipfrule	0,533
	ipbhprp	0,723
	ipmodst	0,504
SE	imptrad	0,518
	impsafe	0,578
	ipstrgv	0,588

Self-enhancement

N	*korrigierter Chi2-Wert*	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
2159	11,456	2	0,995	0,984	0,040	0,012

Dieses Modell hat einen ähnlich hohen Fit wie s*elf-transcendence*. Alle Fit-Indizes liegen unter bzw. über den beschriebenen Schwellenwerten. Mplus gibt keine MI aus. Allerdings beinhaltet das Modell ein Item, das trotz idealem Modell-Fit hinter der „Performance" aller übrigen Indikatoren zurückbleibt: iprspotPO ist das einzige Item, das bisher ein $\lambda<0,5$ aufweist. Damit liegt es noch über dem von Brown (2006) vorgeschlagenen Kriterium einer minimalen standardisierten Ladung von $\lambda=0,4$ (das wäre eine Reliabilität von 0,16, sprich: nur 16 % der Gesamtvarianz im Messmodell können durch das Item erklärt werden), aber relativ gesehen ist es ein auffälliger Befund.

Auch hier können auf Grund der nahezu perfekten Anpassung an die Daten nicht beide Fehlerkovarianzen gleichzeitig geschätzt, geschweige denn ein äquivalentes Modell getestet werden. Partiell sind Fehlerkorrelationen für beide einzelnen Paare ohne rechnerische Probleme möglich. Aber auch hier gilt, dass die geschätzten Kovarianzen nicht signifikant sind. Demnach kann s*elf-enhancement* nicht in seine Unterdimensionen aufgeteilt werden.

Faktor	*Item*	*std. Ladung*
PO	imprich	0,531
	iprspot	0,488
AC	ipshabt	0,751
	ipsuces	0,753

4.4.2 Vorläufige Synopsis

Zur Identifikation hervorstechender Items werden noch einmal jene zusammengefasst, die sowohl besonders hohe ($\lambda>0,7$) als auch besonders niedrige ($\lambda\approx0,5$) Ladungen aufweisen:

Hoch	*std. Ladung*	*Niedrig*	*std. Ladung*
impdiff	0,751	ipmodst	0,524
ipgdtim	0,785	imptrad	0,517
impfun	0,789	imprich	0,531
ipbhprp	0,751	iprspot	0,488
ipshabt	0,751		
ipsuces	0,753		

Tabelle 10: Vorläufige Einteilung relativ guter und mangelhafter Indikatoren

4.4.3 Paarweise Prüfung der Überlagerungen der Messung

		M.I.	*std. EPC*			*M.I.*	*std. EPC*
	self-transcendence - openness to change				*conservation - self-transcendence*		
1	SELFTRAN by ipadvnt(ST)	108,385	-0,273	4	CONS by iphlppl(BE)	42,987	0,191
2	SELFTRAN by ipcrtiv(SD)	59,677	0,213	5	SELFTRAN by ipfrule(CO)	33,787	-0,173
3	SELFTRAN by impfree(SD)	49,558	0,193				
	openness to change - self-enhancement				*self-enhancement - conservation*		
6	CHANGE by iprspot(PO)	34,242	-0,189	10	CONS by iprspot(PO)	88,555	0,248
7	ipadvnt(ST) with imprich(PO)	95,792	0,264	11	SELFENH by ipmodst(TR)	108,528	-0,274
8	SELFENH by ipcrtiv(SD)	22,098	0,150	12	ipbhprp(CO) with iprspot(PO)	26,822	0,154
9	CHANGE by ipsuces(AC)	20,627	0,154	13	CONS by imprich(PO)	34,383	-0,153

Tabelle 11: Paarweiser Vergleich benachbarter higher-order factors*; Modifikationen nummeriert*

		M.I.	*std. EPC*			*M.I.*	*std. EPC*
	openness to change - conservation				self-transcendence – self-enhancement		
14	CONS by ipadvnt(ST)	107,297	-0,255	19	SELFTRAN by imprich(PO)	144,31	-0,301
15	CHANGE by ipmodst(TR)	50,756	-0,182				
16	CHANGE by ipstrgv(SE)	77,073	0,221				
17	ipadvnt(ST) with impsafe(SE)	61,127	-0,213				
18	ipgdtim(HE) with impsafe(SE)	26,490	0,150				

Tabelle 12: Paarweiser Vergleich gegenüberliegender higher-order factors*; Modifikationen nummeriert*

Die paarweise Prüfung aller sieben identifizierbaren Wertetypen beträgt 21 Kombinationen. Dies sprengt den Rahmen der Arbeit. Daher werden nur die paarweisen Überlagerungen der vier *higher-order values* analysiert. Es werden solche Modifikationen berücksichtigt, wo sich der Betrag des standardisierten Parameters um mehr als 0,15 ändert. Eine Schwelle von 0,2 wäre prinzipiell adäquater. Doch dann wäre womöglich die Vorauswahl zu eng. 0,15 scheint ein gangbarer Kompromiss zu sein. In Tabelle 11 und Tabelle 12 sind die ausgewählten Modifikationen zusammengefasst. Zur Lesart derselben sei gesagt, dass die Terme „by“ Faktorladungen und „with“ (Fehler-)Kovarianzen bezeichnen.

Tabelle 13 fasst die Faktorkorrelationen (oder das, was Davidov, Schmidt und Schwartz (2008) *pattern of distances* nannten), die sich aus den paarweisen Vergleichen ergeben, zusammen. Sie stimmen mit denen aus einem Modell mit vier Faktoren überein. Die Hauptachsen zwischen *openness to change* und *conservation* einerseits und *self-enhancement* und *self-transcendence* andererseits beschreiben annähern die Theorie, indem sie den stärksten Gegensatz ausdrücken. Relativ geschieht das beim ersten Paar. Das zweite ist mit $\rho \approx 0$ weniger passend, auch wenn dieser Zusammenhang gegensätzlicher ist als mit den jeweils benachbarten Paaren. Wie im Kapitel der exploratorischen Faktorenanalyse angesprochen, sind die dort erwähnten Faktorkorrelationen nicht sinnvoll interpretierbar. Das ist in einer konfirmatorischen Analyse anders. Und der Befund hier lautet, dass die Theorie und die Daten nicht vollständig übereinstimmen.

		ρ
change with	selfenh	0,536
	selftrans	0,291
	cons	-0,254
selfenh with	cons	0,153
	selftrans	0,051 (SE=0,037)
cons with	selftrans	0,383

Openness to change – self-transcendence

Anhand der standardisierten Residuen, der Korrelationsmatrizen und der kognitiven Interviews werden die vorgeschlagenen Modifikationen (Tabelle 11 und Tabelle 12) schrittweise erläutert.

Korrelationen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
ipcrtiv(SD)	0,205	0,200	0,205	0,178	0,163
impfree(SD)	0,169	0,179	0,190	0,268	0,108
impdiff(ST)	0,182	0,096	0,161	0,163	0,072
ipadvnt(ST)	0,014	-0,047	-0,011	-0,014	-0,062
ipgdtim(HE)	0,162	0,093	0,116	0,247	0,089
impfun(HE)	0,100	0,071	0,075	0,158	0,060
Residuen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
ipcrtiv(SD)	6,127	5,336	6,081	4,356	3,567
impfree(SD)	3,983	4,234	5,085	7,099	1,059
impdiff(ST)	3,785	-1,330	2,655	3,135	-2,349
ipadvnt(ST)	-4,204	-7,866	-5,621	-6,277	-7,912
ipgdtim(HE)	1,957	-2,317	-0,583	6,106	-2,099
impfun(HE)	-1,310	-3,613	-3,039	1,930	-3,589

Modifikationen 2 und 3: Die Self-direction-Items weisen weitaus höhere Korrelationen mit impenvUN und iphlpplUN auf als der Rest der *openness-to-change*-Items. Im Modell werden diese Beziehungen unterschätzt. Die positiven standardisierten Residuen so wie auch die Kreuzladungen deuten darauf hin.

Die Korrelationsmatrix verrät, dass Korrelationen zwischen den SD- und den Universalism-Benevolence-Items über alle Items hinweg in einen gemeinsamen Rahmen fallen, die Kor-

relationen der Stimulation- und Hedonism-Items allerdings dahinter zurückbleiben. Hier könnte es sich um den spezifischen Fall handeln, dass diese Missspezifikation behoben werden könnte, wenn die Self-direction-Items aus dem Konstrukt *openness to change* herausgehoben werden und einen eigenen Faktor bilden.

Modifikation 1: Die Kovarianz zwischen den Universalism-Benevolence-Items und ipadvntST ist durchwegs gering. Mit impenvUN und iphlpplBE gleicht sie sich der von impdiffST und den Hedonism-Items an. Das Modell überschätzt die Korrelationen mit diesem Item, was durch die negative Kreuzladung ausgeglichen wird. Für ipadvntST gilt, dass es besonders wenig Varianz mit den Universalism-Benevolence-Items teilt. Unter Berücksichtigung der kognitiven Interviews ließen sich daraus zwei Schlüsse ziehen: *Zum einen* scheinen gewisse Universalism-Benevolence-Items einer Konvention zu entspringen (siehe Modifikation 4), bestimmten Normen zu folgen, und auch die *higher-order values conservation* und *self-transcendence* korrelieren stark positiv miteinander (siehe Tabelle 13). Wenn also ipadvntST *conservation* relativ krass entgegengesetzt ist, und *conservation* und *self-transcendence* miteinander stark korrelieren, dann sollten auch ipadvntST und *self-transcendence* relativ viel Gegensätzlichkeit beinhalten (siehe auch die Modifikationen 14). *Zum anderen* ist ipadvntST eines der linksschiefen Items. Das bedeutet, dass Befragte tendenziell dieses Portrait als unähnlich bewerteten. Hingegen konnten sie sich mit jenen der *self-transcendence*-Items besonders gut identifizieren (relativ niedrige Mittelwerte). Welcher der beiden Erklärung der Vorzug zu geben ist, lässt sich bei derzeitigen Wissenstand nur auf Grund von Plausibilität bestimmen: Da die erste Erklärung Korrelationen auf der latenten Ebene miteinbezieht, sind Zusammenhänge auf der Ebene der Indikatoren nicht besonders klar zuordenbar, und Schlüsse lassen sich schwer ziehen. Die zweite, methodisch bedingte Erklärung scheint daher nahe liegender (mehr zu den linksschiefen Items siehe Kapitel 4.4.5).

Conservation – self-transcendence

Korrelationen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
ipfrule(CO)	0,050	0,044	-0,012	0,007	0,108
ipbhprp(CO)	0,079	0,134	0,084	0,111	0,234
ipmodst(TR)	0,146	0,137	0,098	0,020	0,179
imptrad(TR)	0,108	0,205	0,061	0,113	0,241
impsafe(SE)	0,099	0,142	0,159	0,173	0,222
ipstrgv(SE)	0,101	0,191	0,139	0,218	0,233
Residuen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
ipfrule(CO)	-2,804	-4,078	-6,173	-5,263	-0,730
ipbhprp(CO)	-3,976	-2,215	-4,138	-2,207	4,215
ipmodst(TR)	1,794	0,927	-0,571	-4,353	3,109
imptrad(TR)	-0,200	4,195	-2,708	0,104	6,520
impsafe(SE)	-1,487	0,129	1,501	2,307	4,425
ipstrgv(SE)	-1,498	2,491	0,347	4,292	5,146

Die Modifikation 4 zeigt, dass das Modell den Zusammenhang von iphlpplBE mit den *conservation*-Items unterschätzt. Wie die folgenden Zitate der kognitiven Interviews zeigen, scheint die Motivation, anderen zu helfen, stark aus einer von Normen geregelten Haltung geprägt zu sein, die gemeinhin als sozial erwartbar wahrgenommen wird (der Text des Items iphlpplBE ist: „Es ist ihm wichtig, den Menschen in seiner Umgebung zu helfen. Er möchte sich um ihr Wohlergehen sorgen.“)

> I: „Was, glaubst du, ist denn die Intention dahinter, warum die Person den anderen helfen möchte; sich um ihr Wohlergehen sorgt?“
> R: „…[…] unvoreingenommen kann man helfen, ohne was dafür zu verlangen. […] Aber ich denke, in einer Freundschaft ist es doch auch so, dass es immer eine Gegenseitigkeit ist. […] dass Freunde für einander da sind, und dass man hilft und natürlich auch Hilfe erwartet, wenn man sie selbst braucht. (Kogn. Int. Nr. 3, iphlpplBE)

> I: „Wieso hilfst du?“
> R: „Wieso helfe ich? Weil das, glaube ich, ein Akt der Nächstenliebe ist, glaube ich, und jeder muss sich gegenseitig helfen. Man ist ja nicht allein auf der Welt – oder? Dafür hat man ja Freunde.“ (Kogn. Int. Nr. 5, iphlpplBE)

> I: [I. klärt einen Sachverhalt, der vom Befragten angesprochen wurde] „Ein unbewusster Antrieb wäre, dass man sich danach gut fühlt, weil man das gemacht hat?“
> R: „Ich meine, man lernt ja dazu. Wenn eine alte Frau auf der Strasse umfällt, dann frage ich sie nicht, ob sie sich bei mir bedanken wird, wenn ich ihr aufhelfe und überlege auch nicht: Wie werde ich mich fühlen? Sondern ich – wie du sagst – mach’s einfach. Ich helfe ihr auf!“ (Kogn. Int. Nr. 9, iphlpplBE)

Normen werden in der Soziologie definiert als: „ … a shared expectation of behaviour that connotes what is considered culturally desirable and appropriate.“ (Scott und Marshall, 2009)

Die Befragten sprechen Erwartungen an, die sogar über den direkten Freundes- bzw. Bekanntenkreis hinausgehen. Die sind nicht nur durch enge Beziehungen entstanden, sondern können durchaus auch auf offener Straße „eingefordert“ werden. Die Befragten scheinen sich hier auf eine Norm zu berufen, die „gewisse Gefälligkeiten“ anbietet, durchaus mit der Erwartung, ebensolche zu erhalten. Normen sind ungeschriebene Regeln, deren relative Wichtigkeit vom *higher-order value conservation* erfasst wird. Der relativ große Zusammenhang ließe sich dadurch erklären. Auch Schwartz und Bardi sprechen diese Verquickung von Werten und Normen an: „Values motivate behavior, but compete with normative pressures“ (Hitlin und Piliavin, 2004: 365). In diesem Fall war die Messung davon direkt betroffen.

Modifikation 5: Der tatsächliche Zusammenhang zwischen *self-transcendence*-Items und dem spezifischen Item ipfruleCO ist geringer als erwartet (mit Ausnahme der Korrelation mit iphlpplBE). Jener Indikator ist, wie auch ipadvntST, ein relativ „schwieriges“ Item und damit linksschief, im Gegensatz zu den *self-transcendence*-Items, mit deren Porträts sich besonders viele Befragten identifizieren konnten. Relativ gesehen vereint damit dieses Item mehr Gegensätzlichkeit – was auf ein methodisches Problem zurückzuführen ist – als die übrigen.

Openness to change – self-enhancement

Modifikation 6 zeigt, dass das Modell den Zusammenhang von iprspotPO und *openness-to-change*-Items überschätzt. IprspotPO hat einen Anteil an einer Form von Motivation, die sich auf *conservation* zurückführen lässt. Dazu liefern die kognitiven Interviews Einblicke:

> I: „Es ist ihr wichtig, von anderen respektiert zu werden. Sie möchte, dass die Leute das tun, was sie sagt.“
> R: „Ähm, das passt für mich jetzt nicht so zusammen, weil ‚Respekt‘ finde ich, ist etwas, das möchte ich schon haben. Das bringe ich auch anderen Leuten entgegen.“ (Kogn. Int. Nr. 3, iprspotPO)

> I: „Was heißt „Respekt“ für dich? Auf gleicher Augenhöhe?“
> R: „Ja, auf jeden Fall. […] bei uns in der Firma haben wir auch eine Putzfrau, die kaum Deutsch kann. Und trotzdem respektier ich sie. Weißt du, ich grüße sie höflich. Ich meine, das macht man so einfach, weil man es tut. Ja? Und das ist wurscht, ob das jetzt die Putzfrau ist, oder der Herr Doktor Irgendwie ist. Das ist eine Art von Respekt.“ (Kogn. Int. Nr. 5, iprspotPO)

> I: „Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt.“
> R: „[…] Also das ist für mich unstimmig. Vielleicht deswegen, weil für mich „Respekt“ etwas anderes bedeutet als – also, das ist mir schon wichtig, dass mir Leute mit Respekt zuvorkommen, weil ich auch den anderen Respekt zeige – aber Respekt in dem Sinne, dass man halt … heißt ja nicht, dass man dem anderen

dann hörig ist, oder sich unterordnet, sondern Respekt einfach, dass man auf gleicher Ebene sich … begegnet.“ (Kogn. Int. Nr. 6, iprspotPO)

Korrelationen	imprich(PO)	iprspot(PO)	ipshabt(AC)	ipsuces(AC)
ipcrtiv(SD)	0,177	0,056	0,260	0,318
impfree(SD)	0,091	0,121	0,219	0,272
impdiff(ST)	0,201	0,013	0,205	0,266
ipadvnt(ST)	0,369	0,114	0,211	0,324
ipgdtim(HE)	0,227	0,110	0,243	0,314
impfun(HE)	0,227	0,107	0,233	0,320
Residuen	imprich(PO)	iprspot(PO)	ipshabt(AC)	ipsuces(AC)
ipcrtiv(SD)	1,662	-3,144	3,798	6,728
impfree(SD)	-2,970	-0,160	1,454	3,806
impdiff(ST)	0,596	-7,162	-2,706	-0,785
ipadvnt(ST)	11,015	-1,902	-1,683	4,953
ipgdtim(HE)	0,753	-3,511	-2,497	0,154
impfun(HE)	0,758	-3,803	-3,194	0,553

Ähnlich wie bei iphlpplBE geht es bei „Respekt“ darum, dass man Respekt von vornherein anderen entgegenbringt, aber man davon ausgeht, dass dies auf Gegenseitigkeit beruht. Auch hier steht eine soziale Norm in den Antworten der Befragten im Hintergrund, die in den Beantwortungsprozess einfließt. Um zu den Korrelationen zurückzukommen, die tiefer als erwartet liegen: *Openness to change* und *conservation* sind gegenüberliegende motivationale Konstrukte und korrelieren damit wenig positiv miteinander. Wenn iprspotPO motivationale Anteile von *conservation* in sich vereint, dann ist die geringere positive Korrelation mit *openness-to-change*-Items damit erklärbar.

Modifikation 7 stellt eine Herausforderung dar, denn oberflächlich gesehen scheint nichts darauf hinzudeuten, dass den Items ipadvntST und imprichPO eine gemeinsame Dimension zugrunde liegt, die über ihre latenten Faktoren hinausgeht. Nun könnte es sich in diesem Fall um ein Artefakt handeln, das wegen der falschen Spezifikation des Faktors *openness to change* zustande gekommen ist. Allerdings käme auch jene Interpretation in Frage, dass es sich hier um zwei linksschiefe Items handelt und die gemeinsame Dimension über die methodische Verknüpfung zustande kommt. Die Ausdrucke der paarweisen Vergleiche zeigen, dass jede Kombination der linksschiefen Items ipadvntST, imprichPO und ipfruleCO Modifikationen hervorruft, allerdings ist die Modifikation 7 nach den hier aufgestellten Kriterien die einzig relevante. Die erwarteten Parameter mit std. EPC≈0,1 für die übrigen Modifikationen ist nicht hoch genug, um in die Vorauswahl zu fallen.

Die Modifikationen 8 und 9 hängen beide mit *openness to change* zusammen und sind daher unter der Annahme zu betrachten, dass die Modifikation aufgrund der Missspezifikation des *higher-order values* zustande kommt. Tatsächlich gibt es in der Lösung mit sieben Faktoren keine Modifikation für eine Kreuzladung zwischen *self-enhancement* und ipcrtivSD. Mit

einem std. EPC von 0,150 liegt sie auch genau am Grenzkriterium für relevante Modifikationen. Diese zwei Gründe sind ausschlaggebend, sie aus der weiteren Betrachtung auszuschließen. Für die Kreuzladung zwischen *openness to change* und ipsucesAC hingegen gibt es in der Sieben-Faktoren Lösung Kreuzladungen zwischen jenem Item und den drei Wertetypen Stimulation, Self-direction und Hedonism. Diese moderate Anpassung muss allerdings, so wie sie ist, stehen gelassen werden, denn im qualitativen Material ist über diesen Zusammenhang nichts zu finden. Eine weiterführende Auslegung ist daher nicht möglich.

Self-enhancement – Conservation

In Modifikation 10 zeigt sich, was bereits bei Modifikation 6 angesprochen wurde:

Korrelationen	imprich(PO)	iprspot(PO)	ipshabt(AC)	ipsuces(AC)
ipfrule(CO)	0,072	0,193	0,085	0,090
ipbhprp(CO)	-0,025	0,242	0,091	0,069
ipmodst(TR)	-0,172	0,003	-0,105	-0,136
imptrad(TR)	-0,081	0,177	0,021	-0,004
impsafe(SE)	-0,015	0,180	0,147	0,060
ipstrgv(SE)	-0,001	0,158	0,134	0,128
Residuen	imprich(PO)	iprspot(PO)	ipshabt(AC)	ipsuces(AC)
ipfrule(CO)	1,353	6,794	1,272	1,481
ipbhprp(CO)	-4,168	8,919	0,588	-0,858
ipmodst(TR)	-9,365	-1,428	-7,480	-8,732
imptrad(TR)	-5,498	6,126	-2,014	-3,050
impsafe(SE)	-2,903	6,386	4,415	-0,364
ipstrgv(SE)	-2,345	5,507	3,960	3,305

iprspotPO korreliert (verglichen mit den übrigen *self-enhancement*-Items) unverhältnismäßig hoch mit *conservation*-Items. Die Kreuzladung balanciert dieses Ungleichgewicht zwischen Modellannahmen und Daten aus. Eine Ausnahme ist allerdings mit dem Item ipmodstTR vorhanden (Modifikation 11). Dieses korreliert als einziges *conservation*-Item (mit der Ausnahme der Korrelation mit iprspotTR) negativ mit den *self-enhancement*-Items. Aus den Protokollen dieser Items kristallisieren sich Schlagwörtern wie Selbstdarstellung, etwas erreichen wollen und Dominanz heraus. Diese stehen dem, was für iprspotTR zutage gefördert wird, motivational gegenüber, wie die folgenden Ausschnitte zeigen:

> I: „Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken."
> R: „4 … weil es ist ja genau das Gegenteil von der anderen Frage, ob ich immer im Vordergrund sein will, wenn man das so nimmt. […] Also, ob ich dann immer die Initiative ergreife, und … das hat für mich ehrlich gesagt keinen Zusammenhang. Weil: Im Vordergrund sein und Initiative ergreifen, das ist ja: dominant sein. Und … präsent sein, oder sowas." (Kogn. Int. Nr. 8, ipmodstTR)

I: „Es ist ihr wichtig, zurückhaltend und bescheiden zu sein. Sie versucht, keine Aufmerksamkeit auf sich zu lenken."
R: „Also ich mein allein schon dadurch, dass ich an der Open Acting Academy [einer Schauspielschule in Wien] bin, sieht man, dass ich Aufmerksamkeit eigentlich schon … bis zu einem gewissen Grad will. Also vor allem, wenn ich was gut kann, dann will ich auch, dass man das sieht! … Ja, und Bescheidenheit bringt einen nur relativ selten irgendwohin, bin ich draufgekommen." (Kogn. Int. Nr. 10, ipmodstTR)

I: „Was glaubst du, wieso sie das machen würde? Woran glaubt diese Person vielleicht?"
R: „Ja, es gibt so Personen, die das einfach brauchen. Die … ich weiß nicht: ich habe mich jetzt noch nie damit auseinandergesetzt, aber … weiß nicht … ist vielleicht auch irgendwie eine Selbstbestätigung. Dass man irgendwie … – vielleicht ist da auch eine gewisse Unsicherheit dabei, dass man denkt, dass die Person denkt: Ich muss eine Bestätigung bekommen, damit ich selber für mich weiß, dass ich toll bin. Und das nicht einfach so hinnehmen kann, ohne großartig aufzufallen." (Kogn. Int. Nr. 6, ipmodstTR)

Zur Stellung des Items innerhalb des Fragebogens wäre anzumerken, dass iprspotTR fünf Items nach „Es ist ihm wichtig, seine Fähigkeiten unter Beweis zu stellen. Er möchte, dass ihn Leute für das bewundern, was er tut." (ipshabtAC) folgt. Respondent Nr. 8 scheint sich auf dieses Item zu beziehen. Es stellt sich die Frage, wie stark der Halo-Effekt des Vorgänger-Items tatsächlich ist. An dieser Stelle kann darüber allerdings nur spekuliert werden.

