universität wien

Diplomarbeit

Fehlerquellen in der international vergleichenden Umfrageforschung- Eine Überprüfung der Äquivalenz zweier Skalen des European Social Surveys

Verfasser

Martin Vogler

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften

(Mag.rer.soc.oec.)

Wien, im Jänner 2010

Studienkennzahl lt. Studienblatt:	A 121
Studienrichtung lt. Studienblatt:	Soziologie: Rechts-, Sozial- und Wirtschaftswissenschaftliche Studienrichtung
Diplomarbeitsbetreuer:	Ao. Univ.-Prof. Dr. Franz Kolland

INHALTSVERZEICHNIS

TABELLENVERZEICHNIS

ABBILDUNGSVERZEICHNIS

Abstract

In der vorliegenden Diplomarbeit wird der European Social Survey auf seine Güte für einen internationalen Vergleich getestet. Es wird überprüft, ob und in welchem Ausmaß eine Vergleichbarkeit der Daten, repräsentiert durch das Konzept der Äquivalenz, zwischen Deutschland, Großbritannien und Österreich möglich ist. Deutschland wurde hierbei in zwei separate Analyseeinheiten aufgeteilt, um einerseits möglichst homogene Untersuchungseinheiten für die Analyse im Sinne eines *Most Different Systems Design* zu verwenden. Die Überprüfung der erreichten Ebene der Äquivalenz des ESS wird exemplarisch anhand von zwei Skalen, die die Einstellung der Befragten zu Migration und die Meinung zu den Auswirkungen dieser für das Einwanderungsland messen durchgeführt. Es wird getestet, ob nationale Unterschiede und Gemeinsamkeiten bezüglich der betreffenden Einstellungen und Meinungen tatsächlich existieren, oder lediglich unter anderem auf Grund der Existenz von Verzerrungen im Ländervergleich zustande kommen.

Für die Beantwortung dieser Frage werden die Stichprobendesigns verglichen, die semantischen und pragmatischen Bedeutungen der Skalenitems, die Verteilungen in den Antworten, die interne und externe Konsistenz der Skalen, die Dimensionalität, die Faktorenstruktur und die Identität der Skalenwerte der zwei Skalen in den einzelnen Ländern untersucht. Für diesen Zweck kommen mehrere statistische Verfahren zur Anwendung, da diese unterschiedliche Vor- und Nachteile bezüglich der Entdeckung von Verzerrungen auf Ebene der Konstrukte und der einzelnen Items im Rahmen eines *post hoc* Ansatzes aufweisen.

Es wurde letztlich festgestellt, dass in allen Analyseeinheiten die Skalen eine vergleichbare Struktur aufweisen. Dies bedeutet, dass die beiden Skalen in den Analyseeinheiten funktional äquivalent sind, also die von ihnen gemessenen Konstrukte eine vergleichbare Beschaffenheit aufweisen. So können statistische Verfahren, die auf Korrelationen und Kovarianzen basieren, für die einzelnen Items angewandt und deren Ergebnisse zwischen den Ländern interpretiert werden. Da allerdings die Messeinheiten der Items nicht äquivalent sind, können keine Aussagen über länderspezifische

Unterschiede auf der Ebene der latenten Konstrukte, deren Zusammenhänge mit anderen Konstrukten sowie Vergleiche von Absolutwerten gemacht werden.

Die hierfür notwendige Existenz identischer Skaleneinheiten und somit eine Messinvarianz aller Items konnte lediglich zwischen Großbritannien und Westdeutschland festgestellt werden. In Großbritannien, Westdeutschland und Österreich weisen fünf der sechs Items identische Skaleneinheiten auf. So wäre es hierbei möglich, die Einstellung der Befragten zur Migration zwischen diesen Untersuchungseinheiten zu vergleichen. Selbiges wurde zwischen Großbritannien und Ost- und Westdeutschland festgestellt. Es können allerdings nur die Meinungen der Befragten zu den Auswirkungen der Migration auf das Einwanderungsland niveauorientiert verglichen werden.

Das Ziel einer jeden international vergleichenden Umfragestudie sollte die Erreichung einer möglichst hohen Ebene der Äquivalenz sein. Diese Diplomarbeit hat jedoch gezeigt, dass dies kein leichtes Vorhaben darstellt, da eine Vielzahl an möglichen Fehlerquellen existiert, die die Ergebnisse eines internationalen Vergleichs verzerren können. So ist es leichter Verzerrungen zu entdecken als deren Ursachen zu bestimmen. Es wurde zudem deutlich, dass die Anwendung von statistischen Methoden nur einen Teilbeitrag zur Untersuchung der erreichten Ebene der Äquivalenz im Rahmen einer Sekundäranalyse leisten kann. Um Verzerrungen von Vergleichen wirklich zu minimieren, ist zudem auch ein adäquates Studiendesign und die Anwendung diverser *a priori* Techniken, wie zum Beispiel Pretests und Methodenexperimente, vonnöten. Nur so kann letztlich eine Minimierung der Verzerrungen beziehungsweise eine Maximierung der Äquivalenz der Daten sichergestellt werden.

1. Einleitung

In den letzten 30 Jahren ist die international vergleichende Forschung zu einem wichtigen Bestandteil diverser Forschungsbereiche geworden.[1] Dies wird sichtbar an der stetig zunehmenden Anzahl an international vergleichenden Studien beziehungsweise Publikationen über die meisten Wissenschaftsdisziplinen hinweg. In diesen international vergleichenden Studien werden ganze Nationalstaaten beziehungsweise deren Einwohner bezüglich verschiedenster Dimensionen verglichen, um Ähnlichkeiten und Unterschiede zwischen diesen zu identifizieren und zu erklären. Diese Entwicklung kann als eine Auswirkung der zunehmenden Globalisierung betrachtet werden, die durch gestiegene Interpendenzen dazu führt, dass Individuen nicht mehr ausschließlich von den Auswirkungen von nationalen Politiken oder Phänomenen betroffen sind.

Die internationalen Studien die in dieser Arbeit behandelt werden, sind solche, die den sozialwissenschaftlichen Wissenschaftsdisziplinen zuzuordnen sind und ihre Daten mittels der Methode der Befragung erheben. Im Fokus dieser Arbeit stehen die „großen" internationalen Umfragestudien, die periodisch durchgeführt werden. Das Themengebiet, das mit den international vergleichenden Umfragen behandelt wird, ist dabei sehr breit gestreut. Es existieren Studien die ihren Fokus auf Zeitverwendung, ökonomischen und politischen Indikatoren, Lebensbedingungen, Lebensqualität, schulische Leistungen und Religiosität haben, um nur einige Themen zu nennen. Bei den meisten dieser Umfragestudien werden die Datensätze einer breiten Öffentlichkeit zugänglich gemacht. Diese erhobenen Daten aus verschiedenen Ländern können beziehungsweise werden so auch von Personen verwendet, die nicht an der Entstehung der betreffenden Studie mitgewirkt haben. Sie werden für wissenschaftliche Zwecke im Rahmen von Sekundäranalysen genützt. Diese zunehmende Nutzung für letztgenannten Zweck, insbesondere auch in einschlägigen Lehrveranstaltungen im universitären Lehrbetrieb, gab den Anreiz für die Thematik der vorliegenden Diplomarbeit.

[1] Auf geschlechtsspezifische Bezeichnungen und Formulierungen wurde in der Arbeit auf Grund der besseren Lesbarkeit verzichtet. Die in dieser Arbeit verwendeten personenbezogenen Ausdrücke umfassen also beide Geschlechter.

1.1. Erkenntnisinteresse der vorliegenden Arbeit

Es wird der European Social Survey (ESS), eine periodisch durchgeführte international vergleichende Umfrage, auf seine Güte für den internationalen Vergleich getestet werden. Es wird exemplarisch anhand von zwei Skalen, die die Einstellung der Befragten zu Migration und die Meinung zu den Auswirkungen dieser für das Einwanderungsland messen, geprüft ob Vergleiche der mit diesen Skalen gemessenen Einstellungen zwischen der deutschen, der englischen und der österreichischen Population möglich sind. Es wird also getestet, ob Unterschiede und Gemeinsamkeiten bezüglich der betreffenden Einstellungen und Meinungen tatsächlich existieren, oder lediglich unter anderem auf Grund einer mangelnden Vergleichbarkeit im Ländervergleich zustande kommen. Wenn letzteres zutrifft und systematische Verzerrungen existieren, sind Vergleiche der mit diesem Instrument erhaltenen Ergebnisse von äußerst fragwürdiger Validität. Die zentrale Frage in dieser Arbeit ist gewissermaßen die, ob soziale Phänomene die in verschiedenen sozialen Systemen beobachtet werden, überhaupt verglichen werden können und vor allem wie es geschafft werden kann, diese durch die Vermeidung von Messfehlern vergleichbar zu machen. Es soll also festgestellt werden ob Verzerrungen existent sind beziehungsweise in welchem Ausmaß Vergleiche der mit den beiden Skalen erhaltenen Ergebnisse über verschiedene Länder hinweg möglich sind.