Modifikation 12 zeigt eine Fehlerkovarianz zwischen ipbhprpCO und iprspotPO an. Diese zwei Items scheinen einen Zusammenhang zu teilen, der unabhängig von den latenten Konstrukten ist. Er könnte dadurch zustande gekommen sein, dass jedes der Items besonders ausgeprägt zwei verschiedene Dimensionen erfasst, die in ihrer Motivation gegenläufig sind (beide messen sozusagen jeweils zwei unabhängige Stimuli). Aus den Interviews auf Seite 98 lässt sich für das Item iprspotPO herauslesen, dass Streben nach „Respekt" wenig damit zu tun hat, dass man will, dass andere das tun, was man ihnen sagt. Beide Dimensionen werden von den Befragten spontan angesprochen. Ähnlich verhält es sich bei iphprpCO:

I: „Es ist ihr wichtig, ein anständiges Leben zu führen. Sie möchte alles vermeiden, was Leute als Fehltritt bezeichnen könnten."
R: „Ja. Ja. Das ist dann 1, also das gleicht mir sehr. - Also, bzw. ohne den 2. Teil: egal, was die Leute sagen. Das ist mir nicht wichtig.
I: „Und wie bewertest du das dann?"
R: „Wenn nur der erste Satz gewesen wäre, dann würde ich sagen: Ja, das ist Antwort 1: gleicht mir sehr. Aber mich stört der 2.[…] Weil, was haben die Leute damit zu tun? […] Nur damit die Nachbarn das Maul nicht aufreißen, oder dass die, äh, Freundinnen jetzt sagen: Wie schaut denn DIE aus! Oder: was tut SIE? - Das kümmert mich überhaupt gar nicht." (Kogn. Int. Nr. 4, ipbhprpCO)

I: „Es ist ihm wichtig, ein anständiges Leben zu führen. Er möchte alles vermeiden, was Leute als Fehltritt bezeichnen könnten."
R: „… Hm. … Eigentlich wieder: der eine Satz: sehr gut. Der andere … eher weniger. Also, ähm, was andere Leute als Fehltritt … - also, ganz ehrlich: es ist mir egal, was andere Leute von mir halten. […]"
I: „Und, wie gleichst du einer Person, die beide Sätze vertritt? […] kannst du da eine Antwort darauf geben?"
R: „Eigentlich nicht. Nein, weil, wie gesagt, für mich sind das, für mich passt das gerade nicht so sehr zusammen. [lacht] Weil, ich ein anständiges Leben habe, aber es ist mir vollkommen egal, was die anderen von mir halten." (Kogn. Int. Nr. 8, ipbhprpCO)

I: „Es ist ihr wichtig, ein anständiges Leben zu führen. Sie möchte alles vermeiden, was Leute als Fehltritt bezeichnen könnten."
R: „Nein. Das würde ich sagen, das gleicht mir zwar etwas, dass ich sag: grundsätzlcih anständig zu sein, z.B. nicht mit einem verheirateten Mann unbedingt mir eine dauerhafte Beziehung aufs Aug' drücken. Das würd ich schon sagen. […] Also, ich hab schon auch wirklich Fehltritte gemacht, wenn du sagst: gesellschaftlich. Das ist mir schon passiert. Also so grundsätzlich ist mir das [nämlich alles zu vermeiden] zu radikal jetzt. Ich finde, du lebst ja auch nicht wirklich, wenn du nicht manchmal sagst: ok, drücken wir ein Auge zu, gehe ich halt da drüber. Wenn du z.B. du verliebst dich in eine Frau und die ist verliebt in dich. Die hat zwei Kinder und ist verheiratet. Ja? Aber ich bin dann halt schon so anständig, dass ich sage: ich würde nie irgendwelchen Kindern den Vater oder die Mutter da irgendwie streitig machen. Und deswegen hör ich dann auf. Nach ein paar Wochen vielleicht." (Kogn. Int. Nr. 11, ipbhprpCO)

Respondentin 11 ist zwischen zwei Werten, die für sie beide in diesem Item zum Tragen kommen, hin und her gerissen. Einerseits sieht sie es als richtig an, sich „grundsätzlich" anständig zu verhalten (in diesem „grundsätzlich" schwingt vielleicht mit, dass sich niemand als unanständig bezeichnen würde). Andererseits bedeutet das nicht, dass Fehltritte nicht zu ihrem Leben dazugehören. Das Ziel anständig zu sein, ist erstrebenswert, nicht jedoch Fehltritte zu vermeiden. Ähnlich verhält es sich bei den anderen beiden Befragten: ein anständiges Leben ist erstrebenswert, nicht jedoch, alles zu vermeiden, worüber Mitmenschen ihre Meinung abgeben könnten. Welcher Stimulus wird letztlich bewertet?

Doppelte Stimuli sind ein methodisch begründetes Phänomen: abgesehen von jeweils zwei Sätzen, die in den Portraits des PVQ verwendet werden, vermitteln sie im Speziellen bei diesen zwei Beispielen auch zwei unterschiedliche Stimuli. Befragten sind gezwungen, Auswege aus dem Dilemma zu finden. In den Interviews vermitteln sie auch ihre Strategien: teilweise geschieht das, indem Antworten gemittelt werden, teilweise enthalten sich Befragten einer Antwort. IpbhprpCO ist auch ein Item, das überdurchschnittlich viele Missings produziert genauso wie iprspotPO (siehe 4.3.3).

ImprichPO (Modifikation 13) wiederum zeigt ein ähnliches Muster wie ipmodstTR in Modifikation 11: Dieser Indikator korreliert als einziges *self-enhancement*-Item schwach negativ mit den *conservation*-Items mit Ausnahme von ipmodstTR und ipfruleCO. Es ist in der Qualität seiner Motivation in Summe weiter entfernt von *conservation* als die übrigen *self-enhancement*-Items. Der Zusammenhang mit ipmodstTR wurde bereits erklärt. Zur Erklärung des einzigen leicht positiven std. Residuums mit ipfruleCO lässt sich sagen, dass das Item mit imprichPO die Messeigenschaft teilt, dass es linksschief ist. Dadurch werden automatisch positivere Zusammenhänge entstehen als mit anderen Items. Gleichzeitig ist es auch automatisch über einen negativen Zusammenhang mit den übrigen Items verknüpft. Auf das Item imprichPO wird noch im Detail auf Seite 112 eingegangen.

Openness to change – conservation

Im Quasizirkumplex gegenüberliegende *higher-order values* sollten der Theorie nach gegensätzliche Motivationen beschreiben, die sich in den Korrelationen der Wertetypen wiederfinden sollten. Die Paare *openness to change* und *conservation* illustrieren das mit einer ausgeprägten negativen Korrelation ρ=-0,254. Ein relativ großer Teil an Gegensätzlichkeit (ca. 24%) kennzeichnet das Item ipadvntST. Wird dieser Pfad (Modifikation 14) zugelassen, verringert das die Korrelation von ρ=-0,254 auf ρ=-0,193. Bekanntermaßen ist ipadvntST ein linksschiefes Item. Die std. Residuen geben allerdings ein sehr diffuses Bild sowohl von positiven als auch von negativen Residuen. Sie machen es dadurch schwer die Modifikation sowohl inhaltlich als auch methodisch genauer zu interpretieren.

Korrelationen	ipfrule(CO)	ipbhprp(CO)	ipmodst(TR)	imptrad(TR)	impsafe(SE)	ipstrgv(SE)
ipcrtiv(SD)	-0,060	-0,125	-0,156	-0,053	-0,020	-0,014
impfree(SD)	-0,110	-0,117	-0,143	-0,057	0,001	0,061
impdiff(ST)	-0,113	-0,184	-0,203	-0,103	-0,092	0,002
ipadvnt(ST)	-0,124	-0,264	-0,251	-0,211	-0,260	-0,132
ipgdtim(HE)	-0,093	-0,107	-0,174	-0,052	0,016	0,069
impfun(HE)	-0,120	-0,123	-0,173	-0,040	-0,049	0,067
Residuen	ipfrule(CO)	ipbhprp(CO)	ipmodst(TR)	imptrad(TR)	impsafe(SE)	ipstrgv(SE)
ipcrtiv(SD)	-1,040	-3,731	-1,850	-0,653	1,075	1,369
impfree(SD)	-3,393	-3,472	-0,936	-0,850	2,196	4,929
impdiff(ST)	-3,042	-7,113	-2,488	-2,524	-1,680	3,137
ipadvnt(ST)	2,762	-1,887	1,156	-2,041	-4,174	4,256
ipgdtim(HE)	-1,820	-1,925	1,354	0,412	4,354	7,129
impfun(HE)	-3,183	-3,112	1,285	0,968	0,954	6,866

Für das Tradition-Item ipmodst kann mit Blick auf die Korrelationsmatrix gesagt werden, dass es verglichen mit den übrigen *self-transcendence*-Items überdurchschnittlich stark negativ mit *openness to change*-Items zusammenhängt (Modifikation 15). Gleichzeitig zeigen die

standardisierten Residuen nur relativ geringe Abweichungen. Davon liegt eine einzige außerhalb des Konfidenzintervalls. Für andere Items, z.B. ipbhprpCO, weist das Modell weitaus höhere Abweichungen auf, die aber weniger relevant zu sein scheinen. Auch diese widersprüchlichen Befunde machen es schwierig, eine Fährte zu finden, die eine schlüssige Erklärung für diese Kreuzladung liefern könnte. Eine mögliche wäre die folgende: die kognitiven Interviews zeigten bei Modifikation 11, dass ipmodstTR in seiner Motivation besonders stark den *self-enhancement*-Items gegenübersteht (von den Befragten wurde besonders „etwas erreichen wollen", also ein klassischer Wert, der zu Achievement passt, hervorgestrichen). Direkte Verbindungen zu Hedonism, Stimulation oder Self-direction können nicht herausgelesen werden. Hingegen könnten die Kovarianzen der Faktoren zur Erklärung herangezogen werden: O*penness to change* und *self-enhancement* stehen sich in ihrer Motivation sehr nahe (ρ=0,536). *Openness to change* und *conservation* (zu dem auch ipmodst gehört) korrelieren mittelstark negativ (ρ= -0,254). Es könnte sein, dass die Gegensätzlichkeit von ipmodstTR zu *self-enhancement* und damit in weiterer Folge zu *openness to change* nicht vollständig über die negative Kovarianz zwischen *conservation* und *openness to change* erklärt werden kann und das Modell daher angepasst werden muss.

Die Security-Items impsafe und ipstrgv sollten laut Theorie wie auch die übrigen *conservation*-Items dem Faktor *openness to change* in ihrer Motivation gegenüberstehen. Die empirischen Zusammenhänge lassen sich aber nicht vollständig mit dieser Annahme vereinen, da die Security- Items tendenziell gar nicht oder sogar leicht positiv mit *openness to change* korrelieren.[35] In diesem Modell ist ipstrgvSE weitaus positiver in seinem Zusammenhang, als es das Modell voraussagen würde (Modifikation 16). Diese Modifikation scheint auf den ersten Blick widersprüchlich, zumal man annehmen sollte, dass ein Bedürfnis nach Sicherheit immer auch Einschränkungen beinhaltet. Sicherheit entsteht auf Kosten eines anderen Gutes, im heutigen Diskurs zum Thema „Terrorismus" sehr oft auf Kosten der bürgerlichen Rechte, und würde damit dem Streben nach Selbstbestimmung widersprechen. Auch der Zirkumplex nach Schwartz weist darauf hin, indem *openness to change* und *conservation* einander gegenüberliegen. Aus den kognitiven Interviews lässt sich allerdings herauslesen, dass ein Ausdruck des Porträts („starker Staat") Interpretationspielräume öffnet. Neben einer faschistischen Konnotation für manche Befragten beinhaltet er auch, dass der Staat für ein gedeihliches Zusammenleben garantiert, was auch von selbstbestimmt eingestellten Befragten gewünscht wird.

35 Hier sollten die Ergebnisse, die die Security-Items betreffen, mit Vorsicht betrachtet werden, weil die Analyse der Messmodelle Security als eigenen Faktor bestätigte. Modellanpassungen sollten immer auch in diesem Licht betrachtet werden.

Zu berücksichtigen ist dabei, dass es sich um ein Phänomen handeln kann, das kulturspezifisch auf jene Länder zutrifft, wo ein Wohlfahrtsstaat, wie in Österreich, neben sozialem Ausgleich auch für bürgerliche Freiheiten steht und das im öffentlichen Diskurs keinen Widerspruch darstellt:

> I: „Es ist ihm wichtig dass der Staat seine Sicherheit gegen alle Bedrohungen wahrt. Er möchte einen starken Staat, der seine Bürger beschützen kann."
> R: „Für mich ist die Sicherheit, die durch den sozialen Staat gegeben ist, wichtig. Und ich hoffe, dass sie auch so bleibt und nicht zum Negativen sich ändert. […] Der starke Staat: weiß ich nicht, ob der nur beschützt. […] Mir gefällt also der starke Staat nicht. Der starke Staat ist für mich schon … eine etwas harte Gesellschaftsform, die mich zwingt […] Wenn ich sagen würde: ein fürsorglicher Staat, dann würde mir das besser gefallen. Oder ein verantwortungsvoller Staat würde mir noch besser. Ein Staat, eine Staatführung, die Verantworung für ihre Mitbürger nimmt. […]
> I: „Wie ähnlich ist dir jemand, der das glaubt?"
> R: „Der ist mir … schon, die Sicherheit ist wichtig, aber die Stärke tut mir weh. … puhuu. Jetzt weiß ich nicht, was überwiegt. Es überwiegt … eher die Sicherheit und die Stärke lasse ich weg. 3. Gleicht mir etwas." (Kogn. Int. Nr. 2, ipstrgvSE)

> I: „Es ist ihr wichtig, dass der Staat ihre Sicherheit gegen alle Bedrohungen wahrt. Sie möchte einen starken Staat, der seine Bürger beschützen kann."
> R: „Gleicht mir sehr! [strahlt] Das kommt nämlich auch von der Zeit, wie ich in Ungarn gelebt habe. Weil dort – damals war ich ja noch jung – da habe ich das Gefühl gehabt, dass der Staat mich, äh, beschützt. […] und, und – Ja! Doch! Und … ja, hier in Österreich selbst habe ich das Gefühl ständig auch. Aber das sind die, die Wurzeln, wo es wichtig war, dass man die Möglichkeit hatte zu studieren und überhaupt. Ja. […]"
> I: „Geht es da ums Gemeinwohl oder eher um deine persönliche Sicherheit?"
> R: „Es geht um meine persönliche Sicherheit, auch. Aber, äh, die Vorraussetzung ist natürlich auch, dass das Gemeinwohl auch gewährleistet wird."(Kogn. Int. Nr. 4, ipstrgvSE)

Die Respondentin Nr. 4 wuchs im sozialistischen Ungarn auf und ist daher durch diesen Hintergrund geprägt, der auch hier zum Ausdruck kommt; sie lebt seit 1978 in Österreich. Die Form des Schutzes, den sie beschreibt, wirkt in diesem Falle eher wie die Garantie für einen gewissen Lebensstandard, z.B. auf höhere Bildung nicht verzichten zu müssen.

> I: „Was ist für dich ein starker Staat?"
> R: „Für mich, ein starker Staat ist einer, der Geld hat, der … wenig Arbeitslose, großes Bruttoinlandsprodukt. Solche Sachen. Wo man gut leben kann. Guter Lebensstandard. Wo man Arbeit kriegt, eine gute Ausbildung kriegt. Eigentlich wie hier [Österreich], wenn man es so nimmt. Das ist für mich ein starker Staat." (Kogn. Int. Nr. 8, ipstrgvSE)

> I: „Es ist ihm wichtig, dass der Staat seine Sicherheit gegen alle Bedrohungen wahrt. Er möchte einen starken Staat, der seine Bürger beschützen kann."

> R: „Gleicht mir. […] Äh, will ich in einem Staat leben, der sicher ist? Ja, natürlich will ich das. Äh, die Frage, wie ich Sicherheit definiere, ist dann die andere Sache.[Aber hier in Wien] also, puh, kannst rausfahren, kannst die Türen offenlassen. Kannst die Kinder im Garten spielen lassen. Brauchst keine Angst haben, dass irgendein Heckenschütze sie über den Haufen schießt. […] Also das ist schon verständlich, aber das Bild, vor dem habe ich Angst … das Bild das mitschwingt, ist ein bisschen so: „Ausländer raus, das Gsindl". […] „brauchen wir jetzt nicht" … So, Strache würde so etwas, das ist eine typische Sprachformulierung." (Kogn. Int. Nr. 1, ipstrgvSE)

Die Antworten reflektieren auch, dass die Sicherheit nicht unbedingt bedingungslos gegeben ist. Allerdings schwebt, wie bei Respondent Nr. 2, auch immer das Ideal mit, dass es sich auch um einen „fürsorglichen" Staat handeln könnte[36], der sich nicht gegen die eigenen Bürger wendet, sondern bestimmte Leistungen erst ermöglicht (Respondentin 4). Damit soll unterstrichen werden, dass der erwartbare negative Zusammenhang zwischen ipstrgvSE und *openness to change* durch diese Konnotation abgeschwächt werden könnte, weshalb das Modell durch eine Kreuzladung korrigiert werden muss.

Modifikationen 17 und 18: das Item impsafeSE scheint jeweils mit ipadvntST und ipgdtimHE methodisch etwas zu verbinden. Betrachtet man die Ergebnisse der Modifikationen des Sieben-Faktoren-Modells (siehe Anhang, Tabellen 25 und 27), dann scheint hier nur die Fehlerkovarianz zwischen impsafeSE und ipadvntST (Modifikation 17) als signifikante auf. Jene zwischen ipgdtimHE und impsafeSE ist ebenso wie Modifikation 9 mit einem std. EPC von 0,150 genau an der Schwelle und wird deshalb hier ebenfalls aus der weiteren Analyse ausgeschlossen. Modifikation 18 hingegen könnte durchaus sinnvoll sein, wenn die kognitiven Interviews betrachtet werden: „Es ist ihm wichtig, in einer sicheren Umgebung zu leben. Er vermeidet alles, das seine Sicherheit gefährden könnte." und „Sie ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Sie möchte ein aufregendes Leben führen." produzieren in den kognitiven Antworten sehr ähnliche Schlagwörter: Risiko, Abenteuer, extreme Sportarten, Sicherheit und Risikovermeidung. Das wäre keine klassische Erklärung für Fehlerkovarianzen, die üblicherweise manifeste Effekte auf der Ebene des Itemwordings und Effekte des Instruments erfasst. Hier ginge es eher darum, dass aufgrund der gegengleichen Bedeutung in der Formulierung ähnliche kognitive Prozesse ablaufen .

36 Ein weiteres Indiz, das für die Hypothese des „fürsorglichen Staates" spricht, sind signifikante positive standardisierte Residuen für ipstrgvSE mit den Benevolence-Items (siehe Tabelle Seite 97).

Self-transcendence – self-enhancement

Korrelationen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
imprich(PO)	-0,142	-0,178	-0,160	-0,059	-0,162
iprspot(PO)	-0,002	0,041	-0,025	0,089	0,060
ipshabt(PO)	0,029	0,057	0,036	0,117	0,042
ipsuces(AC)	0,023	0,053	-0,017	0,101	0,084
Residuen	ipudrst(UN)	impenv(UN)	ipeqopt(UN)	iplylfr(BE)	iphlppl(BE)
imprich(PO)	-7,465	-9,783	-8,299	-3,963	-8,837
iprspot(PO)	-0,778	1,302	-2,043	3,429	2,159
ipshabt(PO)	0,371	2,134	0,835	5,192	1,024
ipsuces(AC)	0,059	1,741	-2,405	4,167	3,273

Für das Gegensatzpaar *self-transcendence* und *self-enhancement* (ρ=0,051) gibt es eine einzige Modifikation (Nr. 19), die das Auswahlkriterium erfüllt. Gleichzeitig ist sie sowohl in Hinblick auf MI als auch auf std. EPC die bisher Gewichtigste. Alle *self-enhancement*-Items korrelieren bis auf die Ausnahme von imprichPO mit den *self-transcendence*-Items entweder fast gar nicht oder sogar leicht positiv anstatt, wie von der Theorie vorgesehen, negativ. Im Modell wiederum wird die Korrelation der Faktoren quasi auf Null geschätzt, was mit den übrigen Korrelationen übereinstimmt. ImprichPO hingegen vereint das gesamte von der Theorie postulierte gegensätzliche Potential. Zur Wiederholung: imprichPO ist eines der linksschiefen Items und damit in die dafür bereits erwähnten speziellen Zusammenhängen verwickelt.

4.4.4 Reliabilitäten

Tabelle 14 fasst die formalen Reliabilitäten und standardisierten Ladungen der Modelle mit vier und mit sieben latenten Faktoren zusammen. Brown (2006: 173) warnt davor, Parameterschätzer zu interpretieren, solange das Modell nicht angepasst ist. Die überwiegende Mehrheit der Items in diesem Fall misst das ihm jeweils zugewiesene Konstrukt unabhängig von der Konfiguration (Differenz≈0). In keiner Version unterschreitet ein Item die Schwelle einer standardisierten Ladung von 0,4, die als Mindestanforderung an die Messeigenschaft gesehen wird. Hingegen verbessert sich die Qualität der Messung bei den Items ipcrtivSD, impsafeSE, ipgdtimHE, impfreeSD, ipstrgvSE, ipadvntST und impfunHE von der Vier- auf die Sieben-Faktoren-Lösung. Das sind mit Ausnahme von impdiffST, das in dieser Auflistung fehlt, exakt jene Items, die gemäß der Analyse der Messmodelle eigene Faktoren bilden können. Die Subsumierung unter die *higher-order values* verzerrt ihre formale Zuverlässigkeit als Indikatoren. Die Ausnahme impdiffST scheint auch ohne einen eigenen Stimulation-Faktor relativ gut zu messen. Letztendlich ändert die Erweiterung auf sieben Faktoren jedoch kaum,

welche Items besonders gut oder schlecht ihre latenten Faktoren messen (Ausnahme ist ipcrtivSD, das sich von $R^2<0,3$ auf $R^2=0,363$ verbessert).

4-Faktoren			Differenz	7-Faktoren		
Item	*std. Ladung*	*Reliabiltäten*		*Reliabiltäten*	*std. Ladung*	*Item*
ipcrtiv(SD)	0,519	0,269	0,094	0,363	0,602	ipcrtiv(SD)
imprich(PO)	0,546	0,298	-0,001	0,297	0,545	imprich(PO)
ipeqopt(UN)	0,582	0,339	0,006	0,345	0,587	ipeqopt(UN)
ipshabt(AC)	0,715	0,511	-0,001	0,510	0,714	ipshabt(AC)
impsafe(SE)	0,573	0,328	0,109	0,437	0,661	impsafe(SE)
impdiff(ST)	0,657	0,432	0,018	0,450	0,671	impdiff(ST)
ipfrule(CO)	0,522	0,273	0,008	0,281	0,530	ipfrule(CO)
ipudrst(UN)	0,574	0,330	-0,006	0,324	0,569	ipudrst(UN)
ipmodst(TR)	0,507	0,257	0,028	0,285	0,534	ipmodst(TR)
ipgdtimHE	0,709	0,503	0,103	0,606	0,778	ipgdtim(HE)
impfreeSD	0,509	0,259	0,078	0,337	0,581	impfree(SD)
iphlpplBE	0,601	0,361	-0,008	0,353	0,594	iphlppl(BE)
ipsucesAC	0,773	0,597	0,004	0,601	0,775	ipsuces(AC)
ipstrgvSE	0,574	0,330	0,120	0,450	0,671	ipstrgv(SE)
ipadvntST	0,619	0,383	0,154	0,537	0,733	ipadvnt(ST)
ipbhprpCO	0,735	0,540	0,033	0,573	0,757	ipbhprp(CO)
iprspotPO	0,470	0,221	-0,003	0,218	0,467	iprspot(PO)
iplylfr(BE)	0,568	0,323	0,004	0,327	0,572	iplylfr(BE)
impenvUN	0,618	0,382	0,005	0,387	0,622	impenv(UN)
imptradTR	0,517	0,267	-0,003	0,264	0,514	imptrad(TR)
impfunHE	0,701	0,492	0,118	0,610	0,781	impfun(HE)

Tabelle 14: Faktorladungen und Reliabilitäten zu einem Modell mit vier und sieben Faktoren (relativ hohe Differenzen zwischen den Lösungen wurden grau hinterlegt)

Die Wertebreite der std. Ladungen beträgt minimal 0,47 und maximal 0,78. Als Kriterium der Güte eines Items wurde jenes Intervall gewählt, bei dem Indikatoren eine std. Ladung von 0,650 über- bzw. 0,550 unterschritten. Damit ist das oberste und unterste Quartil erfasst.

Items mit hohem R^2		Items mit niedrigem R^2	
ipgdtimHE	0,606	imprichPO	0,297
ipsucesAC	0,601	ipfruleCO	0,281
ipadvntST	0,537	ipmodstTR	0,285
ipbhprpCO	0,573	iprspotPO	0,218
impfunHE	0,610	imptrad(TR)	0,264

Besonders gute Indikatoren, gemessen an ihrer Reliabilität, finden sich für die Wertetypen Achievement, Stimulation, und Hedonism, eher schlechte für Power und Tradition. Für Conformity gibt es sowohl ein Item im obersten als auch eines im untersten Quartil (siehe Tabelle 15). Verglichen mit Tabelle 10 hat sich an der Zusammensetzung von guten und mangelhaften Indikatoren nichts geändert, bis auf ipfruleCO, das nach dieser detaillierteren Prüfung ins unterste Quartil gerutscht ist. Niedrige Reliabilitäten deuten an, dass der Indikator schlechte Messeigenschaften besitzt. Den Gründen für problematische Messeigenschaften im untersten Quartil wird mit Hilfe der kognitiven Interviews im folgenden Kapitel nachgegangen.

4.4.5 Bewertung der Messeigenschaft ausgewählter Items unter Zuhilfenahme kognitiver Interviews

In der Einleitung verglich ich mein Vorgehen mit der Arbeit eines Kriminalisten, der von Tatort zu Tatort geht und mit den Zeugen spricht. Die Tatorte sind die ausgewählten Items mit problematischen Messeigenschaften. Jetzt geht es darum, den Hergang der Tat zu verstehen, indem ich die Protokolle meiner Befragten analysiere. Ich stelle mir dabei die Fragen:

- Was passiert in den Köpfen der Befragten, wenn sie die Porträts bewerten sollen?
- Was sind „schlechte" Messeigenschaften? Und wodurch kommen sie zustande?

Kognitive Interviews erlauben einen kleinen Einblick in den Kosmos, der sich zwischen Frage und Antwort abspielt. Sie können mir Antworten auf meine Fragen geben.

Die ersten zwei Items der Liste – imprichPO und ipfruleCO – gehören zusammen mit ipdadvntST zu jenen drei Indikatoren, die linksschief sind. Befragte neigen dazu, sich mit diesen Porträts weniger zu identifizieren. Das ist ein signifikanter Befund, denn Schwartz definiert Werte als sozialpsychologische Konstrukte, die auf *wünschenswerte* Endzustände oder Verhaltensweisen abzielen und zu solchen motivieren (siehe Kapitel 2.2). Wenn Werte tendenziell als unähnlich bewertet werden, weil die dahinterliegende Motivation als nicht wünschenswert gilt, dann stellt sich die Frage, ob solche Indikatoren noch Werte messen. Damit einhergehend sollte geklärt werden, ob Unähnlichkeit schlicht Unähnlichkeit bedeutet oder ob sie eine Messung für die Ablehnung einer bestimmten Assoziation gegenüber Teilinhalten der Items ist (und dadurch das wünschenswerte Element verliert), die womöglich durch einen mehrdeutigen Stimulus hervorgerufen wurde.

Das erste Beispiel illustriert so einen Fall für das Item „Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht." (ipfruleCO), wo ein Teilaspekt eines Items als nicht wünschenswert wahrgenommen wird:

> I: „Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht."
> R: „Ich hab ein großes Problem mit Leuten, die sich allgemein Regeln widersetzen. […] man glaubt immer, dass man sich durchschwindeln kann, wenn keiner zuschaut, aber es schaut eh IMMER jemand zu. Das ist ja das Problem. Also dieses: „es schaut keiner her. Ich gehe jetzt. Jetzt kann ich rauchen." Danach riecht mans, und dann ist der Wurm schon drinnen. Also bin ich eigentlich sehr intolerant. Also, ich hab ein Problem mit blöden Regeln."

I: [liest die Frage noch einmal vor]
R: „Ja, jetzt sag ich dir genau, warum ich ein Problem hab. Ähm, das womit ich ein Problem hab im ersten Satz, ist, ähm … dass die Leute prinzipiell nur kusch sein sollen. Wenn es formuliert wäre: „Sie sollen sich an die Regeln halten, denen sie zugestimmt haben.“ Da bin ich dafür. […] Das ist eigentlich schon, wenn man sichs genau anschaut, das ist dann schon ein bisschen Diktatur. Derjenige, der das glaubt ist dann Diktator […] Also derjenige, der das sagt, stellt die Regeln von sich auf. (Kogn. Int. Nr. 1, ipfruleCO)

Dieser Respondent spricht zwei kognitive Stränge an, die er verarbeitet: Einerseits lehnt er „dumme Regeln“ ab, andererseits entsteht durch den ersten Satz eine emotional aufgeladene Assoziation mit einem Diktator. Diese kommt noch in weiteren Interviews vor:

I: Was fasst denn die erste Frage für dich zusammen: Er glaubt, dass Leute das machen sollten, was man ihnen sagt? … Was ist das für dich für ein Mensch, der das glaubt?
R: Was ist das für ein Mensch? … Dass er mit Menschen zu tun hat, oder zu tun haben will, die ohne zu denken einfach das tun, was ihnen angeschafft wird. … Dass es keine Menschen sind, die hinterfragen, obs vernünftig ist. (Kogn. Int. Nr. 2, ipfruleCO)

I: Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht.
R: Gleicht mir wenig. Weil, man muss nicht das tun, was andere sagen. Wir sind ja da nicht in einer Diktatur! Jeder hat das Recht, das zu tun, was er möchte. (Kogn. Int. Nr. 5, ipfruleCO)

Abgesehen von der Nähe zu „Diktatur“ verbindet die Aussagen, dass es sich bei den „Regeln“ um etwas handelt, das von außen an die Person herangetragen wird. Es ist sozusagen ein oktroyiertes Gesetz, an das sich die porträtierte Person in reflexartigen Weise anpasst. Ein Beispiel für so eine Regel, die von vier Befragten genannt wurde, ist bei roter Ampel (nicht) über die Straße zu gehen, obwohl kein Verkehr herrscht.

I: „Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht.“
R: „Gleicht mir nicht. … Weil das, ja, das würde auch heißen, dass ich nie bei rot über die Straße gehen würde, nur weil die Ampeln jetzt so programmiert sind, und um drei Uhr in der Nacht eh niemand kommt, und ich dann da stehen würde – hehehe. Also, das denke ich nicht. Die Regeln sind da, den Menschen einiges leichter zu machen, aber nicht sie einzuschränken.
I: „Wieso ‚Gleicht mir nicht‘ und nicht ‚gleicht mir ganz und gar nicht‘?“
R: … Ja, weil ich grundsätzlich nicht etwas gegen alle Regeln habe. Es gibt ja auch sinnvollere Regeln, damit hier nicht ein Chaos entsteht. Allerdings muss man halt immer noch den gesunden Menschenverstand noch benützen, und sich nicht blind an etwas halten, ohne darüber nachgedacht zu haben, […] was dahintersteht hinter der Regel. (Kogn. Int. Nr. 6, ipfruleCO)

Neben einer stark emotionalen Ebene und einem abschreckenden Bild, gibt es – wie hier das Beispiel veranschaulicht – auch noch einen anderen Faktor, der es den Befragten schwer macht, dem Item zuzustimmen. Man spricht in der Skalenkonstruktion – wie erwähnt – von Itemschwierigkeit: Wie leicht fällt es den Befragten ein Item zu lösen, ihm also zuzustimmen oder es abzulehnen? Die Itemformulierung scheint mit „immer und überall" sehr spitz zu sein und Schwierigkeiten zu bereiten, sich so eine Person überhaupt vorstellen zu können. Das zeigt das nächste Beispiel:

> I: „Sie glaubt, dass Leute das machen sollten, was man ihnen sagt. Sie meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht."
> R: „Mhm. Ich glaube, es gibt so eine Person nicht."
> I: „Gibt es nicht?"
> R: „Nein, ich glaube nicht. - Viele machen das, was die anderen sagen aus Angst vor Autorität: aus Angst vorm Chef […]. Aber dass, die Leute sich immer an die Regeln halten, auch wenn das keiner sieht … ich kann mir das nicht vorstellen. […] [flüsternd verschwörerisch] Ich fahre auch mim Fahrrad bei rot, wenn es keiner sieht [beide lachen]. Auch wenn ich weiß, dass ich das nicht darf.
> I: „Glaubst du, dass das jemand glauben könnte? So prinzipiell der Ansicht sein?"
> R: „… … Vielleicht […] Vielleicht ein Monk[37]. Aber den kenne ich nicht persönlich. Kann ich nicht sagen. Glaub ich nicht. […] Ich glaube nicht, dass es die gibt."
> I: „Also prinzipiell findest du schon, dass man sich an Regeln halten sollte?"
> R: „Das schon! Auf jeden Fall! Ohne Regeln wird's chaotisch sein. Ohne Regeln ist das … ergibt das ganze keinen Sinn. Aber dass es da wirklich jemanden gibt, der sich an jede Regel hält und nie im Leben was Anderes gemacht hat, das kann ich mir nicht vorstellen." (Kogn. Int. Nr. 7, ipfruleCO)

Ein schwieriges Item in eine Skala einzubauen kann durchaus sinnvoll sein (vgl. Schnell et al., 2008), aber man muss sich die Frage stellen, wie sehr es dann noch die Realität der Befragten erfasst, denen der Fragebogen vorgelegt wird. Die Daten zeigen allerdings keine Bodeneffekte für dieses Item an, d.h. die Breite der Antwortkategorien wird von den Befragten vollständig ausgenützt. Eine während des Sozialismus in Ungarn aufgewachsene Respondentin beschreibt z.B., dass Sie sich mit diesem Porträt identifizieren kann:

> I: „Sie glaubt, dass Leute das machen sollten, was man ihnen sagt. Sie meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht."
> R: „Äh, … nicht immer und überall … aber doch: hauptsächlich! Das heißt: Antwort 2. - Also so ohne Wenn und Aber nicht! Aber es gibt halt Regeln. Es gibt Gesetze, es gibt Regeln, Verhaltensregeln – die sollte man einhalten. Und ich versuche die auch möglichst einzuhalten. […] Ich akzeptiere, wenn mir etwas gesagt wird, dass ich das so tun soll, wenn es mit meinen Prinzipien übereinstimmt.
> I: „Also da schwingt schon auch ein bisschen: Prinzipien und so hinein …"

37 Monk ist der Protagonist einer gleichnamigen TV-Serie. Er leidet an Zwangsstörungen.

R: „Jaja, selbstverständlich! Also ich würde nicht jetzt, ähm, jemanden töten, weil mir einer sagt: äh, mach das! Äh, also so auf so extreme Situationen ist die Frage wahrscheinlich gar nicht ausgelegt, aber … - also nicht unbedingt, aber ich unterwerfe mich den ‚üblichen' Regeln."
I: „Auch wenn niemand hinsieht."
R: … auch wenn NIEMAND hinsieht! Auch! Also, ich, ich, ja. Auch wenn jemand hinschaut – ich machs aber nicht deswegen, weil es jemand sieht! Verstehst du? Sondern – jetzt ein ganz einfaches Beispiel: Ich bin gewohnt, bei schön gedecktem Tisch zu essen. Bin alleine zu Hause, habe für mich alleine gekocht und trotzdem decke ich mir den Tisch, weil das eine Regel für MICH ist. Sieht kein Mensch! Würde jemand sagen: Ja, [lacht] was für ein Schwein! – Interessiert mich nicht! Das sind meine Regeln!" (Kogn. Int. Nr. 4, ipfruleCO)

Die Respondentin macht deutlich, dass die Einhaltung der Regeln insofern befolgt wird, sofern sie von ihr akzeptierte sind oder zumindest in einen akzeptablen Rahmen fallen („üblich"). Regeln, die gegen Prinzipien verstoßen, werden abgelehnt. Dieser Befund spricht an sich für die Messqualität des Indikators, denn er kann scheinbar Personen mit hohen Ausprägungen auf der latenten Variable (Ähnlichkeit mit dem Item) identifizieren.

Bis auf dieses Beispiel empfindet aber kein anderer der Befragten Ähnlichkeit mit dem Porträt. Die ersten Beispiele machten auch deutlich, dass dieses Item Assoziationen hervorruft, die abgelehnt werden, aber in ihrer Qualität nicht mehr mit Konformismus direkt zusammenhängen. Wenn statt eines porträtierten Menschen, der Regeln befolgt, ein Mensch tritt, der diktatorische Eigenschaften besitzt, dann entspricht die Ablehnung des Items nicht mehr der Unähnlichkeit mit einer regelkonformen Person, sondern der Ablehnung einer Diktatur, der ein Machtmotiv nahegelegt wird. Tatsächlich findet sich auch eine Fehlerkovarianz zwischen den Items ipfruleCO und imprichPO sowohl in den vorgeschlagenen Modifikationen der Sieben-Faktoren-Lösung als auch des paarweisen Vergleichs. Darüber hinaus wurde schon erwähnt, dass bei einem menschlichen Wert, der erstrebenswert sein sollte, generell Ablehnungen für die Messung ein Problem darstellen.

Das Item „Reichtum ist ihm wichtig. Er möchte viel Geld und Luxusgüter besitzen." (imprichPO) ist das zweite Item, das tendenziell als unähnlich betrachtet (abgelehnt) wurde. Nach Reichtum und Luxus zu streben, schien in den Interviews nicht mehr ausschließlich als Standard für die Beurteilung von Handlungen zu dienen, sondern wurde abgelehnt, weil dieses Streben entweder zum Selbstzweck oder moralisch verurteilt wiurde.

I: „Wozu könnte diese Person das brauchen? Reichtum und Luxusgüter?"
R: „Och, … wenn jemand das braucht, fürs Selbstgefühl, oder Selbstbestätigung, ja, aber, ehrlich gesagt, diese Personen, die so … sagen wir: ‚luxuriös' sind, oder

das anstreben, äh, sind mir irgendwie fremd. Das ist gar nicht meine Welt. Also diese äh, diese immer präsent, und immer gestylt und immer die beste Gesellschaft, die besten Konzerte, die besten ähm, die besten Vorführungen, die besten Reisen… Das ist nicht meins. Und mit solchen Leuten verstehe ich mich auch nicht sehr gut. Hin und wieder bei unseren Reisen sind wir mit solchen zusammengekommen, die so alles schon gemacht haben … nicht meins.“
I: „Also, die haben viel erreicht? Oder wie würdest du sie beschreiben?
R: „… ja, sie haben etwas erreicht, was für sie wichtig war, aber für mich ist es nicht anstrebenswert. Ich bin zufrieden – wie gesagt, wenn was zusätzlich irgendwie dazukommt, äh … dann: ja. Dann sag ich nicht: ‚Neinnein, das will ich nicht!‘ Das ist klar. Aber nur darauf hinzuarbeiten, dass ich nur in Fünf-Sterne-Hotels, oder nur erste Klasse im Flieger, das wäre dann wirklich Luxus, aber das wäre nicht notwendig. Man kann ohne auch.“ (Kogn. Int. Nr. 4, imprichPO)

I: „Ist das sozusagen maßlos, was diese Person verlangen würde?
R: „Es kommt drauf an, was sie mit dem Geld machen wird. […] Kann ich mir schon vorstellen, dass die Leute, Bankleute, mit Aktien schon reicher werden. […] Aber für mich ganz normal Sterbliche, ganz normal Angestellte, also: das ist ja schon übertrieben!“
I: „Reichtum ist so übermäßig viel? Also mehr als für das, was man braucht, um die Rechnungen zu zahlen?“
R: „Also viel, viel mehr. Das ist riesiger Luxus, und quasi … ich weiß nicht, von einem Punkt zum anderen gehen. Von Florida nach Australien und zurück und das und das; und sich zu treffen mit Leuten, die auch reich sind. Also das ist ja … die obersten Zehntausend. Das ist für mich Reichtum. Mit dem kann ich mich nicht vergleichen.“
I: „Was glaubst du, warum jemand so was hat, oder haben möchte? Oder glaubt, dass es wichtig ist?“
R: Vielleicht vererbt? Vielleicht wenn die Eltern oder Großeltern reich sind, ist sie das gewöhnt wie Paris Hilton, oder so. Sie kann sich ein anderes Leben gar nicht vorstellen, würde ich einmal sagen. […] Und das, wie Paris Hilton: Nein! [lacht] […] Nein, weil ich glaube, sie kann überhaupt die einfachen Leute, was deren Hintergrund ist, die viel arbeiten, die viel Arbeit hineinstecken in ihren Besitz – sie sieht die Leute gar nicht. – Ich sehe das so! Wenn sich jemanden alleine bis nach oben gearbeitet hat, dann ist das schon was Anderes. […] Vielleicht strebt jemand nach Reichtum, der in totaler Armut aufgewachsen ist. Und dann will er sich durch Schule oder Arbeit … sodass sie ganz nach oben kommt. Das respektiere ich voll!“ (Kogn. Int. Nr. 7, imprichPO)

Zu einer moralischen Beurteilung, wie die Porträtierte zu ihrem Reichtum kam, mischt sich in die Bewertung, dass sich die Befragte gar nicht mit der Person identifiziert. Der Indikator kann dadurch „Reichtum“, unabhängig von der materiellen Situation der Befragten, nicht als leitendes Prinzip erfassen. Die Distanz zum Porträt verhindert das. Reichtum als leitendes Prinzip nannte einzig ein Befragter unter allen elf Interviewten:

I: „Wieso gleicht es dir nicht sehr?“
R: „Weil ich glaube, dass ich auch ohne Reichtum sehr gut leben kann. Also ohne viel Geld und Luxusgüter. Ich gehe aber nicht so weit zu heucheln, dass ich das

nicht will. […] Aber … es ist nicht eines meiner primären Ziele.“ (Kogn. Int. Nr. 9, imprichPO)

Hingegen scheint das Streben nach Reichtum für manche der Befragten vielmehr einen Einfluss auf die Suche nach ihrem Glück zu haben und weniger als Ziel für sich selber zu stehen:

I: „Wofür würdest du denn viel Geld, Reichtum, und Luxusgüter brauchen?“
R: „Ich glaube, ich möchte glücklich sein. Und ich kann mir vorstellen, dass ich mit Reichtum und Luxusgütern einen guten Beitrag dazu leisten könnte. Wenn auch nicht den einzigen.“ (Kogn. Int. Nr. 9, imprichPO)

I: „Reichtum ist ihm wichtig. Er möchte viel Geld und Luxusgüter besitzen.“
R: „[atmet tief] … du, ich bin, jeder ist auf seine Art irgendwie reich und Geld ist einfach nicht alles. […] ich glaub, es wär mir einfach lieber, dass ich einfach ein ruhiges Leben führen könnte, mit weniger Geld, aber dafür müsste ich nicht so viel arbeiten – sagen wir so – um einen gewissen äh, Lebensstandard zu halten, oder sowas. Und … deswegen glaub ich auch nicht – also ich bin auch der Meinung: Geld spielt keine Rolle, um glücklich zu sein, und Reichtum auch nicht. (Kogn. Int. Nr. 5, imprichPO)

I: „Reichtum ist ihm wichtig. Er möchte viel Geld und Luxusgüter besitzen.“
R: „Gleicht mir nicht. Weil für mich das ‚etwas Haben‘, und das ‚etwas Besitzen‘ nicht Zufriedenheit bringen würden. Mir wäre lieber Friede und Harmonie. Da verzichte ich lieber auf materielle Güter. Der innere Frieden, die innere Ruhe, die Freude an kleinen Dingen, die man nicht zu bezahlen hat – schau dir den Kirschbaum da draußen an, wie der blüht. Der blüht einfach ohne, dass man Geld dafür ausgibt. Und das ist schön. Unter dem stehe ich und schau mir die Blüten an. […]“
I: „Was ist das für eine Person, oder hast du dir da eine bestimmte Person dabei gedacht?“
R: „Ja, meinen älteren Sohn. Der sagt: Geld ist alles! Und mir drehts jedes Mal den Magen um.“ (Kogn. Int. Nr. 2, imprichPO)

Das letzte Beispiel illustriert auch, dass die gleichzeitige Assoziation von „Glück“ und „Reichtum“ bei manchen Befragten auch eine emotionale Ablehnung hervorruft. Auch hier wird womöglich weniger die Ungleichheit als vielmehr eine Abneigung in der Messung kundgetan. Die Assoziation mit „Glück“ und die Beurteilung dieses Konstrukts verwischt die Messung des eigentlich erwarteten Wertetyps „Macht“. Zwar wurde dieser von manchen Befragten durchaus als Beweggrund für das Streben nach Reichtum genannt, aber eben auch neben dem des moralisch verwerflichen Selbstzwecks und dem Streben nach Glück. Wenn dieser Indikator dadurch mehrere Dimensionen misst, kann er notwendigerweise nicht eine einzige besonders klar messen.