In den Sozialwissenschaften enthalten gemessene Merkmale zumeist sowohl einen „wahren Wert“ wie auch Messfehler. Um trotz dieser Unzulänglichkeiten dennoch wissenschaftliche Erkenntnisse gewinnen zu können, existiert die klassische Testtheorie.

> „Das Grundmodell der klassischen Testtheorie basiert auf der Annahme, dass ein realisierter Messwert aus der Summe eine ‚wahren Wertes’ und einem Messfehler besteht; Messfehler sind also Differenzen zwischen ‚wahren Werten’ und beobachteten Werten. Ein ‚wahrer Wert’ kann als der Mittelwert einer großen Zahl *unabhängiger* Messungen desselben Objektes aufgefasst werden.“(Schnell/Hill/Esser 2008:150)

Von Verzerrungen wird hingegen gesprochen, wenn Störfaktoren auftreten, die die Vergleichbarkeit der Messung über verschiedene Gruppen beeinträchtigen. (vgl.

Harkness/Vijver/Mohler 2003a)[2] Wahrend die Ursachen für die Existenz von Messfehlern in einem bestimmten Studiendesign begründet sind, sind Verzerrungen keine inhärenten Charakteristiken eines Erhebungsinstruments, sondern entstehen erst in dessen Anwendung. (vgl. Van de Vijver 2003a)[3]

Es bestehen unterschiedliche methodische Anforderungen an nationale und internationale Umfragestudien. In der international vergleichenden Forschung hängt die Qualität der Rückschlüsse, die auf Grund der Messungen gezogen werden, von der Qualität einer jeder nationalen Teilstudie ab. Das Ziel von international vergleichenden Umfragen stellt oftmals die Erfassung von Auswirkungen von internationalen Unterschieden und Ähnlichkeiten dar, die durch simultane Variationen von Sozialstrukturen, Rechtssystemen, Sprachen, Politiken, Ökonomien und Kulturen entstehen können. Doch es sind gerade diese Variationen dafür verantwortlich, dass sich eine Einstellungsmessung in einer internationalen Perspektive weitaus schwieriger gestaltet, als die in einem nationalen Umfeld durchgeführte. Werden Unterschiede zwischen einzelnen Ländern gefunden, stellt sich daher stets die Frage, ob es sich hierbei um Messartefakte oder um tatsächlich existierende Unterschiede handelt. Die Existenz von Fehlern beziehungsweise Verzerrungen hat einen Einfluss auf die Vergleichbarkeit von Messwerten über verschiedene Gruppen hinweg. Erst nach sorgfältiger Prüfung der einzelnen Fehler- beziehungsweise Verzerrungsquellen, kann ein real existierender Unterschied angenommen werden. So hat bereits 1932 Rensis Likert seine Bedenken geäußert ob Skalen, wie die von ihm kostruierte, für einen internationalen Vergleich zur Anwendung kommen können.

> „It is certainly reasonable to suppose that just as an intelligence test which has been standardized upon one cultural group is not applicable to another so an attitude scale which has been constructed for one cultural group will hardly be applicable in its existing form to other cultural groups." (Likert 1932:52)

Ein Ziel dieser Arbeit ist es zu zeigen, dass eine Vergleichbarkeit von Messprozeduren und Ergebnissen über Ländergrenzen hinweg nicht ohne weiteres angenommen werden darf beziehungsweise deren Existenz im Rahmen der Studiendurchführung nicht von vornherein sichergestellt ist. Vielmehr muss der hergestellte Typ der Äquivalenz für einen bestimmten Vergleich nachträglich überprüft werden. Dies gilt insbesondere für

[2] In der deutschsprachigen Literatur zu dieser Thematik wird anstatt des Begriffs der Verzerrung auch oftmals der entsprechende englische Begriff Bias verwendet.

[3] Eine detaillierte Begriffsbestimmung und Abgrenzung dieser Begrifflichkeiten wird in Kapitel 3 vorgenommen.

eine komplexe Wissenschaft, wie sie die Sozialwissenschaften darstellen und hierbei besonders auf der Ebene der international vergleichenden Forschung mit ihren großen heterogenen Untersuchungseinheiten. Es können ein Reihe von Anstrengungen unternommen werden, um die Existenz von Verzerrungen im Rahmen einer international vergleichenden Umfrage zu minimieren, völlig ausschließen beziehungsweise eliminieren kann man diese nicht. Dieses Zitat verdeutlicht die Notwendigkeit der Qualitätsüberprüfung jeglicher sozialwissenschaftlicher Forschung.

> „Good science – whether natural science or social science – should never turn a blind eye to its known imperfections. Nor should these imperfections be concealed from potential users. Some might argue that the social sciences are always an order of magnitude error-prone than are the natural sciences. That is disputable, but in any case it provides all the more reason for greater rather than less vigilance in social science methodology." (Jowell/Kaase/Fitzgerald/Gillian 2007: 4)

1.2. Aufbau der Arbeit

Der Aufbau dieser Arbeit lässt sich grob in vier Abschnitte unterteilen. Zu Beginn werden die methodologischen Grundlagen der international vergleichenden Umfrageforschung erörtert. Dies beinhaltet eine Definition was unter international vergleichender Forschung verstanden wird und welche konkurrierenden teilweise synonym für diese verwendeten Begrifflichkeiten für diese Art der Forschung existieren. Anschließend daran wird die Methode des Vergleichs, die historische Entwicklung dieser Methode und deren Stellenwert in den Sozialwissenschaften beschrieben werden.

Der zweite Teil dieser Arbeit behandelt hauptsächlich die in den unterschiedlichen Wissenschaftsdisziplinen existierenden Konzepte und Begriffssysteme für Fehlerquellen der international vergleichenden Umfrageforschung. Es wird dargestellt, was als Fehler in den Sozialwissenschaften aufgefasst wird, zwischen welchen Fehlerquellen unterschieden werden kann, beziehungsweise ob länderspezifische Unterschiede bezüglich der Wahrscheinlichkeit des Auftretens und des Ausmaßes von Fehlern bekannt sind. Dann wird erörtert, welche Auswirkungen können die

verschiedenen Arten von Fehlern beziehungsweise mögliche internationale Variationen bezüglich des Ausmaßes von Fehlern auf die Vergleichbarkeit von Daten haben können.

Der dritte Abschnitt dieser Arbeit beginnt mit einer Vorstellung des Studiendesigns des European Social Surveys und der Begründung warum das Ausmaß der Vergleichbarkeit der erhobenen Daten für Deutschland, Großbritannien und Österreich überprüft wird. Anschließend daran wird festgestellt, welche Möglichkeiten bezüglich der Qualitätsüberprüfung von Daten im Rahmen einer Sekundäranalyse für einen internationalen Vergleich existieren. Dies beinhaltet eine Darstellung der Methoden beziehungsweise die Erörterung der Vor- und Nachteile dieser Methoden für die Zielsetzung dieser Arbeit. Es sollen demzufolge Strategien aufgezeigt werden um eventuell vorhandene Fehlerquellen zu entdecken, zu identifizieren, beziehungsweise diese eventuell zu neutralisieren, so dass fehlerfreie Analysen mit den erhobenen Daten möglich sind.