Zum Vergleich ein Exkurs zum Indikator ipadvntST „Sie ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Sie möchte ein aufregendes Leben führen.“: das dritte linksschiefe Item. Es zählt nicht zum untersten Quartil in Bezug auf seine Reliabilität. An ihm fällt auf,

obwohl eine hohe Risikovermeidung (also eine Tendenz zur Unähnlichkeit mit dem Porträt) unter den Befragten vorherrscht, primär die Lust nach einem Abenteuer bzw. eher „ein aufregendes Leben zu führen“, die meisten anspricht. Hier scheint eher der Stimulus „Ähnlichkeit mit dem Porträt“ als die Ablehnung zu überwiegen. Die Messung eines Wertes ist damit weiter gegeben. Die Ladung ist auch deutlich besser als bei den anderen beiden Items.

> I: „Er ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Er möchte ein aufregendes Leben führen.“
> R: „Ja.: […] dieses nicht-Häuslein bauen, sondern Indiana Jones. Das möchte man. Man möchte derjenige, der mit einem gewissen, also nicht mit einer absoluten Blindheit, aber mit einem gewissen intellektuellen Background trotzdem in den Dschungel geht! […] Ich muss aber nicht in die Welt hinaus. […]“
> I: „Also auch das aufregende Leben?“
> R: „Ja, das aufregende – eigentlich das ist fast noch wichtiger, […].“ (Kogn. Int. Nr. 1, ipadvntST)

Den letzten Satz möchte ich noch in den rechten Kontext rücken: Damit bringt der Respondent zum Ausdruck, dass „aufregendes Leben führen“ in der Bedeutung im Vordergrund steht und auch derjenige Stimulus ist, auf den die Reaktion überwiegend erfolgt.

Die folgenden drei Exzerpte sollen den Befund erhärten, dass zwei Komponenten durch das Item vermittelt werden, die aber nicht im Widerspruch zueinander stehen.

> I: „Sie ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Sie möchte ein aufregendes Leben führen.“
> R: „Ja. Abenteuer im gewissen Maße. Ist halt die Frage, ab wann ist etwas ein Abenteuer. Ähm, aber Risiken bin ich jetzt auch kein Mensch, der viele Risiken eingeht. …“ (Kogn. Int. Nr. 3, ipadvntST)

> I: „Er ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Er möchte ein aufregendes Leben führen.“
> R: „Natürlich will ich Abenteuer erleben. Wer will das nicht?“ (Kogn. Int. Nr. 5, ipadvntST)

> I: „Sie ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Sie möchte ein aufregendes Leben führen.“
> R: „… überhaupt nicht! [lacht] Nein, überhaupt kein Abenteuer, kein großes Risiko.“
> I: „Auch kein aufregendes Leben?“
> R: „Uff, was ist aufregendes Leben? Äh, wir können auch mit ganz normalem Leben aufregendes Leben führen. Da musst du selber etwas machen, dass es was anderes ist. Damit es nicht monoton ist, dass es nicht immer das gleiche ist: dann ist es schon ein bissl aufregend. Oder ein Flug nach Paris. Das ist ja auch aufregend. Aber da muss man nicht abenteuerlich sein, ähm, … von, weiß nicht, Bungeespringen bis zum Kanufahren, oder solche Abenteuer. […] Also das Adrenalin irgendwo in die Höhe zu schieben, also das nicht unbedingt.“ (Kogn. Int. Nr. 7, ipadvntST)

> I: „Er ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Er möchte ein aufregendes Leben führen."
> R: „Erster Satz: 5, zweiter Satz: 1."
> I: „Also, du möchtest ein aufregendes Leben führen?"
> R: „Ja. Generell, wenn du mich fragst: ‚Möchtest du ein aufregendes Leben führen?' Dann sage ich: Jaaa! Ähm … in der Praxis glaube ich, strebe ich nicht aktiv Abenteuer an. Und nehme auch nicht Risiken auf mich dafür … in besonderem Maße."
> I: „Also, eine primäre positive Grundhaltung dazu, aber wenn dann sozusagen die Konsequenzen auch hinzukämen, also wie man's dann wirklich umsetzt, …"
> R: „Ja. Genau. Also, da bin ich eher sehr realistisch und schaue mir die Umstände an und … dann ist das – das hängt dann wieder mit diesem Sicherheitsding zusammen. Mache ich's, wenn ich das Gefühl habe, dass es mir Spaß machen würde […] Ähm, aber wenn die hohe Chance besteht, dass es sich für mich NICHT positiv auswirken wird, dann mache ich es nicht. Das wäre ja dann ein ‚Risiko eingehen'." (Kogn. Int. Nr. 9, ipadvntST)

Bei diesen Antworten zeigt sich, dass prinzipiell „Aufregung" eine erstrebenswerte Sache ist. Dieser Umstand wird durch die Itemformulierung der Neigung zur Risikovermeidung gegenübergestellt und in einem rationalen Prozess bewertet (Respondent 3). Die Itemformulierung lässt noch genügend Spielraum für die Befragten sich in dem Porträt irgendwo wiederzufinden (Respondentin 7). Scharfe, ausschließende Formulierungen wie „immer und überall" oder „sehr" fehlen. Auch blieben emotionale Ablehnungen bis auf jene einer einzigen Respondentin aus. Für sie stand im Gegensatz zu den anderen der Aspekt „Risiken einzugehen", den sie vollständig ablehnt, im Vordergrund. Es gab aber auch Beispiele, wo Befragte beide Sätze kongruent wahrnahmen und wo für sie die Motivation in dieselbe Richtung wirkte.

Zurück zu den übrigen Indikatoren mit problematischen Messeigenschaften. In der Skala gibt es vier Indikatoren, die so formuliert sind, dass sie eine Aussage darüber treffen, wie jemand ist, bzw. welchen Charakter die Person hat. Für alle Items wird im ersten Teilsatz die Formulierung „[Eigenschaft] zu sein" verwendet. Das Item „Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken." (ipmodstTR) ist eines davon. Im Kapitel 2.1 wurde dargelegt, dass Charakterzüge (engl. *traits*) in den Sozialwissenschaften von Werten unterschieden werden. Kann es sein, dass durch die Vermischung unterschiedlicher sozialpsychologischer Konstrukte mangels klarer Formulierung die Messeigenschaft eines Indikators darunter leidet? Bevor die Belege aus den kognitiven Interviews zu dem genannten Item zitiert werden, soll der Kontext zu den übrigen drei Items iplylfrBE „Ihm ist es wichtig, gegenüber seinen Freunden treu zu sein. Er möchte für die Menschen da sein, die ihm nahe stehen." (R^2=0,327), ipcrtivSD „Es ist ihm wichtig, neue Ideen zu entwi-

ckeln und kreativ zu sein. Er unternimmt alles gerne auf seine eigene Art und Weise." (R^2=0,363) und ipsucesAC „Es ist ihm wichtig, sehr erfolgreich zu sein. Er hofft, dass andere Leute seine Leistungen anerkennen." (R^2=0,601) hergestellt werden. Die Messqualitäten von iplylfrBE und ipcrtivSD sind gemessen am R^2 nicht sehr hoch und liegen nur knapp über dem gesetzten Intervall. Die Hypothese würde durch diese Fallbeispiele unterstützt. Doch wie sieht es mit ipsucesAC aus? Trotz Formulierung als Charakterzug weist dieser Indikator eine der höchsten Reliabilitäten der gesamten Skala auf. Er scheint ein besonders starker Indikator für die latente Variable zu sein, die „Kontrolle oder Dominanz über Menschen und Ressourcen" (siehe Tabelle 1, Seite 10) misst und implizit ein wünschenswertes Ziel oder einen wünschenswerten Endzustand – gemäß der Wertekonzeption nach Schwartz – definiert. In den kognitiven Interviews geben die Befragten an, dass für sie das Item sehr deutlich eine Grundhaltung , d.h. die Werthaltung, eines Menschen widerspiegelt. *Einerseits* ist es in sich konsistent (die zwei Teilsätze erfassen ein gemeinsames Konstrukt), andererseits ist mit dem Eigenschaftswort „erfolgreich" auch implizit ein „Streben" verbunden. Bilder, die evoziert werden, zeigen Karrieremenschen, die ehrgeizig und von dieser Grundhaltung „getrieben" sind. Auch wenn „Erfolg" im Bedeutungskontext variabel ist, findet die Mehrzahl der Befragten, dass „Erfolg haben" prinzipiell eine erstrebenswerte Sache ist, wenn dafür eine adäquate Leistung erbracht wurde. Obwohl ipsucesAC vermeintlich wie ein Charakterzug formuliert ist, ist im speziellen Fall von „erfolgreich sein" dieser so angelegt, dass er von sich aus mit Aktivität verbunden ist, durch die dieses Ziel erst erreicht werden kann. Erfolgreich ist man nicht durch einen Charakterzug, sondern durch das, was man tut, um durch die Anerkennung von anderen Erfolg zu erfahren.

Ähnlich steht es mit iplylfrBE: Hier schwanken die Antworten der Befragten zwischen: „treu sein" und „Treue/Loyalität" als wichtige Eigenschaft zu empfinden und andererseits Freundschaften zu pflegen. Die Stimuli treten unterschiedlich stark auf, wobei die Aktivität, nämlich etwas dafür zu tun, oder einzugestehen, dass man dafür mehr tun müsste, eher im Vordergrund steht. Die mangelnde Reliabilität ließe sich teilweise durch diese Ambivalenz erklären und hat auch andere Gründe, auf die hier nicht weiter eingegangen wird. Der Stimulus etwas Bestimmtes „zu sein" steht damit nicht ausschließlich im Vordergrund. Beim Item ipcrtivSD tritt der Fall ein, dass „kreativ Sein" keine rein essentialistische Aussage darstellt, sondern von den Befragten oft in einen Kontext gerückt wird: man ist in einem bestimmten Feld (in der Arbeit, der Küche, beim Hobby, etc.) kreativ – oder auch nicht. Dadurch bleibt den Befragten die Möglichkeit „kreativ Sein" als einen wünschenswerten Endzustand aufzufas-

sen, ohne es ausschließlich als eine Charaktereigenschaft zu sehen, die man hat oder nicht. „Kreativ Sein“ wird auch – wie „erfolgreich Sein“ – mit Handlungen in Verbindung gebracht, die man setzt, um zu sehen, ob man kreativ ist oder nicht. Die mangelhaften Messeigenschaften für dieses Item scheinen eher auf unklare Stimuli zurückzuführen sein.

Anders verhält es sich beim Item ipmodstTR. Zu diesem Indikator äußerten sich die Befragten fast durchgehend in einer Weise, die erkennen lässt, dass hier tiefverwurzelte Charaktereigenschaften angesprochen werden, die die Essenz eines Menschen einfangen:

> I: „Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken.“
> R: „Zurückhaltend und bescheiden zu sein ist definitiv nicht in meiner Natur.“ (Kogn. Int. Nr. 1, ipmodstTR)

> I: „Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken.“
> R: „Das ist die Bescheidenheit, die in meinem Bruder drinnen ist. Und das muss bei mir nicht unbedingt sein. Kann schon sein, dass ich nicht so bescheiden bin.“ (Kogn. Int. Nr. 2, ipmodstTR)

> I: „Was glaubst du, warum jemand, der solche Glaubenssätze vertritt, warum der in erster Reihe stehen wollen würde?“
> R: „Weil ich, äh, … äh, … wie soll ich sagen: zu selbstkritisch bin. […] Äh, ich denke, ich bin kein schöner Anblick da vorne. Also warum soll ich das den Menschen antun, mich anzusehen. Oder anzuhören. Äh, … es gibt nicht so etwas Wichtiges, was ich jetzt unbedingt vor Leuten verkünden möchte. […] Ich, ich … mag keine Aufmerksamkeit.“ (Kogn. Int. Nr. 4, ipmodstTR)

> I: „Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken.“
> R: „Also, ich bin manchmal bescheiden und zurückhaltend. Aber nicht, weil es mir wichtig ist, es zu sein. Und ich versuche … nicht grundsätzlich, keine Aufmerksamkeit zu erregen.“ (Kogn. Int. Nr. 9, ipmodstTR)

Gerade der letzte Respondent macht deutlich, dass „zurückhaltend und bescheiden sein“ in diesem Kontext nicht als erstrebenswertes Ziel oder erstrebenswerter Zustand wahrgenommen wird. Es betrifft vielmehr die Persönlichkeit.

Das Item „Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt.“ (iprspotPO) besitzt die niedrigste Messqualität aller Items. Kein anderes ist so missverständlich formuliert, wie die folgenden Passagen zeigen:

> I: „Respekt hat auch etwas mit Gehorsam zu tun?“
> R: „Nein, Respekt hat mit Anerkennung zu tun. Aber der zweite Teil hat mit Gehorsam zu tun.“ (Kogn. Int. Nr. 1, iprspotPO)

> I: „Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt.“

R: „Pff. … Auf jeden Fall möchte ich respektiert werden. Aber die anderen müssen nicht das tun, was ich sage. Bitte: nein!“ (Kogn. Int. Nr. 2, iprspotPO)

I: „Es ist ihr wichtig, von anderen respektiert zu werden. Sie möchte, dass die Leute das tun, was sie sagt.“
R: „Ähm, das passt für mich jetzt nicht so zusammen, weil ‚Respekt‘ finde ich etwas, das möchte ich schon haben. Das bringe ich auch anderen Leuten entgegen. Aber deshalb will ich nicht, dass Leute tun, was ich sage. Ich möchte keine Führungspersönlichkeit sein. Aber, Respekt hängt damit nicht so zusammen. Respekt kannst du jedem entgegenbringen. … Also: respektieren kannst du jeden. Auch den Obdachlosen auf der Straße, aber, ähm, jemand, der will, dass man auf ihn hört, das ist jemand, der eine Führungsperson sein will oder eine Machtposition haben will.“ (Kogn. Int. Nr. 3, iprspotPO)

I: „Es ist ihr wichtig, von anderen respektiert zu werden. Sie möchte, dass die Leute das tun, was sie sagt.“
R: „Äh, das ist aber nicht für mich unbedingt – vielleicht verstehe ich das Wort ‚respektiert‘ nicht richtig – aber, ‚Respekt‘ bedeutet nicht unbedingt ‚Gehorsam‘ für mich. […] Also, ich möchte als Mensch respektiert werden, mit all meinen Fehlern, aber … aber das bedeutet nicht, dass … dass ich deswegen Befehle gebe oder dass ich berechtigt wäre, jemandem zu sagen, was der zu tun hat. Also, vielleicht ist für mich das Wort Respekt etwas anderes. Weil ich empfinde Respekt gegenüber den Alten, und, ja, ich respektiere die Menschen, aber das hat mit ‚den anderen sagen, was sie tun‘ nichts zu tun.“ (Kogn. Int. Nr. 4, iprspotPO)

I: [nachdem als unmittelbare Antwort auf das Item ein Widerspruch angedeutet wurde] „Widersprechen sich diese beiden Sätze?“
R: „Genau, weil der erste hat schon mit Respekt zu tun, dass du von anderen respektiert wirst. Nur wenn du die anderen manipulierst, dann ist das kein Respekt. Da hat das eine mit dem anderen nichts zu tun, eigentlich.“ (Kogn. Int. Nr. 7, iprspotPO)

„Respekt“ wie die Befragten ihn wahrnehmen, ist sehr stark mit Anerkennung – und wie im paarweisen Vergleich auch herauskam, mit Reziprozität – verbunden. Das ist nicht mit dem zweiten Teilsatz zu vereinbaren, wo es um das Streben nach Macht geht. Im Gegensatz dazu findet Befragter Nr. 8 dennoch eine Assoziation, die beide Konstrukte zusammenbringen kann. Das Zitat macht auch deutlich, wo in dem kognitiven Prozess Schwachstellen des Items zutage treten:

I: „Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt.“
R: „ … Ahm, respektiert werden … wär schon ganz gut. Das bringt einem schon etwas, wenn man respektiert wird. Man, … man kommt …. Moment, ich muss kurz meine Gedanken […] wie ich das jetzt rüberbringen soll.“
I: „Und: Er möchte, dass die Leute das tun, was er sagt?“
R: „… Mhm. Ja, also, für mich kommt das rüber wie Chef-spielen. …. Respektiert werden und das tun, was man sagt, das ist Chef-spielen, eigentlich. […]“
I: „Wo war denn dein Problem, als du das nochmal formulieren musstest? Woran lag es?“

R: „Das Problem lag daran, an … an Respekt, wie ich das formulieren soll. Wie ich das für mich aus Respekt sehe. Wie ich das formulieren soll. Das war mein Problem."
I: „Und formulierst du es anders, als es da in dem Zusammenhang steht, oder bist du eh dabei geblieben?"
R: „Ja, also für mich hat das jetzt schon einen Zusammenhang gehabt. Also wie gesagt. Also wer Chef sein will, will auch Respekt haben. Würde ich mal sagen. Es gibt auch Chefs, die haben, die sind vielleicht zu sehr streng. Das ist dann Angst, eigentlich."
I: „Was passiert denn da, wenn man Respekt vor einer Person hat? Du hast jetzt eine Gegenüberstellung zur Angst gemacht. Das ist es einmal nicht."
R: „Das ist … Ähm … ‚Achtung'! Und … also ich vergleich das jetzt mal so: ich habe paar Professoren [Lehrer einer höheren technischen Lehranstalt in Österreich], ich meine, das sind Autoritätspersonen. Die sollte man respektieren. Ein paar mehr, ein paar weniger. Und … ich fasse es so auf: Es gibt Professoren, die sind engagiert, die machen etwas gerne. Wenn ich sie frage, sie helfen mir auch und sie versuchen mein Problem zu lösen, was ich gerade derzeit in der Schule habe. Und, die respektiere ich, weil sie mir helfen, obwohl ihnen das nichts bringt. Und dann gibt's andere Professoren, und denen ist das komplett egal, und sie kümmern sich gar nicht um uns und [die sagen]: ‚Es ist so! Es ist so! Da habt ihr Pech gehabt!' Und solche Leute respektiere ich nicht. Da mach ich auch nichts gerne für sie. Also wie gesagt, die Leute respektiere ich, wenn sie sagen: ‚Ja, Herr [Name des Befragten], bei uns im Labor fehlt das und das. Könnten sie das bitte in der Mittagspause holen?' – Kein Problem! Gerne! … Das ist für mich Respekt!"
I: „ Also, ein Geben und Nehmen?"
R: „Ja. Aber auf 'ne andere Art und Weise. Auf einer anderen Basis. D.h. … hfff … wenn ich jemanden respektiere, sehe ich das so, dass er ein bisschen über mir ist. […] Das ist für mich ehrlich gesagt ‚Respekt'. Ja." (Kogn. Int. Nr. 8, iprspot-PO)

Zu dieser sehr elaborierten und einsichtigen Definition für die Nuancen von „Respekt" kam der Respondent erst, nachdem er den Raum, der ihm im Interview gegeben worden war, dazu vollständig ausgenützt hatte. Mit dem zweiten Teilsatz „Er möchte, dass die Leute das tun, was er sagt." gewinnt der erste Teil für ihn eine Färbung, die nicht ganz der erste Impuls gewesen zu sein scheint, der durch „Respekt" aufgekommen ist („Chef-spielen" deutet auch an, dass die porträtierte Person aus einer kindlichen Haltung heraus etwas einfordert, was wiederum nicht dem Bild von „Respekt" entspricht, das er gegen Ende entwirft).

Kognitive Interviews, die im englischsprachigen Raum mit dieser Skala durchgeführt wurden, liegen nicht vor. Es ist daher schwer zu beurteilen, ob das Missverständnis auf ein Problem in der Übersetzung oder sogar in der Formulierung des Originals begründet liegt. Auf Englisch lautet das Item: „It is important to her/him to get respect from others. She/he wants people to do what she/he says." Sehr viel klarer wäre das Item, wenn statt „to get respect from others" stehen würde „to demand respect of others". Ins Deutsche übersetzt könn-

te das in etwa so lauten: „Es ist ihm/ihr wichtig Respekt von seinen/ihren Mitmenschen einzufordern."

Auch bei Befragtem Nr. 9 findet sich ein sehr unüblicher Antwortprozess:

> I: „Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt."
> R: „Ich glaube, das gleicht mir. Sagen wir 2. Das ist vielleicht ein bisschen HART formuliert."
> I: „Was ist hart formuliert?"
> R: „Er MÖCHTE, dass die Leute tun, was er sagt. Das ist […] Hört man nicht gerne. Äh, deswegen gebe ich dem auch nur eine 2. Aber es ist doch … es gleicht mir!"
> I: „Was verstehst du denn unter Respekt?"
> R: „Ein gewisses Grundvertrauen in das, was der andere macht, und was er ist. Also … ähm … … … aufgrund dessen, was man von dem anderen kennt … ihn als positiv, und glaubhaft, und als … relevant bewertend. Sagen wir mal so: etwas von jemandem halten, und nicht alles, was er erstmal sagt, komplett zu hinterfragen. Sondern zuerst einmal: Aha, wenn der das sagt, dann hat es auf jeden Fall irgendeine Grundlage."
> I: „Und das schlägt in die gleiche Kerbe: dass man Vertrauen hat und dass man das auch macht, was er sagt."
> R: „Ja. Obwohl, wenn ich darüber nachdenke, ist das so, dass die Einschätzung des Wortes ‚respektieren' sehr stark von dem zweiten Satz gefärbt wurde. Also, natürlich kann man das auch viel allgemeiner formulieren und das so im Sinne von ‚akzeptieren' oder so …. so […] [R. spricht zögerlich] gleichwertig bzw. auch ein … es kann auch ein ‚in Ruhe Lassen' sein." […]
> I: „Du hast dich ja richtig auf den zweiten Satz gestürzt, sozusagen."
> R: „Ich muss sagen, ich war auch überrascht über den zweiten Satz: Der hat mich so richtig erwischt! So, dass ich beim ersten – beim ersten hab ich noch so gedacht: ha, ja, so ein bissl ein Softding. Und dann [lacht] war der zweite irgendwie nochmal sehr stark und klar – extrem eigentlich, finde ich – dadurch ist der erste etwas ins Hintertreffen geraten.
> I: „Und dann erst im Nachhinein hast du das Respektieren erst irgendwie hinein …"
> R: „Ja, wahrscheinlich. Weil ich glaube, nur wenn man sagt: ja, man muss doch die anderen Menschen respektieren, hat das ein biiischen auch was von einer Banalität. So von einem: jo, […] … Und ist nicht sehr konkret, sondern allgemein."
> (Kogn. Int. Nr. 9, iprspotPO)

Das Antwortverhalten deutet darauf hin, dass der stärkste Impuls über den zweiten Teilsatz erfolgt, und der erste eher untergeht. Das steht im Gegensatz zu den Aussagen der übrigen Befragten, für die „Respekt" relativ prominent im Raum stand. Bei Befragtem Nr. 9 fällt auch auf, dass er zwar durchaus einen Unterschied in der Motivation zwischen beiden Sätzen getrennt wahrnimmt. Da aber einer der Stimuli besonders stark in den Vordergrund rückt, wird der zweite außen vor gelassen und fließt auch gar nicht in die Beurteilung mit ein. Das wie-

derum ist eine der beschriebenen Antwortstrategien, um dem Dilemma doppelter Stimuli zu begegnen.

Das letzte Item „Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden.“ (imptradTR) beinhaltet Anteile, die den Befragten nicht nur unähnlich sind, sondern die von ihnen abgelehnt werden oder gar keinen Platz in ihrem Lebenskonzept haben und damit nicht mehr erstrebenswert sein können:

> I: „Traditionen sind ihm wichtig.“
> R: „Bei ‚Traditionen‘ hab ich ein bisschen ein Problem. Das Wort – mit dem Wort ‚Tradition‘ fang ich nicht viel an. […] Tradition? Die Tradition zum Beispiel: Ich muss jemanden grüßen, aber ich muss jetzt nicht der alten Tante, wie’s in der Tradition herkommt, die Hand küssen. Also diese Tradition brauch ich nicht. Die spiel ich höchstens. […] Zum Beispiel die Tradition, dass man die Eltern siezt. Und wie ich schon genannt habe: den Handkuss. Traditionen z.B. auch drinnen in einer strengen Kleidungsvorschrift: Wie habe ich mich wann zu kleiden? Jetzt habe ich eine Krawatte zu tragen? Oder habe ich einen Hut zu nehmen, weil es die Tradition verlangt?“ (Kogn. Int. Nr. 2, imptradTR)

> I: „Traditionen sind ihr wichtig. Sie möchte jene Sitten und Gebräuche befolgen, die ihr durch die Religion oder die Familie überliefert wurden.“
> R: „Gleicht mir wenig. Das hat damit zu tun, dass mir in Bezug auf Religion wenig vermittelt wurde, da ich ja nicht getauft bin. Aber Weihnachtsfeier und so mal ausgenommen, aber das hat bei uns keinen religiösen [Charakter] und da gibt es jetzt auch keine spezielle Tradition, dass wir immer Karpfen essen, oder das und das machen. Ja, so in dem Sinn gibt es jetzt keine Traditionen, die ich so stur befolgen will.“ (Kogn. Int. Nr. 3, imptradTR)

> I: „Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden.“
> R: „Also ich glaub, also Religion eher nicht, nachdem ich Agnostiker bin, ist mir das ziemlich egal. […] Also Religion ist mir eigentlich wurst.“ (Kogn. Int. Nr. 5, imptradTR)

> I: „Traditionen sind ihr wichtig. Sie möchte jene Sitten und Gebräuche befolgen, die ihr durch die Religion oder die Familie überliefert wurden.“
> R: „Also, ich habe prinzipiell …. große Schwierigkeiten, einfach eine Tradition, weil sie eine Tradition ist, weiterzuverfolgen. Das finde ich sehr gefährlich. […] Pff, ja: cum grano salis. Also, wenn … man muss sie hinterfragen, und wenn man sie für sich selber irgendwie anwenden kann und es einem Spaß macht und es zu einem passt, dann natürlich. Aber nicht grundsätzlich.“ (Kogn. Int. Nr. 9, imptradTR)

Respondentin Nr. 3 bringt in ihrem letzten Satz unbewusst zur Geltung, dass es hier um eine aus ihrer Sicht „sture“ Form der Ausübung einer Tradition ist.

Respondentin 11 bringt mit ihrer Aussage etwas Licht in die vertrackte Situation, der sich Befragte gegenübersehen, wenn sie dieses Item beantworten sollen. Ihrer Interpretation nach haben Traditionen in einem urbanen Kulturkreis den Anschein einer Kuriosität. Es ist in gewisser Weise „niedlich", einer gelebten Tradition beizuwohnen:

> I: „Traditionen sind ihr wichtig. Sie möchte jene Sitten und Gebräuche befolgen, die ihr durch die Religion oder die Familie überliefert wurden."
> R: „… [lacht] Naja, leider. Das würde wahrscheinlich schon so auf mich zutreffen. Ich würd das gern machen [lacht]. Aber ich machs nicht. Also, es sind mir Traditionen eben wichtig. Ich fotografier das ja eben auch deswegen – Erstkommunion im Kleid und so, ja? Weil mich das total fasziniert, dass da irgendwelche Mischlingskinder mit einer total melangefarbenen Haut da dieses weiße Kleid, … und ihr Bruder ist aber zwei Köpfe kleiner, ja? Und das ist halt super gewesen. Also mich interessieren Traditionen, aber eher, weil sie mich eben interessieren, weil – ich möcht sie gern verstehen. Ja? […] Jetzt heißt das nicht, dass ich immer nach der Tradition selber leben möchte, also, dass das heißt, dass ich mein Kind auf jeden Fall zu einer Erstkommunion, mit so einem Kleid ausstatten würde. Das heißt das eben eigentlich nicht! […] Weil da ist das ein bisschen ein Unterschied bei mir. … aber sonst würde ich „ja" sagen, grundsätzlich, es ist für mich nicht deswegen, weil es eine Tradition ist, was Schlechtes.
> I: „Und wie ähnlich wäre dir so eine Person?"
> R: „Würde schon irgendwie gleichen.[…] aber eher etwas." (Kogn. Int. Nr. 11, imptradTR)

Was genau beantwortet die Respondentin? Ist es Ihre Neugier einem „unbekannten" Phänomen gegenüber oder die Lust an ungewohnten Sujets? Ihre vorletzte Aussage könnte so interpretiert werden, dass Traditionen für sie primär etwas „Schlechtes" sind, aber womöglich, da sie ein Interesse als Fotografin daran findet, sieht sie über diese erste Annahme hinweg und kann den Traditionen auch etwas Gutes abgewinnen. Es ist höchst fraglich, ob bei dieser Respondentin noch Traditionalism erfasst wird. Es gibt dennoch Personen, bei denen das Item sehr gut ihre Werthaltung erfasst. Respondentin 4 z.B. sieht keine Probleme in dem Item, ausgenommen der „Religion", von der sie „in der letzten Zeit doch recht abgekommen" ist, was sie aber nicht daran hindert, sich dem Porträt sehr ähnlich zu fühlen.

Eine weitere Auffälligkeit tritt zu Tage, wenn die kulturellen Hintergründe der Befragten mitbedacht werden: Die Befragten Nr. 6, 7 und 8 sind entweder in erster oder zweiter Generation Migranten aus dem heutigen Serbien. Für sie haben die genannten Schlagwörter „Traditionen", „Familie" und „Religion" einen ganz eigenen Stellenwert. Ihre Antworten fielen in der Qualität ganz anders aus, als die der übrigen Befragten:

> I: „Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden."

R: „… Ja, also, Tradition … ist mir in dem Sinne wichtig, dass man halt die Mentalität besser versteht. Allerdings möchte ich jetzt nicht alles eins zu eins weiterführen, was meine Eltern oder Großeltern gemacht haben. Von daher würde ich sagen: Gleicht mir etwas."
I: „Was sind Traditionen für dich?"
R: „Ja, Traditionen … das ist auch irgendwie so schwierig … mmh … weiterzuführen, hier, in dem Umfeld. Zum Beispiel wenn meine Eltern aus Zentralserbien gekommen sind und dann hier in Wien wohnen, dann kann man Traditionen, wie zum Beispiel das typische Weihnachtsfeiern, oder solche Sachen auch nicht eins zu eins übernehmen, sondern man lebt das so mehr im … Kopf. Also man geht so in die Kirche und holt sich dann einen kleinen Ast statt einem ganzen Baum und tut statt – was weiß ich – statt eines ganzen Sacks voll Heu unter dem Tisch nur so ein bisschen halt. Man spielt das so ein bisschen nach – in Gedanken."[38]
I: „Also prinzipiell: Die Idee der Tradition wäre gut, aber das lässt sich nicht so ganz umsetzen?"
R: „Ja, weil das ist … diese Tradition hat auch etwas mit der eigenen Identität zu tun und der Identität der Familie." (Kogn. Int. Nr. 6, imptradTR)

I: „Traditionen sind ihr wichtig. Sie möchte jene Sitten und Gebräuche befolgen, die ihr durch die Religion oder die Familie überliefert wurden."
R: „Gleicht mir sehr! Oja. Das ist mir schon wichtig. Ich bin kein streng gläubiger Mensch, aber Traditionen und Sitten … ähm … pflege ich weiter, und schaue auch, dass die Generation danach – also mein Sohn – das weiter auch übernimmt. Ich finde das schon auch wichtig. Weil ohne Tradition haben wir keine Wurzeln. Da wissen wir nicht, wer wir sind. Und ich finde, das ist sehr wichtig. Das muss man schon wissen." (Kogn. Int. Nr. 7, imptradTR)

I: „Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden."
R: „Es ist mir jetzt schon wichtig. Ich mache bei den ganzen Ritualen mit, die wir haben: zu Ostern und zu Weihnachten. … Aber jetzt: Religion wirklich dran glauben? Dass Gott das alles erschaffen hat: das eher weniger. Aber ich respektiere das, und ich … halte mich daran, weil … es ist ein Teil von mir, es ist ein Teil meiner Kultur, meiner Familie und von meinem Volk. Ja. Hat sicher einen Grund, warum sich das so lange gehalten hat." (Kogn. Int. Nr. 8, imptradTR)

Was bei Wienern ohne Migrationshintergrund sehr kritisch gesehen wird, erfährt eine völlig andere Wertigkeit bei den letzten Befragten.

An diesem Indikator kristallisierte sich auch ein Grundproblem der Anwendung des Porträt heraus. Die Anweisung an die Interviewten lautet: „Hier beschreiben wir knapp einige Menschen. Sagen Sie mir bitte, ob ein solcher Mensch Ihnen gleicht oder nicht." In zwei Interviews bewerten die Befragten das Item abhängig davon, ob sie sich eine fiktive Person, auf die das Porträt zutrifft, vorstellen, oder Assoziationen, die durch das Item hervorgerufen werden, mit ihrem eigenen Leben vergleichen:

38 In der serbisch-orthodoxen Tradition wird in Gedenken an die Krippe, in die Christus nach seiner Geburt gelegt wurde, das Weihnachtsmahl auf Stroh verspeist.

I: „Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden."
R: „Gleicht mir. … Also ganz klar: Da sind wir im ganz konservativen Lager. Wir sprechen, also das Bild, das vermittelt wird, ist schon ein zugeknüpfter Klosterschüler, der weiß, was Recht und Ordnung ist, und wie man sich zu benehmen hat. Der einen gewissen Anstand hat, eine courtesy sozusagen. Ähm, höchstkatholisch ist. Also wenn man Religion hört, dann denkt man nicht an einen Buddhisten oder Hindu, sondern man denkt an einen Katholischen. Das ist, was wir da vermittelt bekommen. […] Ich bin nicht konform mit diesem Stereotyp, das als erstes aufscheint. Aber wenn ich es auf mich beziehe, habe ich absolut etwas dafür, wo ich absolut unsere Sitten, wie Familygatherings, wie gemeinsam Ostern feiern, gemeinsam Weihnachten feiern, ähm, die Art, wie miteinander umgegangen wird, da sind da so viele Sachen, die ich schätze. Und meine Form von Religion und Spiritualität ist ja eine viel größere, und da gibt's wie gesagt die Schnittpunkte, die ich sehr pflege. Deswegen muss ich sagen, auf mich bezogen: Ja, schon, nicht fanatisch, aber es gibt Traditionen, die ich schätze.
I: „Gleicht dir?"
R: „Und mit ganz klarer Abzweigung zu dem Bild, das als erstes vermittelt wird. […] also hier muss ich mich wiederum ganz stark trennen von dem, sozusagen, was ich als erstes bekomme. Das ist, sozusagen, Klosterfrau Melissengeist: Nein. Aber auf mein Umfeld gibt es sehr wohl Traditionen, die ich pflege."
I: „Wieso gehst du hier nach dem zweiten Impuls, und bei den anderen Fragen bist du nach dem ersten Impuls gegangen?"
R: „Ich bin immer so ausgegangen: Dort, wo ich den ersten Impuls als allgemein, sozusagen, empfunden hab und der sich nicht so stark abgehoben hat von dem, was ich mache, konnte ich mit dem Allgemeinen gehen. Und dort, wo das Allgemeine und das, was ich mache, das Gegenteil ist, musste ich wechseln. […] Also dort, wo Verbindungen komplett nicht waren, musste ich teilen. Das war aber nur bei drei Fragen der Fall. […] Bei drei Fragen, war es so unterschiedlich, dass man, dass ich da eine Spezifizierung vornehmen musste." (Kogn. Int. Nr. 1, imptradTR)

Befragter Nr. 1 assoziiert mit dem Porträt ein stark emotionales Bild einer Person. Dieses stellt treffend dar, was mit dem Item gemessen werden sollte. Anstatt sich von dem Bild leiten zu lassen, eröffnen die „Traditionen" eine weiteren Beurteilungsschritt im kognitiven Ablauf des Antwortens: Traditionen haben bei dem Befragten neben einer stark negativen Konnotation auch die Bedeutung von Traditionen (Umgangsformen), die innerhalb einer Familie bestehen, die aber nicht unbedingt mit einer größeren Gemeinschaft oder sogar Gesellschaft geteilt werden. Der Respondent spricht als Beispiel für „gute" Traditionen „die Art, wie miteinander umgegangen wird" an, die wahrscheinlich eine sehr spezifische „Tradition" ist. Diese zusätzliche Information wird auch verarbeitet. Warum das passiert, spricht dieselbe Person an anderer Stelle an:

R: „[…] Dort, wo die Begriffe so einfach hingeworfen sind, musst du mit dem ersten intuitiven Impuls gehen, und der ist oft ein sehr archetypisches Allgemein-

> bild, von dem sich fast jedes Individuum trennen möchte. Fast niemand will einer von denen sein. Jeder will das sein, was er selber ist. Nicht?“ (Kogn. Int. Nr. 1)

Befragter Nr. 1 beschreibt in dieser Passage, dass für ihn bei problematischen Items sehr negative Bilder evoziert wurden, von denen er sich distanzieren will. Dieser Zustand ist in seinem kognitiven Interview weitere drei Mal von ihm erwähnt worden (ipfruleCO, ipstrgvSE, ipbhprpCO und imptradTR). Alle vier Indikatoren finden sich in der Liste jener Items, die in der Analyse der fehlenden Werte überdurchschnittlich viele Missings erzeugten.

Es ist schwierig, in den tatsächlichen Prozess der Beantwortung während des Surveys hineinzusehen, um feststellen zu können, ob diese zwei unterschiedlichen Abläufe so, anders oder gar nicht bei Interviewten auftreten. Dennoch, wie bereits angesprochen, machte auch eine weitere Respondentin dieses Dilemma der Beantwortung deutlich:

> I: [Respondentin hat ihre Wertung mit „gleicht mir“ abgegeben] „Was sind denn Traditionen?“
> R: „Ja, das kann … was sein wie die Fastenzeit zu Ostern. Oder eben das Ostereier -Anmalen im Allgemeinen, also religiös. Es kann eben sein, dass sich meine Familie immer trifft am ersten Juliwochenende, weil da alle meine Familienmitglieder Geburtstag haben. … Das würde ich so als Tradition einschätzen, wenn man gemeinsam Eis essen geht. Das ist eine Tradition.“ […]
> I: „Könntest du dir jemanden vorstellen, der das so ganz von sich selber glaubt?“
> R: „Habe ich jetzt keinen, aber ich finds relativ ideenlos, wenn man immer nur dasselbe macht. […] Na, ideenlos, das ist dann auf gar keinen Fall mehr die 2 [Ähnlichkeit mit Portrait].“
> I: „Hat das jetzt das etwas verändert?“
> R: „Nein, zu dem Statement bin ich immer noch eine 2. […] Weil grundsätzlich finde ich Traditionen gut. […] Aber nicht alle.“
> I: „Und wenn man jetzt, sozusagen, ein Bild von jemandem hernimmt, der das macht … dann hättest du gesagt: der ist ideenlos?“
> R: „Ein bisschen. Ja. […] Ja, der ist sehr starr und unbeweglich.“
> I: „Und wie ähnlich wärest du so einem Menschen?“
> R: „[…] Wenn ich mir den jetzt gegenüberstelle, dann sag ich: dem ähnel ich … 3-4. […] Wenn ich mich mit dem Statement vergleiche, dann ist das eine 2.“
> I: „Da [beim Statement] nimmst du dir dann Worte heraus, so wie du sie interpretieren würdest und aus deinem Leben heraus siehst?“
> R: „Genau. Das eine, da vergleiche ich mich mit Situationen, in denen ich war. Und im Anderen vergleiche ich mich mit einem – ich weiß nicht … [Stereotyp].“ (Kogn. Int. Nr. 10, imptradTR)

Respondentin Nr. 2 beschreibt zwei unterschiedliche Bewertungsstrategien (sich eine Person vorzustellen oder Assoziation mit dem eigenen Leben zu vergleichen), die so von anderen Befragten auch verwendet wurden. Über alle Items und Befragten betrachtet, geht die Tendenz eher in die Richtung, dass Assoziationen mit dem eigenen Leben „verglichen“ werden, und dass konkrete Personen seltener in den Bewertungsprozess fließen. Prinzipiell sollte sich

daraus kein Problem ergeben, wenn sich Assoziationen und Personen im Einklang befinden. Problematisch kann dieser Prozess werden, wenn Befragte tendenziell wegen der Ablehnung eines besonders prominenten Reizes auf den „netteren“ Stimulus ausweichen und auf jenen antworten.

Zusammengenommen ergeben die Befunde, dass das Item zurecht eine niedrige Reliabilität und damit einhergehend eine besonders niedrige formale Validität aufweist.

4.5 Fazit

Die Tabelle 16 (Seite 130) fasst die in dieser Arbeit besprochenen Problemfelder und die betroffenen Items zusammen. Es sei noch einmal darauf hingewiesen, dass sie nicht umfassend die Ergebnisse aus den kognitiven Interviews wiedergibt.

Methodologische Probleme

Doppelte Stimuli wurden für die Items ipfruleCO, ipbhprpCO, iprspotPO und imptradTR gefunden. Die Stimuli waren unterschiedlicher Natur: iprspotPO, ipbhprpCO und ipfruleCO wiesen alle Teilaspekte auf, die innerhalb des Items anders bewertet wurden als der Rest. Es ist schwer nachzuprüfen, welcher dieser Stimuli letztendlich bei der Mehrzahl der Befragten überwogen hat. Da ipbhprpCO zu denjenigen Items zählt, die im obersten Quartil der Reliabilitäten vorzufinden sind, ist zweifelhaft, ob doppelte Stimuli automatisch zu schlechten Messeigenschaften führen, zumal iprspotPO und ipfruleCO auch noch in andere Probleme verwickelt sind. Für imptradTR waren die doppelten Stimuli unterschiedliche Formen, in die das Porträt übersetzt wurde.

Normen stellen ein Prolbem dar, wie sich bei zwei Items (iphlpplBE und iprspotPO) herausstellte. Das wirft die Frage auf, ob eine Unterscheidung auf einer operationalen Ebene sinnvoll ist, weil tatsächlich andere Dimensionen gemessen werden. Eine Antwort darauf kann in dieser Arbeit nicht gefunden werden.

Ob *Kontext* einen Einfluss auf die Messung von Werten hat, ist Gegenstand der derzeitigen wissenschaftlichen Debatte. Hechter (1999) und Schwartz vertreten hier zwei ganz gegensätzliche Positionen: während die eine Seite behauptet, Werte wären transsituational und dadurch auch als solche zu erheben, sieht die andere Werte als ausschließlich kontextbasiert vermittelbar. Wenn man sich die Antworten und Antwortstrategien der Teilnehmer ansieht, dann fällt auf, dass Kontext in einem bestimmten Rahmen schlagend wird. Wenn die Befrag-

ten in ihrer Erinnerung kramen, um geeignete Situationen zu finden, die durch das Porträt evoziert werden, dann geschieht dies nicht nach einheitlichen Vorgaben. Und hier geht es gar nicht einmal darum, dass die Befragten *satisficing* anstatt *optimizing* (Krosnick, 1999), betreiben. Die Interviews zeigen, dass sie sehr wohl nachdenken. Aber jeder und jede tut dies in einem eigenen Rahmen. Abhängig von der Verfassung der Befragten und den Assoziationen, die das Porträt auslöst, könnten ganz unterschiedliche Wertungen zustandekommen, zumal Werte scheinbar eher affektiv als kognitiv in uns verankert sind (siehe Kapitel 3.1.1).

Deskriptive Statistik

Die meisten Items sind in Bezug auf ihre Wölbung unauffällig. Einzig das Item iplylfrBE sticht heraus: Es ist stark leptokurtisch. Die kognitiven Protokolle zeigten, dass es in sich eine Tautologie darstellt und noch dazu ein Porträt vermittelt, dem man nur schwer „nicht gleichen“ kann. Das zeigen ebenso Mittelwert und Schiefe. Letztere ist unter allen Items am stärksten ausgeprägt. Es ist fraglich, ob es ein geeigneter Indikator für eine kompakte Version des PVQ darstellt.

Die einzigen linksschiefen Items sind imprichPO, ipfruleCO und ipadvntST. Diese Items werden im Gegenzug zu den übrigen überwiegend als unähnlicher bewertet. Die kognitiven Protokolle könnten so interpretiert werden, dass dies weniger mit einer Unähnlichkeit mit den porträtierten Personen als vielmehr mit einer Ablehnung gegenüber gewissen Inhalten zusammenhängt. Wie versucht wurde darzulegen, macht es einen konzeptionellen Unterschied, ob ein Porträt als unähnlich eingestuft wird, oder ob Aspekte Assoziationen hervorrufen, die abgelehnt werden. Es hat sich gezeigt, dass diese Items innerhalb des CFA-Modells problematisch sind. Im SEM-Ansatz lässt sich das durch Kreuzladungen ausgleichen. Diese wären aber auf einen methodischen und nicht inhaltlichen Grund zurückzuführen.

In der Analyse der fehlenden Werte stellte sich heraus, dass weder Alter noch Geschlecht einen signifikanten Einfluss auf ihre Anzahl haben. Für Bildung gilt, dass dieser Zusammenhang zwar signifikant, aber nur sehr schwach ausgeprägt ist. Hingegen scheinen die Variablen selbst einen Einfluss darauf zu haben, ob sie abgelehnt werden. Überdurchschnittlich ausgeprägt war das bei ipfruleCO, ipudrstUN, ipstrgvSE, ipbhprpCO, iplylfrBE, iprspotPO und imptradTR der Fall. Bis auf ipudrstUN wurden die Items und ihre Probleme, die zu einem eröhten Ausfall führen können (vgl. Krosnick, 1999), bereits ausführlich besprochen. Speziell bei jenem Item bemäkelten mehrere der Befragten, dass zwar prinzipiell Verständnis für andere Menschen gegeben sei, dieses aber auch seine Grenzen habe. Die porträtierte Person

schien manchen Befragten zu naiv. Aus diesem Widerspruch könnte sich ein erhöhter Ausfall erklären.