Der letzte Abschnitt beinhaltet schließlich die Sekundäranalyse, um festzustellen, ob und in welchem Ausmaß die mit den zwei Skalen erhobenen Einstellung und Meinung zu Migration über die Länder Deutschland, England und Österreich hinweg vergleichbar sind oder nicht. Es kommen hierfür mehrere statistische Verfahren zur Anwendung, da diese unterschiedliche Vor- und Nachteile bezüglich der Entdeckung von Verzerrungen auf Ebene der Konstrukte und der einzelnen Items aufweisen. Die Sekundäranalyse beinhaltet die Überprüfung des Ausmaßes der Vergleichbarkeit der Stichprobendesigns, der semantischen und pragmatischen Bedeutungen der Skalenitems, der Verteilungen dieser, der internen und externen Konsistenz der Skalen und der Dimensionalität und der Faktorenstruktur der zwei Skalen in den einzelnen Ländern. Sollten zum Beispiel die Analysen ergeben, dass die konzeptuelle Struktur der Skalen systematisch unterschiedlich zwischen verschiedenen Staaten wäre, würde ein Vergleich und eine Interpretation der erhaltenen Ergebnisse über Ländergrenzen hinweg wenig Sinn machen. Denn um etwas vergleichen zu können, muss der Gegenstand des Vergleichs auf einer gemeinsamen Grundlage basieren. Sollte dies auf diese Skalen zutreffen wird abschließend erörtert, ob Ursachen für diese vorhandene mangelnde Vergleichbarkeit identifiziert werden können, beziehungsweise ob und welchem Ausmaß diese Daten für international vergleichende Analysen verwendet werden können.

2. Methodologische Grundlagen der international vergleichenden Forschung

Am Beginn dieser Arbeit soll ein Einblick in die Methode des Vergleichs, seine Entwicklung und seine Bedeutung innerhalb der Soziologie vorgenommen werden. Anschließend soll präzisiert werden, welcher unter den zahlreichen für transnationale Studien in der Literatur verwendeten Begriffe der geeignetste für den Zweck der vorliegenden Arbeit ist. Es wird zudem dargestellt welche Typen internationaler Forschung existieren und worin sich deren Zielsetzung unterscheidet.

2.1. Der Vergleich als Methode

Die Wichtigkeit der vergleichenden Forschung für die Sozialwissenschaften ist unbestreitbar. Durch den Vergleich können Aspekte eines Landes, einer Kultur und Unterschiede beziehungsweise Gemeinsamkeiten zwischen Ländern und Kulturen entdeckt werden, die zum Teil nicht anhand von länderspezifischen Daten offenbart würden. Der Vergleich könnte als die Suche zur Identifizierung und Erklärung von Ähnlichkeiten und Unterschieden zwischen oder unter Phänomenen bezüglich verschiedener Kategorien definiert werden. (vgl. Kohn 1987)

> „Die vergleichende Soziologie ist nicht etwa nur ein besonderer Zweig der Soziologie, sie ist soweit die Soziologie selbst, als sie aufhört rein deskriptiv zu sein, und danach strebt, sich über Tatsachen Rechenschaft zu geben.“ (Durkheim 1976: 216)

Wie man an der bekannten Aussage Durkheims erkennen kann, hat die Methode des Vergleichs eine lange soziologische Tradition. Viele der heute als soziologische Klassiker bezeichneten „Gründerväter“ des Hauses der Soziologie, wie zum Beispiel Herbert Spencer, Max Weber oder eben Emile Durkheim, um nur einige zu nennen, wandten diese Methode schon ab Mitte des 19. Jahrhunderts an um soziale Phänomene und die dahinterstehenden Gesetzmäßigkeiten zu entdecken und nachzuweisen. Zusätzlich wurde mit der Methode des Vergleichs das Ziel verfolgt, die Soziologie als eigenständige Wissenschaftsdisziplin zu etablieren und somit von zum Beispiel den

Geschichtswissenschaften abzugrenzen. Nur wurde die Methode des Vergleichs früher fast ausschließlich in Form einer historisch-komparativen Methode, oder anders ausgedrückt als vertikaler Vergleich verstanden und angewendet.

Erst nach 1945 wurden verstärkt synchrone Vergleiche durchgeführt und lösten die bis zu diesem Zeitpunkt dominanten diachronen Vergleiche zunehmend ab. (vgl. Zucha 2004:24) Wobei anzumerken ist, dass die durchgeführten Studien bis in die 70er- Jahre zumeist eher eine nationale Problemstellung aufwiesen (vgl. ebd. 25) Heute werden in erster Linie synchrone Vergleiche durchgeführt, sei es jetzt innerhalb eines Landes oder über mehrere Länder hinweg. So werden zum Beispiel einzelne Untergruppen eines Landes auf Grundlage ihrer Bildung oder verschiedene Länder oder einzelne Untergruppen verschiedener Länder miteinander verglichen. Letztlich werden in den Sozialwissenschaften immer Zusammenhänge zwischen verschiedenen Variablen anhand der Methode des Vergleichs untersucht. Diese Feststellung besitzt nicht nur für die sogenannte quantitative Ausrichtung der Soziologie Gültigkeit, sondern auch für die oftmals hierzu in Gegensatz betrachtete qualitative Ausrichtung wie das nachfolgende Zitat untermauern soll.

> „Comparison across categories (e.g. in the sense of comparing the proportion of cases within a category) is the basic building brick of any social science methodology, including so called „qualitative" ones. What tends to change is simply the nature of the categories and how information about them is collected (or how they are measured)." (MacInnes 2006:101)

Dieses Zitat gilt für die gesamten Sozialwissenschaften und ist demzufolge auch für die internationale Sozialforschung von Bedeutung. Denn von der Grundidee ist die internationale Ausrichtung der Sozialwissenschaften der nationalen Ausrichtung sehr ähnlich. Nur das es bei ersterer hauptsächlich darum geht Nationalstaaten anstatt einzelne Untergruppen eines Landes miteinander zu vergleichen.[4] Der wichtigste Unterschied zwischen der national und international orientierten Sozialforschung besteht also in der Einbeziehung der makrosozialen Analyseebene der Nationalstaaten. Diese makrosoziale Analysedimension stellt die mit Abstand wichtigste Variable in der internationalen Forschung dar. (vgl. Hantrais/Mangen 1998)

[4] Es handelt sich jedoch auch um internationale Sozialforschung, wenn keine ganzen Nationalstaaten verglichen werden. So ist es zum Beispiel auch möglich, dass gewisse Untergruppen verschiedener Länder miteinander verglichen werden.

2.2. Definitionen, Geschichte, Zweck und Arten von international vergleichender Forschung

In der englischsprachigen Literatur existieren verschiedene Bezeichnungen für diese spezielle Art der Forschung wie zum Beispiel *cross-national, cross-cultural, inter-national, cross-societal* oder *comparative research*. In der deutschsprachigen Literatur werden demzufolge Begriffe wie internationale, international vergleichende oder interkulturelle Forschung hierfür verwendet. Stein Rokkan zufolge, sind die englischen Begriffe jeweils einem unterschiedlichem Untersuchungsziel bzw. Studiendesign zuzuordnen. (vgl. Rokkan 1993) „[…]the prefix cross- stresses the objects of comparison while the prefix inter- relates to a characteristic of the research organization." (ebd. 9) In der deutschen Sprache wären aber sowohl *cross-national*, wie auch *inter-national* mit inter-national beziehungsweise international zu übersetzen. Eine sprachliche Abgrenzung dieser Art erscheint in der deutschen Sprache weder sinnvoll noch in der Literatur üblich zu sein.

In der vorliegenden Arbeit wird der Begriff international vergleichende Umfrageforschung verwendet werden, da bei den existierenden großen transnationalen Umfrageforschungsinstrumenten die Abgrenzung der Untersuchungseinheiten jeweils die Staatsgrenzen darstellen und nicht etwa kulturelle Grenzen. Die international vergleichende Forschung stellt eine spezielle Art der vergleichenden Forschung dar. Oftmals jedoch wird der Begriff vergleichende Forschung mit dem Begriff der internationalen Forschung gleichgesetzt, als wenn die einzig möglichen Vergleiche auf internationaler Ebene stattfinden könnten. (vgl. Kohn 1987) Wie bereits dargelegt ist dies aber nicht zutreffend, da der Vergleich wenn auch oftmals implizit, letztlich der Grundbaustein einer jeden soziologischen Methode ist. Die folgende breite Definition von Kohn enthält drei Merkmale anhand derer international vergleichende Forschung charakterisiert werden kann.