	Methodologische Probleme					Deskriptive Statistik			Reliabilitäten/ Kommunalitäten		
	Skalierung	Doppelte Stimuli	Normen	Kontext	Soziale Erwünschtheit	Kurtosis	Schiefe	Missings (überdurch-schnittlich)	unterstes Quartil	in 4-Faktormodell unterschätzt	in EFA über-/ unterschätzt
ipcrtivSD	siehe Schiefe				Gilt für alle? Klärung durch Theorie nötig (siehe Kapitel 3.1.5)					x	
imprichPO							rechtsschief		x		überschätzt
ipeqoptUN											
ipshabtAC											
impsafeSE										x	
impdiffST											
ipfruleCO		x					rechtsschief	x	x		
ipudrstUN								x			
ipmodstTR									x		
ipgdtimHE										x	
impfreeSD										x	
iphlpplBE			x								
ipsucesAC											
ipstrgvSE								x		x	
ipadvntST							rechtsschief			x	überschätzt
ipbhprpCO		x						x			
iprspotPO		x	x					x	x		überschätzt
iplylfrBE						spitz	linksschief	x			
impenvUN											
imptradTR		x						x	x		
impfunHE										x	

Reliabilitäten

Items mit der geringsten Messqualitäten sind imprich,PO ipfruleCO, ipmodstTR, iprspotPO und imptradTR. Probleme, die in größerer Tiefe mittels kognitiver Interviews analysiert wurden, sind der Tabelle 17 zu entnehmen.

	Problemdimensionen				
imprichPO	Teilaspekte abgelehnt	mangelnder Bezug zur Lebenswelt	doppelte Stimuli	Messung von „Streben nach Glück“?	dient nicht mehr als Standard zur Beurteilung, sondern wird selber beurteilt
ipfruleCO	Teilaspekte abgelehnt	sehr schwieriges Item; mangelnder Bezug zur Lebenswelt	doppelte Stimuli		Normen
ipmodstTR					*Trait*/Charakterzug
iprspotPO	Teilaspekte mit gegensätzlicher Konnotation		doppelte Stimuli		Normen
imptradTR	Teilaspekte abgelehnt	Bezug zu Lebenswelt je nach Population und Teilaspekt	zwei Formen von Porträts		

Tabelle 17: Problemkategorien für ausgewählte Items aus kognitiven Protokollen

Eines der Ziele dieser Arbeit bestand darin, unterschiedliche quantitative Methoden und ihre Ergebnisse zu vergleichen. EFA und CFA unterscheiden sich in der Anzahl der Faktoren, die mit gutem Fit extrahiert/modelliert werden können. Anderseits stimmen die Reliabilitäten (vier Faktoren, CFA) und Kommunalitäten (EFA) weitgehend überein. Ausnahmen gab es bei drei Items (imprichPO, ipadvntST, iprspotPO), bei denen die EFA die Ladungen mit einer Differenz von ca. 0,1 überschätzte (vgl. Tabellen 21 und 26 im Anhang). EFA und CFA unterscheiden sich auch in Bezug auf die Faktorkorrelationen: während die EFA sowohl mit Mplus als auch mit SPSS gerechnet ein diffuses Bild ergeben, ist es bei CFA klar und stimmt mit der Theorie von Schwartz mit Einschränkungen überein.

Beurteilung der Messmodelle

In der Fassung des ESS sind für Österreich für die zweite Welle des ESS maximal sieben Wertetypen von einander unterscheidbar. Schmidt et al. (2007) und Davidov et al. (2008) haben gezeigt, dass drei Wertetypen zusammengeführt werden müssen, um Resultate in der konfirmatorischen Faktorenanalyse zu berechnen. In der vorliegenden Arbeit konnten diese Ergebnisse bestätigt werden. *Conservation* und *openness to change* lassen sich jeweils in die Faktoren Security und Conformity-Tradition einerseits, und Self-direction, Stimulation und Hedonism andererseits teilen. Letztere Aufteilung ist allerdings nicht zufriedenstellend, was es problematisch macht, sie als getrennte Faktoren wahrzunehmen.

Im paarweisen Vergleich der *higher-order values* stellte sich heraus, dass von den 19 schwersten Modellmissspezifikationen neun mit jenen Indikatoren zu tun haben, die die schlechtesten Messeigenschaften aufweisen. Das entspricht der Hälfte der durch das Modell erzeugten am stärksten über- bzw. unterschätzten Parameter. Weitere fünf entfallen auf Items, die entweder zu dem obersten Quartil (wie ipadvntST) oder den mittleren 50% (wie ipstrgvSE und impsafeSE) gehören. Das wiederum bedeutet, dass gute Messeigenschaften alleine nicht reichen, um ein Modell gut darzustellen. Es scheint deshalb essentiell zu sein, nicht auf einer quantitativen Ebene bei der Beurteilung von Indikatoren und Modellen stehenzubleiben, sondern qualitativ in die Tiefe zu gehen.

5 Schlussbetrachtung

Bei Mixed-Methods ist es immer spannend, sich am Ende zu überlegen, zu welchem Anteil QUALI und QUANTI sich überlagern oder in der Analyse auseinanderstreben. Im vorliegenden Forschungsdesign ist diese Frage sinnlos, da nur überlappende Ergebnisse ausgewertet wurden. Hingegen lässt sich sehr wohl über den Mehrwert des Ansatzes diskutieren. Aus meiner Sicht beschreibt die vorliegende Arbeit einen gangbaren Weg, wie QUANTI und QUALI für die Entwicklung eines Erhebungsinstrumentes zusammengeführt werden können.

Die vorliegende Arbeit ist keine umfassende Betrachtung und Bewertung der Skala. Das sehr breit angelegte und arbeitsintensive Design lässt bei gegebenen Ressourcen nur eine begrenzte Bearbeitungskapazität zu. Das mag auch ein Nachteil des Designs sein. Um die Stärken beider Methoden zu verknüpfen, wird es außerdem von den Analysten verlangt, sich gänzlich auf das jeweilige Paradigma einzulassen. Dies wurde nur in bescheidenem Maß für den qualitativen Teil erreicht.

In dieser Arbeit ging es in der Verknüpfung der Methoden sehr stark um einen Kompromiss zwischen Methodenvielfalt und Erkenntnisinteresse, konkret darum, wie man im Sinne einer soziologischen Betrachtungsweise quantitative Ergebnisse als kodifiziertes soziales Handeln *deutend verstehen* kann. Es wurde auch versucht, die Komplexität abzubilden, die mit der Erfassung von menschlichen Werten einhergeht, indem die angewandte Methodenvielfalt gezeigt wurde. Auch machten theoretische Widersprüche in der Erfassung von menschlichen Werten sichtbar, womit sich Forschende auf dem Gebiet konfrontiert sehen. Diese vielschichtigen Gründe machen vielleicht auch verständlicher, warum ein Zugang zu ihrer Erfassung oder der Beurteilung derselben vielschichtig sein könnte, wenn nicht sogar sein sollte.

5.1 Bewertung der Methode

Während in Publikationen und in der sozialwissenschaftlichen Forschung üblicherweise kognitive Interviews und konfirmatorische Faktorenanalyse getrennt behandelt werden (Schmidt et al., 2010: 1148), ist die vorliegende Arbeit neben der von Latcheva (2009) eine der wenigen, die beides integriert. Auch in der gängigen Literatur zu Mixed-Methods (Teddlie und Tashakkori, 2009; 1998; Creswell und Plano Clark, 2007) findet sich kein beschriebenes Design, das konkret auf die vorliegende Arbeit zutreffen könnte, weshalb dafür ein eige-

ner Begriff erfunden wurde (*Backtracking*). Die Forschungsfrage, wie und ob quantitative und qualitative Methoden verknüpft angewendet werden, um Messinstrumente zu entwickeln und/oder sie zu evaluieren, scheint darin kein Thema zu sein. Es wirkt fast so, als sei diese Forschungsfrage nicht Teil des gängigen Universums von Mixed Methods. Das ist erwähnenswert, denn die einzelnen Methoden (CFA und kognitive Interviews), die in dieser Arbeit verwendet werden, sind jede für sich ca. 20-30 Jahre alt und damit nicht „neu" – zumindest in keinem absoluten Sinn. Auch ist die gemeinsame Verwendung von QUALI und QUANTI für die Entwicklung eines Fragebogens keine Neuheit, sondern fixer Bestandteil der Praxis.

Eine Eigenart dieser Verknüpfung ist vielleicht, dass sie stark durch das quantitative Paradigma gefärbt ist – was gerade auch für die kognitiven Interviews gilt. Die gemeinsame Anwendung wurde spezifisch für ein quantitatives Feld entwickelt. Beispielhaft für die Sichtweise, die in diesem Paradigma herrscht, ist womöglich bei Tourangeau et al. zu erkennen, die es geradezu als Schwäche sehen, „[…] that cognitive interviews become the basis for nonquantitative essentially impressionist analyses of the interviews – hardly a desirable situation" (Tourangeau et al., 2000: 327). Daher fordern sie für die quantitativen Verfahren rigorose Methodenstrenge in Form von Codierschlüsseln und die Entwicklung von Standards. Diese Rhetorik ist wohl schwer mit jener des qualitativen Paradigmas zu verknüpfen, das einen Schwerpunkt auf semi- bis unstrukturierte Datenerhebung, eine ganzheitliche Beschreibung der Situation und einer engen, empathischen Beziehung zwischen Interviewer und Interviewten legt (Teddlie und Tashakkori, 2009). Womöglich ist das ein Grund, warum diese spezielle Verknüpfung nicht als Mixed-Methods anerkannt wird. Vielleicht sind 20 Jahre auch noch nicht lang genug, damit Neuheiten aus einer Forschungstradition in einer anderen ankommen.

Angebracht scheint die Verknüpfung dennoch zu sein. Wie Latcheva (2009) beschrieben hat, sind manche quantitative Kriterien wie eine minimale std. Ladung von 0,4 nicht hinreichende Indikatoren, um verzerrte Messungen angemessen wahrzunehmen. Ebenso konnten wir sehen, dass selbst Indikatoren mit guten Messeigenschaften in Missspezifikationen verwickelt sein können. Wäre es in diesen Momenten nicht nützlich zu wissen, worauf diese zurückzuführen sind, um entsprechende Gegenstrategien umzusetzen? Andererseits: Wie können wir sicher sein, ob identifizierte Probleme tatsächlich solche sind? Um eine Verknüpfung von QUALI und QUANTI für die vorliegende Arbeit zu begründen, möchte ich noch einmal über Vor-, Nachteile und ergänzende Wertschöpfung für beide Seiten reflektieren:

Quantitative Analysen erlauben es, komplexe Zusammenhänge stark reduziert darzustellen. Es ist faszinierend zu beobachten, wie aus einer Reihe von Indikatoren und einer großen

Menge an repräsentativen Antworten dazu ein hoch abstraktes Modell der erfahrbaren Wirklichkeit berechnet werden kann. Gerade der SEM-Ansatz bietet hier durch seine Verknüpfung zwischen mathematischen Gleichungssystemen und einer graphischen Übersetzung die Möglichkeit, sehr intuitiv hoch komplexe Zusammenhänge darzustellen.

Dennoch lässt sich sagen, dass all diese Abstraktion den Blick auf Informationen verklärt, die für das Verständnis eines Modells notwendig scheint. Zum Beispiel ist eine Faktorladung relativ bedeutungslos, wenn die ihr zugrunde liegende Dimension inhaltlich nicht vollständig erfasst wird oder der Entstehungszusammenhang zwischen Frage und Antwort nicht geklärt ist. Was steckt denn alles in dieser verknappten Darstellung durch eine Zahl? Was wurde zuvor durch die Befragten oder die Analyse zusammengeworfen, was qualitativ ganz unterschiedlich zu bewerten ist? Qualitative Verfahren wie kognitive Interviews können hier helfen, das zu expandieren, was zuvor in einer einzigen Zahl zusammengefasst wurde. Das, was sehr abstrakt dargestellt wurde, wird in einem qualitativen Interview aufgebrochen und in individuelle Einzelbedeutungen zerlegt. Die Analyse qualitativer Protokolle stellt damit auch gar keinen Anspruch auf Allgemeingültigkeit, sondern die Interviews sprechen nur beispielhaft für die Befragten. Gleichzeitig ist das Protokoll aber reich an Inhalten, die den Forschenden sowohl Einblicke in kognitive Prozesse als auch Hinweise auf verschiedene inhaltliche Dimensionen liefern können. Sie sind wie Spiegel, in denen sich das kognitive Geschehen der Befragten, das üblicherweise für jeden außenstehenden Beobachter unsichtbar ist, für andere sichtbar reflektiert wird.

5.2 Probleme und Grenzen der Arbeit

Die gesamte Skala umfasst 21 Items, von denen hier nur fünf in der Tiefe analysiert wurden. Und selbst diese Analyse ist nicht vollständig, denn die Protokolle der kognitiven Interviews lassen noch mehr Analysen zu und öffnen weitere Bereiche, wo sie Items als problematisch identifizieren. Ebenso könnten mit Hilfe eingehenderer Testung der einzelnen Messmodelle spezifischere Aussagen zu problematischen Items gemacht werden, um die Vorauswahl für die qualitativen Protokolle treffsicherer zu gestalten. Für die quantitative Analyse wäre es auch von Vorteil gewesen, über mehr Items in der Skala zu verfügen, damit die Qualität der einzelnen Messmodelle besser zu überprüfen wäre.

Wie die Arbeit zeigte, ist die Vermittlung von Ergebnissen für QUANTI relativ effizient zu gestalten, während sie für QUALI extensiv ausfällt. Letzteres bedeutet einen hohen Aufwand in der Verarbeitung, da immer sehr viel Arbeit anfällt, die Aussagen der Befragten zu

interpretieren und diese Interpretation immer auch glaubwürdig dem Leser vermittelt werden muss. Dafür braucht es genügend Zitate, die nicht immer vorhanden sind. Dann wiederum passiert es, dass man als Interviewer sich sehr wohl erinnert (oder zu erinnern meint), wie ein bestimmtes Zitat zu interpretieren wäre, dies aber durch den reinen Wortlaut nicht unbedingt klar vermittelt werden kann.

Generell gilt für das Triangulationsdesign, dass Forschende im Verlauf des Forschungsprozesses entweder Ergebnisse der unterschiedlichen Methoden nicht konvergieren oder diese sich überhaupt nicht verknüpfen lassen. Dies war bei der vorliegenden Arbeit auch ein Thema: Die Fülle an Problemen in beiden Methoden machten das Feld undurchschaubar. Man sah gewissermaßen den Wald vor lauter Bäumen nicht mehr. Letztlich konnte das Mixed-Methods Design dennoch angewendet werden, weil die Fülle der Problemfelder mit einer Methode rigoros eingeschränkt wurde und anschließend nur konvergierende Ergebnisse mittels Methodenmix analysiert wurden. Jene, die nicht übereinstimmten, wurden gar nicht erst erwähnt.

Darüber hinaus liefert *Backtracking,* wie es in dieser Forschungsarbeit angewandt wurde, auch keine klare Anweisung dafür, was man machen könnte, um die Items zu verbessern. Da es sich um ein rein diagnostisches Instrument handelt, kann es keine Gewissheit liefern, wie ein „richtiges“ Porträt aussähe. Für die Formulierung neuer Porträts bräuchte es evidenzbasiertes Wissen zum Verständnis von Fragen (vlg. Saris und Gallhofer, 2007). Dennoch können die hier erzielten Resultate weiterverwendet werden: Sie liefern Ansätze, wo Änderungen wirksam sein könnten, und Kontextinformation, in welche Richtung sie gehen könnten.

6 Literaturverzeichnis

Ajzen, I. (1991). The Theory of Planned Behavior. *Oraganizational Behavior and Human Decision Processes*, *50*, 179-211.

Akremi, L., Baur, N., und Fromm, S. (Eds.). (2011). *Datenanalyse mit SPSS für Fortgeschrittene 1*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Backhaus, K., Erichson, B., und Weiber, R. (2011). *Fortgeschrittene Multivariate Analysemethoden*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Beierlein, C., Davidov, E., Schmidt, P., Schwartz, S. H., und Rammstedt, B. (k.A.). Testing the Discriminant Validity of Schwartz's Portrait Value Questionnaire Items. A Replication and Extension of Knoppen and Saris (2009). *Survey Research Methods*.

Bernard, M. M., Olson, J. M., und Maio, G. R. (2003). Effects of intorspection about reasons for values: extending research on values-as-truisms. *Social Cognition*, *21*(1), 1-25.

Bilsky, W., und Koch, M. (2000). On the content and structure of values: universals or methodological artefacts? *Paper presented at the Fifth International Conference on Logic and Methodology* (Vol. IV). Cologne, Germany.

Blair, Johnny, Conrad, F., Ackermann, A. C., und Claxton, G. (2006). The Effect of Sample Size on Cognitive Interview Findings. Url: http://www.abtassociates.com/presentations/aapor06_sample_size_cognitive_interviews.pdf (Zugriff: 7.10.2011)

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Brown, T. (2006). *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Buss, D. (1986). Can a social science be anchored in evolutionary biology? Four problems and a strategic solution. *Revue Europeenne des Sciences Sociales*, *24*, 41-50.

Campbell, D. T., und Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*, 81–105.

Computing Scores for the 10 Human Values. (k.A.). Url: http://ess.nsd.uib.no/ess/doc/ess1_human_values_scale.pdf (Zugriff: 20.3.2012)

Costa, P. T., und McCrae, R. R. (1989). *Manual for the NEO Personality Inventory: Five Factor Inventory/NEO-FFI.* Odessa, FL: PAR.

Creswell, J. W., und Plano Clark, V. L. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, California: Sage publication, Inc.

Crowne, D. P., und Marlowe, D. (1964). *The approval motive.* New York: Wiley.

Damarin, F., und Messick, S. (1965). *Response styles as personality variables: A theoretical integration*. Princton: Educational Testing Service.

Davidov, E, Datler, G., Schmidt, P., und Schwartz, S. H. (2011). Testing the invariance of values in the Benelux countries with the European Social Survey: Accounting for ordinal and nonnormal distribution of the data. In Jaak Billiet, E. Davidov, und P. Schmidt (Eds.), *Cross-Cultural Analysis: Methods and Applications* (pp. 149-168). Routledge.

Davidov, Eldad. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the european social survey. *Survey Research Methods*, *2*(1), 33-46.

Davidov, Eldad. (2010). Testing for comparability of human vlaues across countries and time with the third round of the European Social Survey. *Internation journal of comparative sociology*, *51*, 171-191.

Davidov, Eldad, und De Beuckelaer, A. (2010). How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument. *International Journal of Public Opinion Research*, *22*, 485-510.

Davidov, Eldad, Schmidt, P., und Schwartz, S. H. (2008). Bringing Values Back In. *Public Opinion Quarterly*, *72*, 420-445. Url: http://poq.oxfordjournals.org/content/72/3/420.abstract (Zugriff: 1.3.2012)

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

Eysenck, H. J., und Eysenck, S. B. G. (1964). *The manual of the Eysenck Personality Inventory.* London: U of London Press.

Feather, N. T. (1973). Effects of response anonymity on assessment of own and school value systems and satisfaction with school. *British Journal of Education Psychology*, *43*, 140-150.

Fischhoff, B., Slovic, P., und Lichtenstein, S. (1980). Knowing what you want: Measuring labile values. In T. S. Wallsten (Ed.), *Cognitive processes in choice and decision behavior*. Hillsdale, N.J.: Erlbaum.

Fontaine, J. R. J., Poortinga, Y. H., Delbeke, L., und Schwartz, S. H. (2008). Structural Equivalence of the Values Domain Across Cultures : Distinguishing Sampling Fluctuations From Meaningful Variation. *Journal of Cross-Cultural Psychology*, *39*, 345-365.

Goldsmith, R. E., Stith, M. T., und White, J. D. (1987). The Rokeach Value Survey and social desirability. *Journal of Social Psychology*, *127*, 553-554.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* (pp. 258–348). New York: Columbia University Press.

Hathaway, S. R., und McKinley, J. C. (1951). *MMPI manual.* New York: Psychological Corporation.

Hechter, M. (1992). Should Values Be Written out of the Social Scientist's Lexicon? *Sociological Theory*, *10*, 214-230.

Hechter, M. (1993). Values research in the social and behavioral sciences. In M. Hechter, L. Nadel, und R. Michod (Eds.), *The Origin of Values* (pp. 1-28).

Hechter, M., Ranger-Moore, J., Jasso, G., und Home, C. (1999). Do Values Matter? An Analysis of Advance Directives for Medical Treatment. *European Sociological Review*, *15*, 405-430.

Hitlin, S., und Piliavin, J. A. (2004). Values: Reviving a Dormant Concept. *Annual Review of Sociology*, *30*, 359-393.

Hofstede, G. (1980). *Culture's consequences: International dlferences in work-related values.* Beverly-Hills, CA: Sage.

Holm, K. (1970). Zuverlässigkeit von Skalen und Indizes. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *22*, 356-386.

Horn, J. L., und MacArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research,*, *105*, 117-144.

Hu, L., und Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Kelly, K., Silverman, B. I., und Cochrane, R. (1972). Social desirability and the Rokeach Value Survey. *Journal of Experimental Research in Personality*, *6*, 84-87.

Kilmann, R. H. (1975). A scaled-projective measure of interpersonal values. *Journal of Personality Assessment*, *39*, 34-40.

Kluckhohn, C. (1951). Values and value-orientations in the theory of action: An exploration in definition and classification. In T. Parsons und E. Shils (Eds.), *Toward a general theory of action.* (pp. 388-433). Cambridge. MA: Harvard University Press.

Knoppen, D., und Saris, W. E. (2009a). Do we have to combine Values in the Schwartz' Human Values Scale? Do we have to combine Values in teh Schwartz' Human Values Scale? A Comment on the Davidov Studies. *Survey Research Methods*, *3*, 91-103.

Knoppen, D., und Saris, W. E. (2009b) Evaluation of the Portrait Values Questionnaire using SEM: A New ESS Proposal. Paper prepared for the QMSS2 seminar at Bolzano (Bozen), Italy.

Konty, M. (2002). *Values, deviance and conformity: measuring values with the factorial survey method*. University of Arizona.

Kristiansen, C. M., und Zanna, M. P. (1991). Value relevance and the value-attitude relation: value expressiveness versus halo effects. *Basic and Applied Social Psychology*, *12*, 471–483.

Krosnick, J. a. (1999). Survey research. *Annual review of psychology*, *50*, 537-567.

Latcheva, R. (2009). Cognitive interviewing and factor-analytic techniques: a mixed method approach to validity of survey items measuring national identity. *Quality Quantity*, *45,* 1175-1199.

Lubke, G. H., und Muthén, B. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514-534.

Maio, G R, und Olson, J. M. (1998). Values as truisms: evidence and implications. *Journal of Personality and Social Psychology*, *74*, 294-311.

Maio, G. R., und Olson, J. M. (2000). What is a “value-expressive” attitude? In G R Maio und J. M. Olson (Eds.), *Why We Evaluate: Functions of Attitudes* (pp. 249–269). Mahwah, NJ: Lawrence Erlbaum.

McClelland, D. C. (1991). *The personal value questionnaire.* Boston: McBer und Company.

Melech, G. (2001). *Value development in adolescence.* The Hebrew University of Jerusalem.

Messick, S. (1995). Validity of Psychological Assessment. Validation of Inferences From Persons’ Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, *50*, 741-749.

Meuleman, B., Davidov, E., Schmidt, P., und Billiet, J. (2012). Social location and value priorities. A European-wide comparison of the relation between social-structural variables and human values. In O. Gabriel und S. Keil (Eds.), *Society and democracy in Europe.* Routledge.

Mohler, P. P., und Wohn, K. (2005). Persönliche Wertorientierungen im European Social Survey. *ZUMA-Arbeitsbericht Nr. 2005/01*.

Moosbrugger, H., und Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion*. (H. Moosbrugger und A. Kelava, Eds.)*Psychologie Schweizerische Zeitschrift Für Psychologie Und Ihre Andwendungen*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-20072-4

Morris, C. W. (1956). *Varieties of human value.* Chicago: University of Chicago Press.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, *40*, 120-123.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrica*, *54*, 557-585.

Muthén, B., du Toit, S. H. C., und Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Url: http://gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf (Zugriff: 1.1.2012)

Paulhus, D. (2002). Socially desirable responding. In H. I. Braun, D. N. Jackson, und D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement.* Routledge.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. Robinson, P. Shaver, und L. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Peytcheva, E. A. (2008). *Language of administration as a source of measurement error: Implications for surveys of immigrants and cross-cultural survey research*. University of Michigan.

Porst, R. (2011). *Fragebogen. Ein Arbeitsbuch.* VS Verlag für Sozialwissenschaften.

Presser, S., und Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, *24*, 73-104.

Prüfer, P., und Rexroth, M. (2000). Zwei - Phasen - Pretesting. *ZUMA-Arbeitsbericht Nr. 2000*, (8), 1-21.

Prüfer, P., und Rexroth, M. (2005). Kognitive Interviews. *ZUMA-How to*, *15*.

Roccas, S., Sagiv, L., Schwartz, S. H., und Knafo, a. (2002). The Big Five Personality Factors and Personal Values. *Personality and Social Psychology Bulletin*, *28*, 789-801.

Rokeach, M. (1973). *The nature of human values*. The Free Press.

Rothgeb, J., Willis, G., und Forsyth, B. (2001). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results. *2001 Proceedings of the Section on Survey Methods.* Alexandria, VA: American Statistical Association.

Sackheim, H. A., und Gur, R. C. (1978). Self-deception, other-deception and consciousness. In G. E. Schwartz und D. Shapiro (Eds.), *Consciouness and self-regulation: Advances in research* (pp. 139-197). New York: Plenum Press.

Saris, W. E., und Gallhofer, I. N. (2002). Report on the MTMM experiments in the pilot studies and Proposals for Round 1 of the ESS. Url: http://www.europeansocialsurvey.org (Zugriff: 7.10.2011)

Saris, W., und Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley-Interscience.

Schafer, J. L., und Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schmidt, P., Bamberg, S., Davidov, E., Hermann, J., und Schwartz, S. H. (2007). Die Messung von Werten mit dem “Portraits Value Questionnaire". *Zeitschrift für Sozialpsychologie*, *38*, 261–275.

Schmidt, Peter, Herrmann, J., und Kelle, U. (2010). Introduction. *Quality und Quantity*, *45*(6), 1145-1150.

Schnell, R., Hill, P. B., und Esser, E. (2008). *Methoden der empirischen Sozialforschung*. Oldenburg.

Schwartz, S H, Melech, G., Lehmann, A., Burgess, S., Harris, M., und Owens, V. (2001). Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-Cultural Psychology*, *32*, 519-542.

Schwartz, S H, & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross Cultural Psychology*, *32*, 268-290.

Schwartz, S. H., und Bilsky, W. (1987). Toward a psychological structure of human values. *Journal of Personality and Social Psychology*, *53*, 550-562.

Schwartz, S. H., und Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of Personality and Social Psychology*, *58*, 878-891.

Schwartz, Shalom H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Advances in Experimental Social Psychology*, *25*, 1-65.

Schwartz, Shalom H. (1994a). Are There Universal Aspects in the Structure and Contents of Human Values? *Journal of Social Issues*, *50*, 19-45.

Schwartz, Shalom H. (2003). A Proposal for Measuring Value Orientations across Nations. http://www.europeansocialsurvey.org/index.phpoption=com_docman&task=doc_view&gid=126&Itemid=80 (Zugriff: 22.11.2011)

Schwartz, Shalom H, und Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal or Research in Personality*, *38*, 230-255.

Schwartz, Shalom H, und Sagiv, L. (1995). Identifying Culture-Specifics in the Content and Structure of Values. *Journal of Cross-Cultural Psychology*, *26*, 92-116.

Schwartz, Shalom H. (2007). Value orientations: measurement, antecedents and consequences across nations. In R. Jowell, C. Roberts, R. Fitzgerald, und G. Eva (Eds.), *Measuring Attitudes Cross-Nationally. Lessons From the European Social Survey.* (pp. 169-204). Sage publications Ltd.

Schwartz, Shalom H., Verkasalo, M., Antonovsky, A., und Sagiv, L. (1997). Value priorities and social desirability: Much substance, some style. *British Journal of Social Psychology*, *36*, 3-18.

Scott, J., und Marshall, G. (2009). A Dictionary of Sociology. Retrieved from http://www.oxfordreference.com (Zugriff: 3.32012)

Seligman, C., und Katz, A. N. (1996). The Ontario Symposium: The Psychology of Values. In C. Seligman, J. M. Olson, und M. P. Zanna (Eds.), (pp. 53-75). Mahwah, NJ: Lawrence Erlbaum.

Spini, D. (2003). Measurement Equivalence Of 10 Value Types From The Schwartz Value Survey Across 21 Countries. *Journal of Cross-Cultural Psychology*, *34*, 3-23.

Steinmetz, H., Baeuerle, N., und Isidor, R. (k.A.). Testing the Circular Structure of Human Values: A Meta-Analytical Structural Equation Modeling Approach. *Survey Research Methods*.

Steinmetz, H., Schmidt, P., Booh, A. T., Wieczorek, S., und Schwartz, S. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality und Quantity*, *43*, 599-616.

Tan, L., und Grace, R. C. (2008). Social desirability and sexual offenders: a review. *Sexual abuse : a journal of research and treatment*, *20*, 61-87.

Tashakkori, A., und Teddlie, C. (1998). *Mixed Methodology*. Thousand Oaks, California: Sage publication, Inc.

Teddlie, C., und Tashakkori, A. (2009). *Foundations of Mixed Methods Research. Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Thousand Oaks, California: Sage publication, Inc.

Tourangeau, R, Rips, L. J., und Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tourangeau, Roger. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research*, *15*, 3-7.

Van de Vijver, F., und Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, California: Sage.

Vandenberg, R. J., und Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Verkasalo, M., Lönnqvist, J.-E., Lipsanen, J., und Helkama, K. (2009). European norms and equations for a two dimensional presentation of values as measured with Schwartz ' s 21-item portrait values questionnaire. *European Journal of Social Psychology*, *39*, 780-792.

Weber, M. (1980). *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie.* (J. Winckelmann, Ed.) (5th ed.). Tübingen: J. C. B. Mohr (Paul Siebeck).

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality und Quantity*, *38*, 681-702.

Wiggins, J. S. (1959). Interelationships among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology*, *23*, 419-427.

Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In W. Grove und D. Cicchetti (Eds.), *Thinking clearly about Psychology: Essays in honor of Paul Meehl* (pp. 89-113). Minneapolis: University of Minnesota Press.

Williams. (1968). Values. In E. Sills (Ed.), *International Encyclopedia of the Social Sciences.* New York: Macmillan.

Willis, Gordon B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage publication.

Willis, G. B., Schechter, S., und Whitaker, K. (2000). A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell Us? *American Statistical Association, Proceedings of the Section on Survey Research Methods*.

Wilson, T. D., Dunn, D. S., Kraft, D., und Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 287-343). Orlando, FL: Academic Press.

Yan, T., Kreuter, F., und Tourangeau, R. (2010). Evaluating Survey Questions : A Comparison of Methods. Url: http://www.jpsm.umd.edu/jpsm/people/faculty/rtourangeau/Evaluation of Survey Questions A Comparison of Methods.pdf (Zugriff: 8.3.2012)

7 Anhang

Die Batterie der Items wird mit der Anweisung „Hier beschreiben wir knapp einige Menschen. Sagen Sie mir bitte, ob ein solcher Mensch Ihnen gleicht oder nicht.“ eingeleitet. Für Frauen gibt es eine eigene, sprachlich an das Geschlecht angepasste Version.

Nr.	Item	Korrespondierender Wertetyp	Wortlaut
1	ipcrtiv	SD	Es ist ihm wichtig, neue Ideen zu entwickeln und kreativ zu sein. Er unternimmt alles gerne auf seine eigene Art und Weise.
2	imprich	PO	Reichtum ist ihm wichtig. Er möchte viel Geld und Luxusgüter besitzen.
3	ipeqopt	UN	Für ihn ist es wichtig, dass jeder Mensch auf dieser Welt gleich behandelt werden sollte. Er glaubt, dass jeder Mensch die gleichen Chancen im Leben haben sollte.
4	ipshabt	AC	Es ist ihm wichtig, seine Fähigkeiten unter Beweis zu stellen. Er möchte, dass ihn Leute für das bewundern, was er tut.
5	impsafe	SE	Es ist ihm wichtig, in einer sicheren Umgebung zu leben. Er vermeidet alles, das seine Sicherheit gefährden könnte.
6	impdiff	ST	Er liebt Überraschungen und sucht immer nach Neuem. Für ihn ist es wichtig, im Leben möglichst viele verschiedene Dinge zu unternehmen.
7	ipfrule	CO	Er glaubt, dass Leute das machen sollten, was man ihnen sagt. Er meint, dass Leute sich immer und überall an Regeln halten sollten, selbst wenn es niemand sieht.
8	ipudrst	UN	Es ist ihm wichtig, Leuten zuzuhören, die von ihm verschieden sind. Auch wenn er mit ihnen nicht übereinstimmt, möchte er sie trotzdem verstehen.
9	ipmodst	TR	Es ist ihm wichtig, zurückhaltend und bescheiden zu sein. Er versucht, keine Aufmerksamkeit auf sich zu lenken.
10	ipgdtim	HE	Spaß zu haben ist ihm wichtig. Er verwöhnt sich gerne.
11	impfree	SD	Es ist ihm wichtig, selbst zu entscheiden, was er tut. Er ist gerne frei und unabhängig von anderen.
12	iphlppl	BE	Es ist ihm wichtig, den Menschen in seiner Umgebung zu helfen. Er möchte sich um ihr Wohlergehen sorgen.
13	ipsuces	AC	Es ist ihm wichtig, sehr erfolgreich zu sein. Er hofft, dass andere Leute seine Leistungen anerkennen.
14	ipstrgv	SE	Es ist ihm wichtig dass der Staat seine Sicherheit gegen alle Bedrohungen wahrt. Er möchte einen starken Staat, der seine Bürger beschützen kann.
15	ipadvnt	ST	Er ist auf Abenteuer aus und nimmt dafür Risiken auf sich. Er möchte ein aufregendes Leben führen.
16	ipbhprp	CO	Es ist ihm wichtig, ein anständiges Leben zu führen. Er möchte alles vermeiden, was Leute als Fehltritt bezeichnen könnten.

17	iprspot	PO	Es ist ihm wichtig, von anderen respektiert zu werden. Er möchte, dass die Leute das tun, was er sagt.
18	iplylfr	BE	Ihm ist es wichtig, gegenüber seinen Freunden treu sein. Er möchte für die Menschen da sein, die ihm nahe stehen.
19	impenv	UN	Er ist überzeugt, dass sich die Menschen um ihre Umwelt kümmern sollten. Umweltschutz ist ihm wichtig.
20	imptrad	TR	Traditionen sind ihm wichtig. Er möchte jene Sitten und Gebräuche befolgen, die ihm durch die Religion oder die Familie überliefert wurden.
21	impfun	HE	Er nutzt jede Gelegenheit, zu seinem Vergnügen zu kommen. Es ist ihm wichtig, Dinge zu tun, die ihm Vergnügen bereiten.

Tabelle 18: Labels, Wording und Wertetypen der ESS-PVQ-Items

Gütlige Antworten

gleicht mit sehr	gleicht mir	gleicht mir etwas	gleicht mir wenig	gleicht mir nicht	gleicht mir ganz und gar nicht
1	2	3	4	5	6

Fehlende Werte

Verweigerung	Weiß nicht	Keine Antwort
7	8	9

Tabelle 19: Antwortkategorien für die ESS-PVQ-Items (im Survey als Showcard den Befragten übergeben)

	IPCRTIV	IMPRICH	IPEQOPT	IPSHABT	IMPSAFE	IMPDIFF	IPFRULE
IPCRTIV	1,000						
IMPRICH	0,178	1,000					
IPEQOPT	0,207	-0,171	1,000				
IPSHABT	0,289	0,386	0,031	1,000			
IMPSAFE	-0,021	-0,011	0,159	0,132	1,000		
IMPDIFF	0,398	0,215	0,169	0,241	-0,089	1,000	
IPFRULE	-0,079	0,079	-0,001	0,093	0,279	-0,100	1,000
IPUDRST	0,219	-0,128	0,363	0,035	0,082	0,175	0,059
IPMODST	-0,169	-0,183	0,129	-0,141	0,243	-0,196	0,319
IPGDTIM	0,341	0,240	0,121	0,261	0,001	0,470	-0,104
IMPFREE	0,378	0,088	0,190	0,230	0,011	0,331	-0,103
IPHLPPL	0,158	-0,168	0,365	0,039	0,224	0,064	0,096
IPSUCES	0,348	0,418	0,002	0,584	0,056	0,294	0,069
IPSTRGV	0,001	-0,002	0,158	0,127	0,449	0,001	0,261
IPADVNT	0,305	0,374	-0,005	0,233	-0,257	0,513	-0,121
IPBHPRP	-0,128	-0,032	0,096	0,082	0,398	-0,182	0,400
IPRSPOT	0,078	0,288	-0,021	0,371	0,164	0,034	0,202
IPLYLFR	0,184	-0,054	0,284	0,137	0,174	0,161	0,002
IMPENV	0,209	-0,173	0,386	0,055	0,146	0,097	0,036
IMPTRAD	-0,044	-0,094	0,066	0,006	0,263	-0,116	0,289
IMPFUN	0,328	0,237	0,084	0,263	-0,062	0,475	-0,129

	IPUDRST	IPMODST	IPGDTIM	IMPFREE	IPHLPPL	IPSUCES	IPSTRGV
IPUDRST	1,000						
IPMODST	0,162	1,000					
IPGDTIM	0,159	-0,182	1,000				
IMPFREE	0,176	-0,150	0,412	1,000			
IPHLPPL	0,348	0,182	0,087	0,119	1,000		
IPSUCES	0,019	-0,153	0,335	0,294	0,066	1,000	
IPSTRGV	0,114	0,231	0,076	0,075	0,257	0,121	1,000
IPADVNT	0,020	-0,254	0,394	0,240	-0,071	0,331	-0,129
IPBHPRP	0,067	0,401	-0,119	-0,114	0,243	0,077	0,414
IPRSPOT	-0,002	-0,003	0,113	0,122	0,062	0,349	0,167
IPLYLFR	0,275	0,023	0,229	0,266	0,372	0,097	0,217
IMPENV	0,358	0,148	0,097	0,203	0,376	0,047	0,199
IMPTRAD	0,129	0,276	-0,050	-0,055	0,260	-0,014	0,357
IMPFUN	0,109	-0,178	0,645	0,335	0,057	0,337	0,065

	IPADVNT	IPBHPRP	IPRSPOT	IPLYLFR	IMPENV	IMPTRAD	IMPFUN
IPADVNT	1,000						
IPBHPRP	-0,248	1,000					
IPRSPOT	0,115	0,259	1,000				
IPLYLFR	-0,019	0,122	0,090	1,000			
IMPENV	-0,041	0,143	0,054	0,363	1,000		
IMPTRAD	-0,224	0,390	0,165	0,129	0,222	1,000	
IMPFUN	0,487	-0,135	0,103	0,161	0,075	-0,053	1,000

Kommunalitäten			
SD	ipcrtiv	Wichtig originelle Ideen zu haben und kreativ zu sein	0,332
PO	imprich	Wichtig reich zu sein, Geld und wertvolle Sache zu haben	0,420
UN	ipeqopt	Wichtig, dass alle gleich behandelt werden und gleiche Chancen haben	0,359
AC	ipshabt	Wichtig seine Fähigkeiten zu zeigen und bewundert zu werden	0,541
SE	impsafe	Wichtig in sicherer Umgebung zu leben	0,339
ST	impdiff	Wichtig neue und unterschiedliche Dinge im Leben zu probieren	0,429
CO	ipfrule	Wichtig das zu machen, was einem gesagt wird und sich an Regeln zu halten	0,292
UN	ipudrst	Wichtig unterschiedliche Leute zu verstehen	0,327
TR	ipmodst	Wichtig bescheiden zu sein und keine Aufmerksamkeit auf sich zu lenken	0,307
HE	ipgdtim	Wichtig Spaß zu haben	0,555
SD	impfree	Wichtig eigene Entscheidungen zu machen und frei zu sein	0,298
BE	iphlppl	Wichtig anderen zu helfen und sich um sie zu kümmern	0,364
AC	ipsuces	Wichtig erfolgreich zu sein, und dass Leute Leistungen anerkennen	0,564
SE	ipstrgv	Wichtig, dass Staat stark ist, um die Sicherheit zu gewährleisten	0,375
ST	ipadvnt	Wichtig Abenteuer zu suchen und ein aufregendes Leben zu haben	0,473
CO	ipbhprp	Wichtig sich anständig zu benehmen	0,547
PO	iprspot	Wichtig den Respekt der anderen zu bekommen	0,298
BE	iplylfr	Wichtig loyal zu Freunden sein und für sie da sein	0,326
UN	impenv	Wichtig sich um die Natur und Umwelt zu kümmern	0,399
TR	imptrad	Wichtig Traditionen und Bräuchen zu folgen	0,287
HE	impfun	Wichtig Vergnügen und jene Dinge zu suchen, die Spaß bereiten	0,593

		SD ipcrtiv	PO imprich	UN ipeqopt	AC ipshabt	SE impsafe	ST impdiff	CO ipfrule
		Wichtig originelle Ideen zu haben und kreativ zu sein	Wichtig reich zu sein, Geld und wertvolle Sache zu haben	Wichtig, dass alle gleich behandelt werden und gleiche Chancen haben	Wichtig seine Fähigkeiten zu zeigen und bewundert zu werden	Wichtig in sicherer Umgebung zu leben	Wichtig neue und unterschiedliche Dinge im Leben zu probieren	Wichtig das zu machen, was einem gesagt wird und sich an Regeln zu halten
N	Gültig	2221	2226	2222	2221	2225	2220	2191
	Fehlend	35	30	34	35	31	36	65
Mittelwert		2,421	3,765	2,082	2,877	2,379	2,976	3,593
Schiefe		0,755	-0,133	0,979	0,470	0,784	0,292	-0,103
Kurtosis		0,217	-0,744	0,863	-0,455	0,135	-0,713	-0,817
Standardfehler der Kurtosis		0,104	0,104	0,104	0,104	0,104	0,104	0,105

		UN ipudrst	TR ipmodst	HE ipgdtim	SD impfree	BE iphlppl	AC ipsuces	SE ipstrgv
		Wichtig unterschiedliche Leute zu verstehen	Wichtig bescheiden zu sein und keine Aufmerksamkeit auf sich zu lenken	Wichtig Spaß zu haben	Wichtig eigene Entscheidungen zu machen und frei zu sein	Wichtig anderen zu helfen und sich um sie zu kümmern	Wichtig erfolgreich zu sein, und dass Leute Leistungen anerkennen	Wichtig, dass Staat stark ist, um die Sicherheit zu gewährleisten
N	Gültig	2208	2204	2232	2228	2218	2217	2176
	Fehlend	48	52	24	28	38	39	80
Mittelwert		2,394	3,191	2,504	2,025	2,289	2,909	2,555
Schiefe		0,784	0,170	0,684	1,042	0,773	0,429	0,691
Kurtosis		0,503	-0,712	-0,177	0,915	0,927	-0,275	0,062
Standardfehler der Kurtosis		0,104	0,104	0,104	0,104	0,104	0,104	0,105

		ST ipadvnt	CO ipbhprp	PO iprspot	BE iplylfr	UN impenv	TR imptrad	HE impfun
		Wichtig Abenteuer zu suchen und ein aufregendes Leben zu haben	Wichtig sich anständig zu benehmen	Wichtig den Respekt der anderen zu bekommen	Wichtig loyal zu Freunden sein und für sie da sein	Wichtig sich um die Natur und Umwelt zu kümmern	Wichtig Traditionen und Bräuchen zu folgen	Wichtig Vergnügen und jene Dinge zu suchen, die Spaß bereiten
N	Gültig	2218	2196	2194	2210	2215	2221	2223
	Fehlend	38	60	62	46	41	35	33
Mittelwert		3,734	2,831	3,012	1,770	2,223	2,874	2,798
Schiefe		-0,180	0,495	0,324	1,489	0,884	0,487	0,496
Kurtosis		-0,791	-0,361	-0,429	3,047	0,629	-0,571	-0,497
Standardfehler der Kurtosis		0,104	0,104	0,104	0,104	0,104	0,104	0,104

Mplus Faktorladungen EFA			Faktor			
			change	cons	self-enh	self-trans
SD	ipcrtiv	Wichtig originelle Ideen zu haben und kreativ zu sein	0,205	-0,285	0,209	-0,342
PO	imprich	Wichtig reich zu sein, Geld und wertvolle Sache zu haben	0,164	-0,509	0,004	0,303
UN	ipeqopt	Wichtig, dass alle gleich behandelt werden und gleiche Chancen haben	0,032	0,037	0,004	-0,594
AC	ipshabt	Wichtig seine Fähigkeiten zu zeigen und bewundert zu werden	-0,018	-0,732	-0,032	-0,068
SE	impsafe	Wichtig in sicherer Umgebung zu leben	-0,04	-0,095	-0,527	-0,136
ST	impdiff	Wichtig neue und unterschiedliche Dinge im Leben zu probieren	0,505	-0,111	0,171	-0,163
CO	ipfrule	Wichtig das zu machen, was einem gesagt wird und sich an Regeln zu halten	-0,091	-0,131	-0,499	0,067
UN	ipudrst	Wichtig unterschiedliche Leute zu verstehen	0,075	0,053	-0,027	-0,552
TR	ipmodst	Wichtig bescheiden zu sein und keine Aufmerksamkeit auf sich zu lenken	-0,112	0,169	-0,475	-0,081
HE	ipgdtim	Wichtig Spaß zu haben	0,727	-0,049	-0,056	-0,072
SD	impfree	Wichtig eigene Entscheidungen zu machen und frei zu sein	0,292	-0,185	0,132	-0,303
BE	iphlppl	Wichtig anderen zu helfen und sich um sie zu kümmern	0,005	-0,006	-0,213	-0,525
AC	ipsuces	Wichtig erfolgreich zu sein, und dass Leute Leistungen anerkennen	0,104	-0,714	-0,014	-0,024
SE	ipstrgv	Wichtig, dass Staat stark ist, um die Sicherheit zu gewährleisten	0,117	-0,058	-0,577	-0,115
ST	ipadvnt	Wichtig Abenteuer zu suchen und ein aufregendes Leben zu haben	0,501	-0,205	0,24	0,114
CO	ipbhprp	Wichtig sich anständig zu benehmen	-0,1	-0,089	-0,7	-0,023
PO	iprspot	Wichtig den Respekt der anderen zu bekommen	-0,009	-0,467	-0,256	0,056
BE	iplylfr	Wichtig loyal zu Freunden sein und für sie da sein	0,136	-0,046	-0,102	-0,502
UN	impenv	Wichtig sich um die Natur und Umwelt zu kümmern	-0,009	-0,001	-0,083	-0,611
TR	imptrad	Wichtig Traditionen und Bräuchen zu folgen	-0,011	0,026	-0,505	-0,103
HE	impfun	Wichtig Vergnügen und jene Dinge zu suchen, die Spaß bereiten	0,803	-0,011	-0,077	0,043