> „The broadest possible definition of cross-national research is any research that transcends national boundaries.[…] I prefer to restrict the term cross-national to studies that are explicitly comparative, that is, to studies that utilize systematically comparable data from two or more nations." (Kohn 1987: 714)

Das erste Merkmal der international vergleichenden Forschung stellt also die Involvierung von zwei oder mehr Nationen einer Untersuchung dar. Das zweite Merkmal stellt der Umstand dar, dass diese Art der Forschung, ausdrücklich vergleichend ist. Sie definiert sich sozusagen dadurch, dass etwas, also Phänomene mannigfaltiger Natur miteinander verglichen werden. Dies soll durch den in dieser Arbeit verwendeten Term der international vergleichenden Forschung untermauert werden. Das dritte Merkmal stellt eine gewisse Systematik dar, von der Gebrauch gemacht wird, um letztlich vergleichbare internationale Daten zu erhalten. Die Systematik, die notwendig ist, dass internationale Daten auch wirklich wissenschaftlichen Standards gerecht werden und somit vergleichbar sind werden jedoch etwas später detaillierter erläutert werden. So muss zum Beispiel bereits zu Beginn einer Studie festgelegt werden welche Länder verglichen werden und das Studiendesign dementsprechend entwickelt werden. (vgl. Kohn 1987)

> Einer präziseren Definition von Hantrais und Mangen zufolge beinhaltet die internationale Forschung die Untersuchung von „particular issues of phenomena in two or more countries with the express intention of comparing their manifestations in different sociocultural settings"(Hantrais/Mangen 1998:1). Der Begriff der internationalen Forschung impliziert zudem, dass „one or more units in two or more societies, cultures or countries are compared in respect of the same concepts and concerning the systematic analysis of phenomena, usually with the intention of explaining them and generalising from them" (ebd.: 2).

In dieser Definition wird besser ersichtlich, dass international vergleichende Forschung letztlich dieselben, oder wie später noch dargestellt werden wird, äquivalente Konzepte und Phänomene systematisch über mehrere Länder hinweg analysiert. So können Ursachen beziehungsweise Auswirkungen von sozialen Phänomenen besser erklärt werden. Zudem ist es möglich auf nationaler Ebene vorläufig bestätigte Hypothesen zu testen und so deren Geltungsbereich zu erhöhen.

Die Ziele von international vergleichender Forschung unterscheiden sich von nationalen Studien. Auch wenn die erhaltenen Daten durchaus auch auf nationaler Ebene ausgewertet werden, besteht ihre Hauptnutzung darin, Nationalstaaten beziehungsweise deren Einwohner bezüglich verschiedenster Dimensionen zu vergleichen, um Ähnlichkeiten und Unterschiede zwischen diesen zu identifizieren, zu analysieren und zu erklären. International vergleichende Forschung unterscheidet sich von herkömmlicher nationaler Forschung also dadurch, dass stets zwei Ebenen der Analyse

existieren. Es existieren stets mikro- und auch makrosoziale Erhebungs- und Analyseeinheiten, die zur Erklärung von sozialen Phänomenen herangezogen werden.

> "The goal of most comparative social science is to produce explanations of macrosocial phenomena that are general but also show an appreciation of complexity. In other words, comparative social scientists recognize that a good social scientific explanation is relevant to a variety of cases (if for no other reason than because it uses general explanatory concepts), but at the same time they recognize that social phenomena are complex and that a general explanation is a partial explanation at best." (Ragin 1987:54)

2.2.1. Geschichte der international vergleichenden Umfrageforschung

Die Umfrageforschung hat eine lange internationale und interkulturelle Tradition. Beginnend mit den 50er-Jahren haben sich Sozialwissenschaftler zunehmend mit der Möglichkeit bzw. den methodischen Schwierigkeiten von international vergleichender Umfrageforschung befasst. (vgl. Rokkan 1993)[5] Bis in die späten 70er- Jahre bestand die Mehrheit dieser Studien aus demographischen Arbeiten und Verhaltensstudien. (vgl. Harkness/Vijver/Mohler 2003a) Das Problem bei den meisten vergleichenden Umfragestudien bis zu diesem Zeitpunkt bestand in dem Umstand, dass die jeweiligen Studiendesigns bezüglich eines bestimmten Themas auf nationaler Ebene entwickelt wurden. (vgl. Hoffmeyer-Zlotnik/Harkness 2005) So waren letztlich das Studiendesign, die Durchführung und deren Intention zwischen den einzelnen Ländern nicht wirklich vergleichbar und die gewonnenen Daten mussten nachträglich harmonisiert, also so umrekodiert werden, um Vergleiche überhaupt durchführen zu können. (vgl. Harkness/Vijver/Mohler 2003a)

> Unter Harmonisierung versteht man eine Methode „for equating conceptually similar but operationally different variables that are collected as part of separate surveys for purposes of cross-cultural or cross national research. Also referred to as 'ex post harmonization'." (Johnson 2003:352)

Eine Vergleichbarkeit der Daten beziehungsweise die Validität der so gewonnen Erkenntnisse war also auf Grund des Fehlens einer vergleichbaren Operationalisierung der Variablen in Anbetracht der derzeitigen vorherrschenden methodischen Standards nicht gegeben. So wurden früher oftmals Umfragestudien verschiedener Länder mit derselben Thematik aber unterschiedlichen auf nationaler Ebene konstruierten

[5] Für eine umfassende Darstellung der Geschichte der international vergleichenden Umfrageforschung wird ein Blick auf die Homepage von GESIS unter folgendem Link empfohlen: http://www.gesis.org/en/services/data/portals-links/comparative-survey-projects/?0

Erhebungsinstrumenten miteinander verglichen und dies als international vergleichende Forschung bezeichnet. (vgl. Hoffmeyer-Zlotnik/Harkness 2005)[6] Dies soll allerdings nicht bedeuten, dass diese Problematik in der gegenwärtigen methodologischen Debatte vollständig verschwunden wäre. Denn eine nachträgliche Harmonisierung besonders bei sozialstatistischen Merkmalen, wie zum Beispiel der Schulbildung, ist auf Grund des Fehlens eines einheitlichen internationalen Klassifikationsschemas vielfach noch notwendig.

Im Jahr 1970 entstand mit dem Eurobarometer jedoch ein neuer Typus eines multinationalen Umfragestudie, mit dem vergleichbare Daten systematisch auf Individual- und Haushaltsebene gesammelt wurden. Der Eurobarometer, der von der Europäischen Kommission in Auftrag gegeben wurde und seit seiner Entstehung mindestens zweimal im Jahr durchgeführt wird, erfasst die Einstellungen und Wertorientierungen der Bürger der Mitgliedsstaaten der Europäischen Union (vormals Europäische Gemeinschaft). Der Eurobarometer hat gewissermaßen den Weg für eine weitere Reihe von internationalen Umfrageprojekten geebnet, von denen an dieser Stelle nur die aus europäischer Sicht bedeutsamsten erwähnt werden. Seit 1981 wird der World Value Survey (WVS), mit dem ebenfalls Einstellungen und Wertorientierungen erhoben werden, alle 2 Jahre durchgeführt. Seit 1985 wird das International Social Survey Programme (ISSP), mit welchem Einstellungen und das Verhalten gemessen werden, jährlich durchgeführt. Das jüngste Mitglied der Familie der „großen“ international vergleichenden Umfragestudien stellt der European Social Survey (ESS) dar, der seit 2002 alle zwei Jahre durchgeführt wird und mit dem Einstellungen, Meinungen und Verhalten gemessen werden. Es existieren noch viele weitere vergleichbare Studien, wie zum Beispiel die PISA Studie, die Europäische Werte-Studie (EVS) und EU-SILC.

In den letzten 30 Jahren ist die international vergleichende Forschung zu einem wichtigen Bestandteil des akademischen Bereichs geworden. Dies wird sichtbar an der stetig zunehmenden Anzahl der international vergleichenden Studien beziehungsweise Publikationen über die meisten Wissenschaftsdisziplinen hinweg. Das Themengebiet, das mit den international vergleichenden Umfragen behandelt wird, ist dabei sehr breit gestreut. Es existieren Studien, die ihren Fokus auf Zeitverwendung, ökonomischen und

[6] Warum dies problematisch für die Validität eines Vergleichs ist, wird in Kapitel 3 dargestellt.

politischen Indikatoren, Lebensbedingungen, Lebensqualität, schulische Leistungen und Religiosität haben, um nur einige Themen zu nennen. Analog zu den fokussierten Themen weisen die diversen internationalen Umfragen auch große Unterschiede bezüglich der verwendeten Methoden und Studiendesigns auf, um vergleichbare Daten für die teilnehmenden Substichproben aus den einzelnen Nationalstaaten zu erhalten. Dies betrifft sowohl die Art der Vorgaben, die den teilnehmenden Ländern von den verantwortlichen Forschern gemacht werden, sowie die Art der Organisation des Forschungsablaufs. Während bei manchen Studien wie zum Beispiel dem World Value Survey zwar ein Master-Fragebogen in einer Sprache konzipiert wird und Mindeststichprobengrößen für die Substichproben vorgegeben vorgegeben werden, wird auf Übersetzungsanweisungen für den Fragebogen an die einzelnen teilnehmenden Länder und zentralen Koordinierungsstelle zur Überwachung des Forschungsablaufs verzichtet. Bei anderen Studien wie zum Beispiel dem European Social Survey existiert hingegen praktisch kein Bereich des Forschungsablaufs, für den keine spezifische Vorgaben für die nationalen Forschungsteams gemacht werden und die nicht durch eine zentrale Koordinierungsstelle kontrolliert und dokumentiert werden. (vgl. Lynn, P./Japec, L./Lyberg L. 2006)

Diese Entwicklung der letzten 30 Jahre kann als eine Auswirkung der zunehmenden Globalisierung betrachtet werden, die durch gestiegene Interpendenzen dazu führt, dass ein Individuum zum Beispiel nicht mehr ausschließlich von den Auswirkungen von nationalen Politiken oder Phänomenen betroffen ist. (vgl. Hantrais/Mangen 1998) Dies hat zu einem Anstieg potentieller Finanzierungsquellen für solche Studien geführt, von denen die Europäische Union die wichtigste darstellt. Dies hatte wiederum zur Folge, dass nach den ersten Anfängen sich die Methodologie dieser speziellen Art von vergleichender Forschung durch den gestiegen Anreiz und den neuen Möglichkeiten stets weiterentwickelt hat. Seit den Anfängen der international vergleichenden Umfrageforschung stehen hauptsächlich grundsätzliche methodologische Fragen im Fokus der Sozialwissenschaftler. Zum Beispiel die Frage, ob soziale Phänomene die in verschiedenen sozialen Systemen beobachtet werden, überhaupt verglichen werden können und vor allem wie es geschafft werden könnte, diese durch die Vermeidung von systematischen Messfehlern vergleichbar zu machen. (vgl. Hoffmeyer-Zlotnik/Harkness 2005:5) Gegenwärtig steht die Entwicklung von einheitlichen Klassifikationsschemata zur Erfassung von sozialstatistischen Merkmalen im Fokus des Interesses. Dies ist

besonders für den europäischen Raum zutreffend und hat den Zweck Ausprägungen sozialstatistischer Merkmale auf internationaler Ebene vergleichbarer zu gestalten.

2.2.2. Verschiedene Arten von international vergleichender Forschung und deren Zweck

Es existieren verschiedene Arten der international vergleichenden Forschung, die zumeist nach dem unterschiedlichen Erkenntnisinteresse und dem dadurch unterschiedlichen Studiendesign unterschieden werden. Es werden vier verschiedene Typologien vorgestellt, die die international vergleichende Forschung nach jeweils anderen Gesichtspunkten kategorisieren. Da diese Typologien kombinierbar sind, tragen sie zu einem besseren Verständnis dazu bei, wie facettenreich diese Forschungsart sein kann, beziehungsweise welcher Typ in dieser Arbeit behandelt wird.

Eine fundamentale Unterteilung der international vergleichenden Forschung stammt von Charles C. Ragin. (vgl. Ragin 1987) Dieser unterscheidet nach angewandter Forschungsstrategie grob in fallorientierte und variablenorientierte vergleichende Forschung. Bei der fallorientierten Ausrichtung der vergleichenden Forschung werden lediglich wenige Fälle betrachtet. Diese könnte auch als qualitativ historisch bezeichnet werden. Es steht sozusagen die Komplexität des jeweiligen Kontexts im Mittelpunkt des Forschungsinteresses. Es werden die einzelnen Fälle als Ganzes miteinander verglichen und somit die Komplexität von sozialen Phänomenen zu Lasten einer möglichen Generalisierung zur Forschungsmaxime erhoben.[7] Es werden alle möglichen Konstellationen miteinander verglichen, was letztlich bedeutet, dass nur wenige Fälle, die zumeist eine extreme Ausprägung eines Phänomens aufweisen und einen typischen Fall repräsentieren sollen, in die Analyse aufgenommen werden können. Es werden historisch signifikante soziale Phänomene interpretiert und versucht, deren Ursachen zu bestimmen. Eine kausale Argumentation ist bei dieser Forschungsausrichtung eher nicht möglich, da zumeist mehrere mögliche Erklärungen für das Auftreten eines bestimmten

[7] Ein Beispiel für eine fallorientierte vergleichende Studie wäre die Studie „Party and Society". (vgl. Alford 1963) Darin werden die Effekte von Urbanisierung und Industrialisierung auf den Zusammenhang zwischen sozialer Klasse und Parteienzugehörigkeit in englischsprachigen Demokratien mit einem vergleichbaren Wahlsystem untersucht. Ein anderes Beispiel wäre „The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World". (vgl. Moore 1966) Moore untersucht darin die Ursachen für Diktatur und Demokratie anhand von 8 Ländern, wobei die Länder jeweils paarweise verglichen werden.

sozialen Phänomens existieren und zudem in unterschiedlichen Kontexten unterschiedliche Ursachen-Wirkungs-Verhältnisse existieren können.

In der vergleichenden Forschung, die als variablenorientiert bezeichnet wird und bei der quantitative Methoden zur Anwendung kommen, ist das Ziel die Formulierung von allgemeingültigen Aussagen.(vgl. Ragin 1987) Es werden Hypothesen getestet, die die kausalen Beziehungen zwischen verschiedenen Eigenschaften von Analyseeinheiten beschreiben. Diese Charakteristiken haben die Form von Variablen und es kommen alle statistischen Methoden der modernen Soziologie zur Anwendung. Durch die Natur der statistischen Verfahren, hierbei besonders der multivariaten Verfahren, hat die variablenorientierte vergleichende Forschung einen quasiexperimentellen Charakter. Die unabhängigen Variablen werden zwar nicht vom jeweiligen Forscher kontrolliert, jedoch können Drittvariablen durch diese Methoden entfernt und kontrolliert werden und somit der Effekt jeder einzelnen unabhängigen Variable für das Auftreten eines Phänomens geschätzt werden. Der Nachteil der statistischen Methoden ist allerdings, dass vereinfachende Annahmen getätigt werden, die zur Lasten der Komplexität des einzelnen Falles gehen. Denn es können zumeist nur wenige Variablen, die vom Sozialwissenschaftler als relevant für die Problemstellung erachtet werden, in der Untersuchung berücksichtigt werden. Dass der jeweilige Kontext in den Hintergrund tritt, wird durch den Umstand kompensiert, dass es bei der variablenorientierten Ausrichtung quasi keine Limitierung bezüglich der Anzahl an untersuchten Fällen und Variablen gibt. (vgl. Ragin 1987)

Eine andere mögliche Unterscheidung von international vergleichender Forschung erfolgt nach dem Studiendesign. So können diese nach Studien unterschieden werden, die ein *Most Similar Systems Design* haben und solchen die ein *Most Different Systems Design* haben. (vgl. Przeworski/Teune 1970:31-46) Die Länder stellen beim *Most Similar Systems Design* zumeist die Ausgangsebene der Analyse dar und Variationen zwischen diesen werden durch systemische Faktoren erklärt, die a priori festgelegt werden. Es wird so vorgegangen, dass Länder verglichen werden, die viele ökonomische, kulturelle und politische Charakteristiken gemeinsam haben. Das Ziel dieses Designs besteht im Entdecken von Unterschieden und Gemeinsamkeiten zwischen diesen Ländern, auch wenn der Fokus eher auf der Suche nach Unterschieden besteht. Der Grundgedanke hierbei ist, dass die Gemeinsamkeiten dieser Länder nicht

die gefundenen Unterschiede verursachen können. Wenn Unterschiede zwischen diesen sonst so ähnlichen Ländern entdeckt werden, wird davon ausgegangen, dass die Zahl der möglichen erklärenden Faktoren die diese Unterschiede verursachen, minimiert wird. Ein Problem dieses Ansatzes ist der Umstand, dass zwar einige mögliche Ursachen für gefundene Unterschiede ausgeschlossen werden können, aber dennoch noch immer eine Menge alternativer Erklärungsalternativen übrigbleiben.