Tabelle 23: Ladungsmatrix EFA (Rotation: geomin; Mplus)

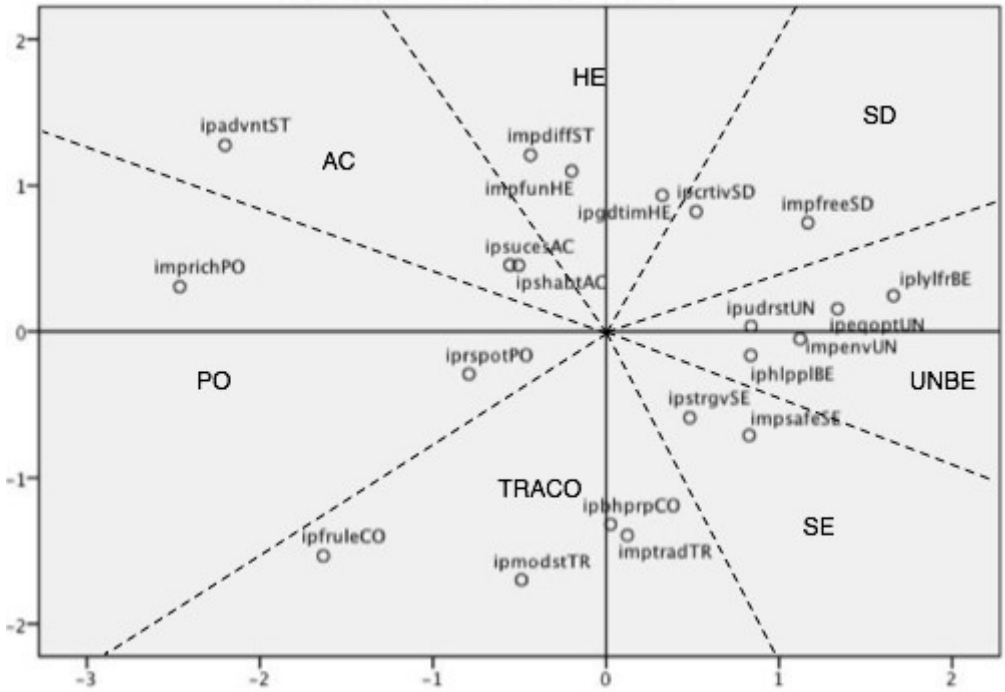


Abbildung 7: MDS mit unzentrierten Daten

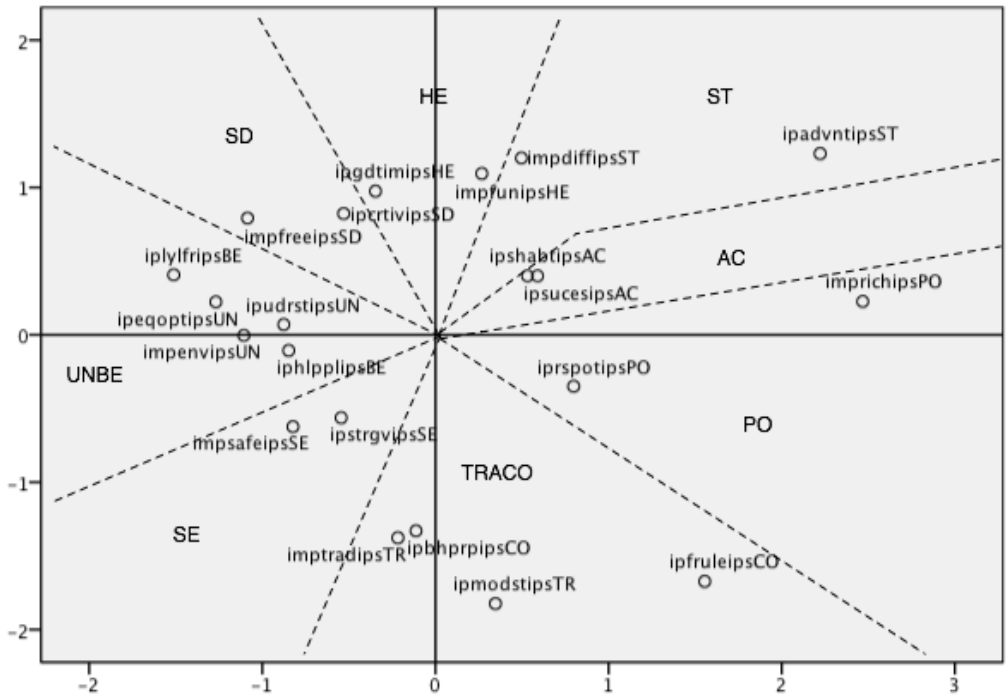


Abbildung 8: MDS mit zentrierten Daten

Wertetyp	*Item*	*Itembeschreibung*	*Häufigkeit missing*	*Item*	*Itembeschreibung*	*Häufigkeit missing*	*Item*	*Itembeschreibung*	*Häufigkeit missing*	*Mittelwert der Wertetypen (missings)*	*Mittelwert aller 21 Items (missings)*
SD	ipcrtiv	Wichtig originelle Ideen zu haben und kreativ zu sein	35	impfree	Wichtig eigene Entscheidungen zu machen und frei zu sein	28				31,5	**42,4**
PO	imprich	Wichtig reich zu sein, Geld und wertvolle Sache zu haben	30	iprspot	Wichtig den Respekt der anderen zu bekommen	**62**				**46**	
UN	ipeqopt	Wichtig, dass alle gleich behandelt werden und gleiche Chancen haben	34	ipudrst	Wichtig unterschiedliche Leute zu verstehen	**48**	impenv	Wichtig sich um die Natur und Umwelt zu kümmern	41	41	
AC	ipshabt	Wichtig seine Fähigkeiten zu zeigen und bewundert zu werden	35	ipsuces	Wichtig erfolgreich zu sein, und dass Leute Leistungen anerkennen	39				37	
SE	impsafe	Wichtig in sicherer Umgebung zu leben	31	ipstrgv	Wichtig, dass Staat stark ist, um die Sicherheit zu gewährleisten	**80**				**55,5**	
TR	imptrad	Wichtig bescheiden zu sein und keine Aufmerksamkeit auf sich zu lenken	**52**	ipmodst	Wichtig Traditionen und Brächen zu folgen	35				**43,5**	
HE	ipgdtim	Wichtig Spaß zu haben	24	impfun	Wichtig Vergnügen und jene Dinge zu suchen, die Spaß bereiten	33				28,5	
BE	iphlppl	Wichtig anderen zu helfen und sich um sie zu kümmern	38	iplylfr	Wichtig loyal zu Freunden sein und für sie da sein	**46**				42	
ST	impdiff	Wichtig neue und unterschiedliche Dinge im Leben zu probieren	36	ipadvnt	Wichtig Abenteuer zu suchen und ein aufregendes Leben zu haben	38				37	

Tabelle 24: Häufigkeiten und Mittelwerte fehlender Werte der 21 PVQ-Items (Fett und größer dargestellte Werte liegen über dem Durchschnitt der 21 Items)

			2004 (n = 2.256)				
		Itemformulierung	*openness to change*	*conservation*	*self-enhancement*	*self-transcendence*	Communality
SD	ipcrtiv	Wichtig originelle Ideen zu haben und kreativ zu sein	,179	-,214	,261	-,355	,332
PO	imprich	Wichtig reich zu sein, Geld und wertvolle Sache zu haben	,191		**,496**	,306	,420
UN	ipeqopt	Wichtig, dass alle gleich behandelt werden und gleiche Chancen haben				**-,607**	,359
AC	ipshabt	Wichtig seine Fähigkeiten zu zeigen und bewundert zu werden			**,758**		,541
SE	impsafe	Wichtig in sicherer Umgebung zu leben		**,525**		-,122	,339
ST	impdiff	Wichtig neue und unterschiedliche Dinge im Leben zu probieren	**,539**	-,141		-,155	,429
CO	ipfrule	Wichtig das zu machen, was einem gesagt wird und sich an Regeln zu halten		**,521**	,118	,095	,292
UN	ipudrst	Wichtig unterschiedliche Leute zu verstehen				**-,555**	,327
TR	ipmodst	Wichtig bescheiden zu sein und keine Aufmerksamkeit auf sich zu lenken		**,488**	-,179		,307
HE	ipgdtim	Wichtig Spaß zu haben	**,748**				,555
SD	impfree	Wichtig eigene Entscheidungen zu machen und frei zu sein	,259	-,140	,169	-,318	,298
BE	iphlppl	Wichtig anderen zu helfen und sich um sie zu kümmern		,203		**-,530**	,364
AC	ipsuces	Wichtig erfolgreich zu sein, und dass Leute Leistungen anerkennen			,711		,564
SE	ipstrgv	Wichtig, dass Staat stark ist, um die Sicherheit zu gewährleisten	,156	**,596**			,375
ST	ipadvnt	Wichtig Abenteuer zu suchen und ein aufregendes Leben zu haben	**,539**	-,209	,147	,122	,473
CO	ipbhprp	Wichtig sich anständig zu benehmen		**,715**			,547
PO	iprspot	Wichtig den Respekt der anderen zu bekommen		,241	**,481**		,298
BE	iplylfr	Wichtig loyal zu Freunden sein und für sie da sein				**-,516**	,326
UN	impenv	Wichtig sich um die Natur und Umwelt zu kümmern				**-,626**	,399
TR	imptrad	Wichtig Traditionen und Bräuchen zu folgen		**,516**			,287
HE	impfun	Wichtig Vergnügen und jene Dinge zu suchen, die Spaß bereiten	**,837**	,112			,593

Vollständig erklärte Varianz = 40,12%; KMO = 0,840

Kreuzladungen			M.I.	E.P.C.	std E.P.C.	stdYX E.P.C.				M.I.	E.P.C.	Std E.P.C.	stdYX E.P.C.
traco	by	imprich	46,975	-0,327	-0,230	-0,171	st	by	iphlppl	20,760	-0,122	-0,108	-0,114
traco	by	ipeqopt	16,127	-0,157	-0,111	-0,108	st	by	ipsuces	17,930	0,173	0,154	0,128
traco	by	impdiff	20,824	0,273	0,192	0,145	st	by	ipstrgv	21,724	0,187	0,166	0,143
traco	by	iphlppl	55,147	0,268	0,189	0,199	st	by	iprspot	58,812	-0,327	-0,291	-0,237
traco	by	ipadvnt	20,822	-0,317	-0,223	-0,159	st	by	iplylfr	13,997	0,093	0,083	0,094
traco	by	iprspot	101,049	0,449	0,316	0,258	st	by	impfun	18,596	0,462	0,410	0,314
unbe	by	imprich	136,177	-0,658	-0,397	-0,295	he	by	impsafe	16,682	-0,144	-0,135	-0,116
unbe	by	ipshabt	14,360	0,189	0,114	0,091	he	by	impdiff	46,573	0,622	0,583	0,440
unbe	by	impdiff	115,307	0,646	0,389	0,294	he	by	ipmodst	31,672	-0,206	-0,193	-0,146
unbe	by	ipfrule	17,855	-0,264	-0,159	-0,117	he	by	iphlppl	13,081	-0,095	-0,089	-0,094
unbe	by	ipgdtim	15,296	0,208	0,125	0,104	he	by	ipsuces	15,927	0,153	0,144	0,120
unbe	by	ipsuces	10,667	0,159	0,096	0,080	he	by	ipstrgv	16,689	0,147	0,138	0,118
unbe	by	ipadvnt	115,302	-0,749	-0,451	-0,321	he	by	ipadvnt	46,579	-0,721	-0,676	-0,481
unbe	by	imptrad	21,978	0,295	0,178	0,131	he	by	iprspot	22,871	-0,188	-0,176	-0,144
unbe	by	impfun	15,290	-0,226	-0,136	-0,104	he	by	iplylfr	33,187	0,141	0,132	0,150
sd	by	imprich	17,600	-0,260	-0,180	-0,134	poac	by	ipfrule	20,183	0,212	0,155	0,114
sd	by	impdiff	103,842	1,028	0,711	0,537	poac	by	ipmodst	103,162	-0,466	-0,341	-0,259
sd	by	ipmodst	30,729	-0,284	-0,196	-0,149	poac	by	ipbhprp	22,755	0,225	0,165	0,127
sd	by	ipgdtim	29,492	0,582	0,403	0,334	poac	by	iplylfr	21,802	0,140	0,103	0,117
sd	by	iphlppl	24,898	-0,209	-0,145	-0,153	se	by	imprich	69,115	-0,387	-0,296	-0,221
sd	by	ipsuces	32,321	0,311	0,215	0,179	se	by	ipeqopt	12,491	-0,144	-0,110	-0,107
sd	by	ipadvnt	103,832	-1,192	-0,824	-0,587	se	by	ipshabt	11,574	0,140	0,107	0,085
sd	by	iprspot	35,602	-0,346	-0,239	-0,195	se	by	impdiff	47,519	0,344	0,264	0,199
sd	by	iplylfr	26,772	0,202	0,140	0,159	se	by	ipudrst	18,411	-0,180	-0,138	-0,131
sd	by	impfun	29,487	-0,634	-0,438	-0,336	se	by	ipmodst	31,430	-0,569	-0,435	-0,330
st	by	imprich	32,705	0,262	0,232	0,173	se	by	iphlppl	37,999	0,232	0,177	0,187
st	by	ipshabt	15,717	-0,165	-0,146	-0,117	se	by	ipadvnt	47,511	-0,399	-0,305	-0,217
st	by	impsafe	21,715	-0,183	-0,163	-0,141	se	by	iprspot	65,511	0,353	0,270	0,221
st	by	ipmodst	38,952	-0,276	-0,245	-0,186	se	by	imptrad	15,437	0,412	0,315	0,232
st	by	ipgdtim	18,610	-0,425	-0,377	-0,313							

Tabelle 26: Modifikationen für Kreuzladungen (7-Faktorenmodell, Mplus)

Fehlerkovarianzen			M.I.	E.P.C.	Std E.P.C.	stdYX E.P.C.
impsafe	with	ipshabt	16,207	0,098	0,098	0,129
impdiff	with	ipcrtiv	26,269	0,145	0,145	0,161
ipfrule	with	imprich	19,212	0,152	0,152	0,118
ipmodst	with	ipudrst	23,095	0,127	0,127	0,131
impfree	with	imprich	11,785	-0,088	-0,088	-0,095
impfree	with	ipgdtim	46,498	0,143	0,143	0,231
iphlppl	with	imprich	20,299	-0,106	-0,106	-0,124
iphlppl	with	impfree	12,028	-0,063	-0,063	-0,101
ipsuces	with	impsafe	11,433	-0,078	-0,078	-0,118
ipsuces	with	iphlppl	13,076	0,067	0,067	0,116
ipstrgv	with	impdiff	13,588	0,102	0,102	0,121
ipadvnt	with	imprich	76,885	0,283	0,283	0,263
ipadvnt	with	ipshabt	16,268	-0,113	-0,113	-0,135
ipadvnt	with	impsafe	33,305	-0,168	-0,168	-0,203
ipadvnt	with	ipfrule	14,414	0,126	0,126	0,114
ipadvnt	with	ipgdtim	39,756	-0,187	-0,187	-0,259
ipadvnt	with	impfree	12,284	-0,092	-0,092	-0,118
iprspot	with	ipcrtiv	16,093	-0,109	-0,109	-0,110
iprspot	with	ipshabt	10,767	0,102	0,102	0,107
iprspot	with	impdiff	20,465	-0,134	-0,134	-0,126
iprspot	with	ipbhprp	21,669	0,130	0,130	0,141
iplylfr	with	ipcrtiv	11,630	-0,065	-0,065	-0,099
iplylfr	with	ipmodst	20,273	-0,099	-0,099	-0,123
iplylfr	with	ipgdtim	22,472	0,081	0,081	0,149
iplylfr	with	impfree	13,794	0,063	0,063	0,106
iplylfr	with	ipadvnt	16,401	-0,085	-0,085	-0,123
impenv	with	imprich	30,707	-0,141	-0,141	-0,154
imptrad	with	iphlppl	10,331	0,079	0,079	0,088
imptrad	with	ipstrgv	13,994	0,114	0,114	0,113
imptrad	with	iprspot	15,765	0,132	0,132	0,104
imptrad	with	impenv	18,756	0,114	0,114	0,120
impfun	with	ipadvnt	37,830	0,199	0,199	0,255

Tabelle 27: Modifikationen für Fehlerkovarianzen (7-Faktorenmodell, Mplus)

Abstract – Deutsche Fassung

Werte sind für die Soziologie seit der Zeit Max Webers im 19. Jahrhundert bedeutsam. Sie sind Konzepte, die sowohl eng mit der Individualität von Menschen verknüpft als auch durch die Gesellschaft mitgeprägt sind. Das macht sie einerseits brauchbar für die Vorhersage von menschlichem Sozialverhalten und andererseits durch ihre mangelnde Abgrenzbarkeit schwer erfassbar. Mehrere Forscher haben bereits ihre Messung thematisiert, die bis dato durch wechselseitige Erfolge und Rückschläge gekennzeichnet ist. Shalom Schwartz entwickelte eine Theorie für eine Skala, die die universelle Struktur von Übereinstimmung und Gegensätzen von Werten erfasst. Die Validität und Reliabilität der empirischen Messung durch unterschiedliche Versionen der Skala unterliegen einer anhaltenden Debatte. Die vorliegende Arbeit füllt eine Lücke in dieser wissenschaftlichen Diskussion: Einerseits verknüpft sie bekannte methodologische Probleme, die bei der Messung von Werten auftreten können, mit empirischen Ergebnissen. Andererseits rückt sie die vernachlässigten Dimensionen kognitiver und inhaltlicher Aspekte der Messung in den Vordergrund, indem kognitive Protokolle mangelhafte Messeigenschaften eines Datensatzes des European Social Surveys erklären. Unterstützt werden diese qualitativen Befunde durch quantitative Analysen mit deskriptiven, explo-ratorischen und konfirmatorischen Methoden. Diese identifizieren problembehaftete Indikatoren und liefern gleichzeitig zusätzliche Hinweise, die die Ergebnisse der qualitativen Studie bestätigen. Dieses Vorgehen verwertet die Vorzüge sowohl der quantitativen als auch der qualitativen Methoden ohne ihre Nachteile in Kauf nehmen zu müssen. Ergebnisse, die sich methodisch überschneiden, deuten auf die Brauchbarkeit des Konzepts hin. Sie zeigen auch, dass Probleme, die als ausgeräumt galten, und bekannte methodologische Bedenken in der Messung weiterhin bestehen und eine Anpassung derselben nötig scheint. Das Design der vorliegenden Arbeit liefert rein beschreibende Anhaltspunkte, wo die Entwicklung der Skala ansetzen könnte, und nicht, wie eine Änderung aussehen könnte.

Abstract – English Version

Human Values are an area of interest for Sociology since Max Weber's time in the 19th century. They are concepts partly intervoven with the individuality of human beings and partly imprinted by society. That specific feature makes them on the one side instrumental to predicting human social behaviour, but on the other side problematic to capture due to their fuzzy delineation. Until now several researchers have conceptualized their measurement, which has been alternating between accomplishments and setbacks. Shalom Schwartz developed a theory for a scale that covers the universal structure of conflicts and congruencies between human values. The validity and reliability of different versions of the scale are currently debated. The study at hand fills in a gap in this academic discussion: on the one hand it connects existing methodological concerns that may arise during the process of measurement of values with empirical results. On the other hand it highlights the neglected aspects of cognitive processes and content of the measurement by explaining the poor performance of measurement quality of a given dataset from the European Social Survey through cognitive protocols. Qualitative findings are supported by quantitative analysis comprising descriptive, exploratory and confirmatory methods. These identify problematic indicators and simultaneously offer confirming information for the qualitative results. This proceeding utilizes the merits of both quantitative and qualitative approaches whithout hazarding their drawbacks. Results that overlap methodologically indicate the utility of this approach. They also show that problems, which had been considered to be eliminated from the measurement, and methodological concerns still persist. An adaptation of the scale seems to be required. The study at hand can give descriptive but not prescriptive advice for the further development.

Curriculum Vitae

Name:	Georg Kessler
Geboren:	Wien, 10. Oktober 1982
Schulausbildung:	1993 – 2001 Bundesrealgymnasium in 1090 Wien
Hochschulausbildung:	2004 – 2012 Soziologie, Universität Wien 2006 – 2007 Medizin, Medizinische Universität Wien SS 2009 Studienaufenthalt in Rovaniemi (Finnland), University of Lapland
Fachliche Schwerpunkte:	Spezielle Soziologien: Alterssoziologie und Migrationssoziologie selbst gelegter Fokus: quantitative Methoden mit Schwerpunkt auf Strukturgleichungsmodellen
Sprachkenntnisse:	Deutsch (Muttersprache), Englisch (sehr gut), Spanisch (Fortgeschritten), Französisch (Anfänger), Serbisch (Anfänger)

Wissenschaftlicher Werdegang:

Sommer 2009	4-monatiges Praktikum an der Universtät Lappland (Finnland) bei Prof. Anu Valtonen
SS 2010	30-stündiges Tutorium an der Universität Wien mit Prof. Peter Schmidt (Uni Gießen) für "Strutkturgleichungsmodelle mit Amos18".
Sommer 2010	Teaching Assistant auf der ECPR Summerschool Ljubljana für den 2-wöchigen Kurs "Confirmatory Factor Analysis and Structural Equation Modelling" mit Prof. Peter Schmidt und Bart Meulemann, MA (Uni Gent)
WS 2010	30-stündiges Tutorium an der Universität Wien mit Prof. Jost Reinecke (Uni Bamberg) für "Strutkturgleichungsmodelle in Längs- und Querschnittsdaten".
SS 2011	30-stündiges Tutorium an der Universität Wien mit Prof. Peter Schmidt (Uni Gießen) für "Entwicklung von Messinstrumenten".
WS 2012	30-stündiges Tutorium an der Universität Wien mit Prof. Jost Reinecke (Uni Bamberg) für "Längsschnitt- und Mischverteilungsmodelle".

e-Mail-Kontak: kessler.georg@gmail.com