Beim *Most Different Systems Design* werden Faktoren der Systemebene in Form von Unterschieden zwischen sehr heterogenen Ländern zu Beginn der Studie nicht als Erklärungsfaktoren für soziale Phänomene betrachtet. Es werden Variationen von Verhalten und Einstellungen auf Individualebene betrachtet, wobei angenommen wird, dass die Individuen einer homogenen Population entstammen. Diese Annahme wird in mehreren Forschungsschritten getestet. Sollte diese Annahme nicht zurückgewiesen werden, bleibt die Analyse auf intrasystemischer Ebene. Sollten intersystemische Unterschiede bezüglich des Verhältnisses zwischen unabhängigen Variablen und der abhängigen entdeckt werden, müssen systemische Faktoren als mögliche Erklärung für die gefundenen Unterschiede in Betracht gezogen werden. Während Studien mit einem *Most Similar Systems Design* nach einer Identifikation von relevanten systemischen Faktoren für die Erklärung von zum Beispiel nationalen Unterschieden streben, geht es bei Studien mit einem *Most Different Systems Design* um die Elimination irrelevanter systemischer Faktoren. Was beide Vorgehensweisen jedoch gemein haben ist, dass Variablen auf Mikro- und Makroebene kombiniert werden können, um letztlich Hypothesen zu prüfen beziehungsweise diese zu generalisieren.

Eine weitere Typologie, die sich für das in dieser Arbeit existierende Erkenntnisinteresse besser eignet, stammt von Melvin L. Kohn (vgl. Kohn 1987). Dieser zufolge kann zwischen vier verschiedenen Typen von international vergleichender Forschung unterschieden werden, die sich je nach ihrer Zielsetzung und Vorgangsweise, unter anderem ersichtlich durch die Art der Verwendung der Kategorie „Land“ und die Auswahl der Länder, unterscheiden. Denn je nachdem welche Länder miteinander verglichen werden, werden die gefundenen Ergebnisse eher Ähnlichkeiten oder Unterschiede offenbaren. (vgl. Tab. 1)

Tabelle 1: Verschiedene Ansätze für internationale Vergleiche

Country as	Object of study	Context of study	Unit of analysis	Part of larger system
Primary purpose	Idiographic – understand each country in own terms	Test abstract hypothesis or dimension across countries	Seek relations among dimensions on which nations vary	Interpret each country as part of transnational system
Country selection	Compare any, all or similar countries	Maximise diversity on one dimension	Diversity within a common framework	Maximise diversity on all dimensions

Quelle: Kohn 1987 in Lobe/Livingstone/Haddon (2007: 39)

Im ersten Typus stellen Nationalstaaten das Objekt der Studie dar. Die Zielsetzung in dieser Art von internationaler vergleichender Forschung stellt in erster Linie ein Vergleich von bestimmten Institutionen in verschiedenen Ländern dar, wobei es im Wesentlichen darum geht, länderspezifische Informationen zu sammeln. Das Ziel ist eher die Erkenntnis zu erlangen worin sich Länder unterscheiden, als die Formulierung generalisierbarer Hypothesen. Eine Auswahl ähnlicher Länder erscheint hier am sinnvollsten zu sein, besonders wenn zum Beispiel die Auswirkungen gewisser Politiken untersucht werden sollen.

Im zweiten Typus stellen Nationalstaaten den Kontext der Studie dar. Hier wird hauptsächlich die Generalisierbarkeit von gefundenen Ergebnissen beziehungsweise Hypothesen bezüglich eines sozialen Phänomens getestet und interpretiert, wie gewisse soziale Institutionen operieren beziehungsweise funktionieren. Wenn nach einer Generalisierung von Aussagen gestrebt wird, sollte eine Auswahl möglichst zahlreicher und heterogener Länder bezüglich der interessierenden Dimension, die höchste Universalität der des untersuchten Phänomens zur Folge haben.

Im dritten Typus stellen Nationalstaaten die Analyseeinheiten dar. Bei dieser Art des internationalen Vergleichs wird versucht, Beziehungen zwischen verschiedenen Charakteristiken von Ländern zu entdecken beziehungsweise nachzuweisen. Die Zielsetzung stellt hierbei eine Klassifizierung von Ländern entlang einer oder mehrerer Dimensionen anhand gewisser Charakteristiken dar, wie zum Beispiel das Bruttoinlandsprodukt. Bei diesem Typus sollten die Länder so ausgewählt werden, dass die Heterogenität bezüglich der interessierenden Dimensionen repräsentiert wird.

Der vierte Typus hat transnationalen Charakter, dass heißt Nationalstaaten werden als Subsysteme eines größeren internationalen Systems angesehen. Es wird angenommen, dass Länder durch gewisse Prozesse, wie zum Beispiel die Globalisierung systematische Beziehungen zueinander aufweisen. (vgl. Lobe/Livingstone/Haddon 2007) Das Ziel ist es, eine transnationale Erklärung zu formulieren, die in Beziehung zu den einzelnen nationalen Kontexten gesetzt wird. Die Auswahl der Länder sollte bezüglich der relevanten Dimensionen eine möglichst hohe Diversität aufweisen.

Schließlich kann zwischen die strukturorientierten und levelorientierten Studien unterschieden werden. (vgl. Vijver 2003a) Strukturorientiert ist eine Studie, wenn die Fragestellung lautet, ob ein Instrument, das gleiche Konstrukt in verschiedenen Ländern misst. Anders ausgedrückt geht es vornehmlich darum, Bedeutungsunterschiede und Gemeinsamkeiten von diversen sozialwissenschaftlichen Konzepten, wie zum Beispiel Religiosität zu entdecken. Der zweite Typus, die levelorientierten Studien, bezeichnet Untersuchungen, bei denen die Entdeckung und Darstellung von länderspezifischen Mittelwertsunterschieden im Mittelpunkt des Erkenntnisinteresses steht.

Da wie gezeigt wurde verschiedene Formen der international vergleichenden Forschung existieren, scheint eine Präzisierung dahingehend angebracht, dass unter diesem Begriff in dieser Arbeit auf Grund der speziellen Zielsetzung, internationale Umfragen verstanden werden. Es handelt sich bei dem in dieser Arbeit behandelten Forschungstypus also um international vergleichende Umfrageforschung, die variablen- und strukturorientiert ist und die Nationalstaaten als Kontext beziehungsweise als Objekt der Studie behandelt. Ein Ziel dieser Umfragen kann das Testen von Hypothesen beziehungsweise Theorien mit dem Zweck der Generalisierung dieser sein. Zweck der Untersuchung kann allerdings auch die Suche, Analyse und Erklärung von Unterschieden und Gemeinsamkeiten zwischen Ländern bezüglich der Struktur der Beziehung von diversen Phänomenen zueinander sein. Zumeist handelt sich bei dem verursachenden Faktor um ein makrosoziales Phänomen, wie zum Beispiel die Sozialstruktur eines Landes, das letztlich die Beschaffenheit eines sozialen Phänomens determiniert.

3. Fehlerquellen der international vergleichenden Umfrageforschung

3.1. Definitionen und Darstellung der verschiedenen Konzepte: Fehler, Verzerrung und Äquivalenz

Zu Beginn dieser Arbeit wurde argumentiert, dass in den Sozialwissenschaften, gleichgültig ob national oder international, letztlich immer verglichen wird. Es wurde dargestellt, dass sich die Kategorien, hierbei besonders das Hinzufügen der Variable Land und die Ziele dieser Ausrichtungen unterscheiden. Die Unterschiede zwischen nationaler und international vergleichender Umfrageforschung bestehen allerdings nicht nur darin, sondern zeigen sich auch in unterschiedlichen methodischen Anforderungen. So hat die international vergleichende Umfrageforschung mit zusätzlichen methodischen Problemen die die Vergleichbarkeit von Daten gefährden können zu kämpfen, die bei nationalen Studien relativ unbedeutend sind. Es könnte gesagt werden, dass sich die Fehlerquellen mit der Anzahl der Länder, die untersucht werden sollen gewissermaßen vervielfachen. In der international vergleichenden Forschung hängt nämlich die Qualität der Rückschlüsse, die auf Grund der Messungen gezogen werden, von der Qualität einer jeder nationalen Teilstudie ab. (vgl. Jowell/Kaase/Fitzgerald/Gillian 2007)

Zudem existieren linguistische, kulturelle und konzeptuelle Barrieren welche einen internationalen Vergleich erschweren. Dies führt dazu, dass diverse Konzepte, die in der nationalstaatlichen Umfrageforschung verwendet worden sind, in anderen Ländern nicht in dieser Form existieren beziehungsweise anders interpretiert werden. Neben diesen Barrieren existieren auch länderspezifische Unterschiede und Vorlieben beziehungsweise Erfahrungswerte bezüglich der verschiedenen Erhebungsmodi und des Stichprobendesigns (vgl. ebd.). Dies ist auch bei Kodierungsvorgängen, dem Interviewer-Training, den sozioökonomischen Klassifikationen, sowie bei den üblichen Rücklaufquoten zu beobachten. Linguistische und konzeptuelle Unterschiede stellen das häufigste Hindernis eines Vergleichs der einzelnen Länderergebnisse dar. (ebd.).

Im folgenden Teil dieser Arbeit werden die Begriffsysteme Fehler, Verzerrung und Äquivalenz dargestellt und definiert. Diese Typologien werden in unterschiedlichen Wissenschaftsdisziplinen verwendet. Während die im folgenden Teil beschriebene Fehler-Typologie für gewöhnlich in der Soziologie zur Anwendung kommt, haben die Typologien der Verzerrung und der Äquivalenz ihren Ursprung in der auf Umfragen basierenden Ausrichtung der vergleichenden Psychologie. (vgl. Braun 2003b) Es erscheint wichtig, dass diese Terminologien und die typologische Abgrenzungen definiert und klar abgegrenzt werden.[8] Denn diese Typologien weisen trotz einiger Gemeinsamkeiten unterschiedliche Schwerpunktsetzungen bezüglich möglicher Fehlerquellen im Verlauf einer Umfrage auf.

Bevor die Begriffsysteme Fehler, Verzerrung und Äquivalenz dargestellt werden, sei noch einmal darauf hingewiesen, dass in der vorliegenden Arbeit primär die Vergleichbarkeit der Ergebnisse zweier Skalen behandelt wird. In den Sozialwissenschaften wird häufig versucht soziale Phänomene zu messen und zum Beispiel die Auswirkungen auf das Verhalten von diesen Individuen festzustellen. Diese sozialen Phänomene stellen allerdings zumeist abstrakte theoretische Konzepte dar, die nicht direkt beobachtbar sind, weil sie mehrere Dimensionen enthalten.[9] Sie können also nicht direkt gemessen werden. Es ist die Aufgabe des Sozialwissenschaftlers eine Konzeptspezifikation vorzunehmen, um die relevanten Dimensionen zu bestimmen, aus denen sich das Konzept zusammensetzt. (vgl. Schnell/Hill/Esser 2008:128-129) Jede dieser Dimensionen wird als Konstrukt bezeichnet, die zumeist noch nicht direkt messbar sind, beziehungsweise es nicht möglich ist alle relevanten Aspekte einer Dimension mittels einer Variablen zu erfassen. (vgl. Harkness/Van de Vijver/Mohler 2003) Im Verlauf der Operationalisierung erfolgt schließlich die Zuweisung von manifesten Variablen um diese latenten Konstrukte messbar zu machen. Diese manifesten Variablen in Form von Fragebogenfragen werden als Indikatoren bezeichnet. Indikatoren werden also verwendet um latente Konstrukte zu erfassen und diese repräsentieren wiederum ein zugrundeliegendes abstraktes theoretisches Konzept.

[8] Die einzelnen Begriffe werden hinsichtlich ihrer in der internationalen Umfrageforschung üblichen Bedeutungen verwendet und müssen nicht notwendigerweise mit den in nationalen Studien vorherrschenden Verwendungsweisen deckungsgleich sein.

[9] In der englischsprachigen Literatur wird auch zwischen *concepts-by-intuition* und *concepts-by-postulation* unterschieden. (vgl. Saris/Gallhofer 2007b) *Concepts-by-intuition* bezeichnen einfache Konzepte deren Bedeutung klar ist und die direkt gemessen werden können. Diese Konzepte beinhalten Urteile, Gefühle, Bewertungen, Normen und Verhalten. *Concepts-by-postulation* hingegen entsprechen der in der deutschsprachigen Soziologie üblichen Verwendungsweise von Konzepten. Sie müssen aus der Theorie definiert werden und sind häufig mehrdimensional.

Skalen sind Spezialfälle von Indizes, für die Beurteilungskriterien existieren, ob ein Indikator zu einer Skala gehört oder nicht. (vgl. Schnell/Hill/Esser 2008:181-182)

3.1.1. Fehler

Ein Fehler stellt die Differenz zwischen „wahren“ und beobachteten Werten dar. Groves unterscheidet zwischen vier verschiedenen Typen von Fehlern in Umfragen (vgl. Groves 1989; Braun 2003b):

- *Sampling error*
- *Coverage error*
- *Nonresponse error*
- *Measurement error: Instrument bias*, *Interviewer bias*, *Mode bias* und *Respondent bias*

Die ersten drei genannten Fehlertypen befassen sich mit Fehlern, die mit der Stichprobenziehung in Zusammenhang stehen und lediglich der letzte Typus bezieht sich auf das eigentliche Instrument. Erwähnenswert erscheint noch der Umstand, dass sowohl der *Sampling* als auch der *Coverage error* unabhängig vom Verhalten der zur interessierenden Grundgesamtheit gehörenden Individuen ist, was für den *Nonresponse* beziehungsweise den *Measurement error* nicht zutreffend ist. (vgl. Braun 2003b)

Sampling error

Sampling error entsteht durch den Umstand, dass für Umfragen Stichproben gezogen werden und nicht die gesamte interessierende Population befragt wird. Der Stichprobenfehler ist die Abweichung der auf Basis einer Stichprobe geschätzten Werte von den realen Populationswerten. Der Stichprobenfehler kann allerdings nur exakt berechnet werden, wenn die folgenden 3 Fehlertypen vernachlässigbar wären, was in der Umfrageforschung selten der Fall ist.

Coverage error

Coverage error entsteht, wenn nicht alle Befragten, die der interessierenden Grundgesamtheit angehören, eine von Null unterschiedliche Auswahlwahrscheinlichkeit haben, in der Stichprobe enthalten zu sein. Wenn keine komplette Liste der Untersuchungseinheiten existiert, besteht die Gefahr, dass gewisse Mitglieder der interessierenden Grundgesamtheit keine Chance haben in die Stichprobe zu gelangen. Dies hat je nach Forschungsfrage und Forschungsdesign unterschiedlich gravierende Auswirkungen auf den *Coverage error*. Sollte zum Beispiel eine Gruppe von Individuen, die eine gemeinsame Ausprägung bezüglich eines für das Untersuchungsziel relevanten Variable aufweisen, eine Auswahlwahrscheinlichkeit von Null kann dies beträchtliche Auswirkungen auf die Qualität der Erklärung eines sozialen Phänomens haben. Die so erhaltenen Ergebnisse wären nicht „repräsentativ" für die Population, für die eigentlich Aussagen getroffen werden sollten.[10] Werden verschiedene Stichprobendesigns in verschiedene Ländern verwendet, stellt dies an sich noch kein Problem dar, solange es sich bei den Stichproben um Auswahlverfahren handelt, denen eine einheitliche Definition der interessierenden Population zugrunde liegt. (vgl. Häder/Gabler 2003) Zudem müssen diese Designs alle den Prinzipien einer wahrscheinlichkeitsbasierten Prozedur entsprechen und die Auswahlwahrscheinlichkeit der Einheiten muss bekannt sein.(vgl. Häder/Lynn 2007) Dann können die Substichproben der einzelnen Länder gewichtet werden und somit vergleichbar gemacht werden. Ein identisches Stichprobendesign für alle teilnehmenden Länder an einer Umfrage ist demnach nicht notwendig und auch ziemlich unwahrscheinlich wie folgendes Beispiel verdeutlichen soll. Es existieren in einigen Ländern, wie zum Beispiel in Norwegen, öffentlich zugängliche Melderegister die für die Stichprobenziehung genutzt werden können und dies stellt in diesen Fällen sowohl aus ökonomischer wie auch aus soziologischer Sichtweise gewissermaßen den Königsweg dar. (vgl. Häder/Lynn 2007) Diese Länder sind an dieses Stichprobendesign gewöhnt und es existieren wenig Alternativen. In anderen Ländern, wie zum Beispiel in Österreich existiert so ein Register zwar, ist aber jedoch nicht öffentlich zugänglich und kann somit nicht für die Stichprobenziehung verwendet werden. Für diese Länder,

[10] Für eine Diskussion des inflationär, uneinheitlich und oft missbräuchlich verwendeten Begriffs der „Repräsentativität" siehe Schnell/Hill/Esser 2008:304-306.

sowie für jene bei denen erst gar kein zentrales Melderegister existiert, sind jedoch andere Designs entwickelt und Praxis geworden.[11]

Nonresponse error

Nonresponse error entsteht, weil nicht alle für die Befragung vorgesehenen Individuen, die die Bruttostichprobe bilden, tatsächlich an dieser teilnehmen. Der *Nonresponse error* ergibt sich aus der Differenz zwischen der Bruttostichprobe und der Nettostichprobe, also dem Unterschied zwischen der vorgesehenen Stichprobengröße und der tatsächlich realisierten Stichprobengröße. Der Begriff *Nonresponse error* muss allerdings dahingehend präzisiert werden, dass in der gängigen Literatur zwischen *Unit-Nonresponse* und *Item-Nonresponse* unterschieden wird. (vgl. Schnell/Hill/Esser 2008:353-357) *Item-Nonresponse* bezeichnet den Umstand, dass Befragte sich weigern auf bestimmte Fragen eine Antwort zu geben. Diese Fehlerquelle ist dieser Typologie zufolge dem *Measurement error* zuzuordnen. In dieser Typologie wird unter *Nonresponse error* lediglich das Auftreten von *Unit-Nonresponse* erfasst. Die Hauptursachen für das Auftreten von *Unit-Nonresponse* in Umfragen sind, dass kein Kontakt mit der zu befragenden Person hergestellt werden kann, die Weigerung der zu Befragenden an dieser Umfrage teilzunehmen und ein Interviewabbruch durch den Befragten. (vgl. Braun 2003b) Das Auftreten von *Nonresponse error* kann verschiedene Ursachen und Auswirkungen auf die Qualität einer Umfrage haben, die im internationalen Vergleich differieren können. Die Ursachen betreffen zum Beispiel Merkmale der Befragten, Interviewlänge, Thema der Umfrage, Erhebungsmodus, Fragebogendesign, sowie Thema der Umfrage. (vgl. Couper/Leeuw 2003) Die Auswirkungen können ähnlich gravierend wie die des bereits besprochenen *Coverage errors* sein. Sollten sich die Gruppen der Befragten und der Nicht-Befragten stark bezüglich eines für die Untersuchung relevanten Merkmales unterscheiden, stellt dies einen systematischen Fehler dar und hat somit einen Einfluss auf die Qualität und die Generalisierbarkeit und gefährdet somit die Vergleichbarkeit der Erkenntnisse über mehrere Länder hinweg. Die Schwierigkeit das Ausmaß des *Nonresponse errors* zu bestimmen besteht in dem Umstand, dass zumeist über den eigentlich zu Befragenden keinerlei Information vorliegt. Somit lassen sich eventuell vorhandene systematische Fehler schwer ausschließen. Dies zeigt die Wichtigkeit der Erreichung einer ähnlichen

[11] Für eine Darstellung von drei Stichprobendesigns, die in verschiedenen Ländern häufig verwendet werden siehe Kapitel 5.1.

hohen Ausschöpfungsrate, welche das Spiegelkonzept des *Nonresponse errors* darstellt. Doch ist gerade diese in den einzelnen Ländern aus verschiedenen Gründen rückläufig und vor allem auf unterschiedlichem Niveau. (vgl. Couper/Leeuw 2003) Die unterschiedlichen Ausschöpfungsraten können als ein Indikator für eine unterschiedliche Befragungsakzeptanz in den einzelnen Ländern aufgefasst werden. Doch selbst wenn diese sich auf demselben Niveau befinden würden, wäre dies noch keine Garantie für eine etwaige Abstinenz von einem systematischen Fehler durch *Nonresponse*. Dies könnte lediglich sichergestellt werden, wenn die Ursachen für *Nonresponse* auf eine länderübergreifende Weise einheitlich genauer hinterfragt und dokumentiert würden.

Measurement error

Measurement error liegt dann vor, wenn der mit einem Erhebungsinstrument gemessene Wert nicht mit dem „wahren" Wert übereinstimmt. Eine Befragung beinhaltet sowohl soziale wie auch kognitive Prozesse. (vgl. Sudman/Bradburn/Schwarz 1996) Eine Befragung stellt eine spezielle soziale Situation dar, in der Befragte und Interviewer sozial interagieren. Sie unterscheidet sich von einer alltäglichen Situation durch eine asymmetrische Strukturierung der Kommunikation und eine theoriegeleitete Kontrolle der Situation. (vgl. Atteslander 2003) Der Befragte wird gebeten, ihm unbekannten Personen eine Auskunft über Dinge zu geben, die seine Privatsphäre betreffen. Der Forscher präsentiert dem Befragten eine Reihe von Stimuli in Form von Fragen und Antwortkategorien entweder direkt mittels eines schriftlichen Fragebogens oder indirekt über eine dritte Person, dem Interviewer, auf die der Befragten in Form einer Antwort reagieren soll. Die Aufgabe des Befragten ist es, die verschiedenen Stimuli wahrzunehmen und zu interpretieren, aus seinem Gedächtnis die relevanten Informationen abzurufen um eine Reaktion zu generieren und diese Informationen an die vorgegebenen Antwortkategorien anzupassen. (vgl. Braun, M. 2003a) Die Abgabe der Antwort erfolgt schließlich nach einer Kosten- Nutzen- Abwägung zwischen den zwei motivationalen Grundbedürfnissen der sozialen Anerkennung und der Vermeidung von Missbilligung. (vgl. Esser 1986) Ein Problem mit dem die international vergleichende Forschung konfrontiert ist, stellen allerdings variierende kulturelle Normen, Werte und Erfahrungen dar, die diese Antwortprozesse beeinflussen. (vgl. Johnson u.a. 1997) Messfehler werden schließlich durch zwischen den einzelnen

Interviews einer Befragung unterschiedlich strukturierte kognitive, kommunikative und motivationale Prozesse verursacht.

Die vier Gründe, die für Unterschiede bezüglich dieser Prozesse verantwortlich sein können, können das Erhebungsinstrument, der Interviewer, der Befragte und der Erhebungsmodus identifiziert werden. Dementsprechend werden die Messfehler in *Instrument bias*, *Interviewer bias*, *Respondent bias*, und *Mode bias* unterteilt. So eindeutig diese Unterscheidungen auf den ersten Blick sein mögen, wird im Verlauf der Darstellung der einzelnen Messfehlertypen doch deutlich, dass eine eindeutige Trennung ihrer Effekte auf Grund von Interpendenzen zwischen den kognitiven, kommunikativen und motivationalen Prozessen nicht immer möglich ist. Dies betrifft besonders die Unterscheidung zwischen *Instrument bias* und *Respondent bias*. So existiert in der Literatur kein Konsens darüber, ob systematische Anwortverzerrungen die durch Antworttendenzen verursacht werden, eine Folge von Charakteristiken des Erhebungsinstruments oder von Persönlichkeitsmerkmalen darstellen. (vgl. Moors 2008)

Instrument bias

Instrument bias beinhalten Messfehler die durch Charakteristiken des Fragebogens verursacht werden. So kann das Fragebogendesign, die Formulierung einzelner Fragen, deren Reihenfolge und die Antwortkategorien die Antworten verzerren und so Messfehler verursachen.[12] Antwortverzerrungen durch das Erhebungsinstrument betreffen die ersten zwei der zuvor genannten Prozesse der Antwortbildung, also die Beschaffenheit, Wahrnehmung und Interpretation der Stimuli in Form einer Fragebogenfrage und den kognitiven Informationsverarbeitungsprozess. In den einzelnen Ländern existiert eine unterschiedliche Vertrautheit mit Umfrageprozeduren und dies kann auch zu einer unterschiedlichen Wahrnehmung des Erhebungsinstruments beitragen. (vgl. Kleiner/Pan 2006) Es kann nicht davon ausgegangen werden, dass ein Erhebungsinstrument, das in einem bestimmten kulturellem Umfeld, wie zum Beispiel den mitteleuropäischen Ländern, entwickelt wird ohne weiteres auf andere Länder übertragbar ist.

[12] Für die Vor- und Nachteile diverser Prozesse das Fragebogendesign für international vergleichende Umfrageforschung betreffend vgl. Harkness/Van de Vijver./Johnson 2003; Smith 2003.

Wörter, Sätze und Satzgebilde haben eine semantische und pragmatische Bedeutung. Semantische Bedeutung bezeichnet die Bedeutung oder die Bedeutungen, die mit einzelnen Wörtern oder gewissen Wortkombinationen assoziiert wird. Pragmatische Bedeutung bezeichnet demgegenüber, dass die Bedeutung durch die Abhängigkeit von dem was gesagt wird von dem Kontext in dem es gesagt wird determiniert wird. Dies kann dazu führen, dass ähnliche Phänomene eine unterschiedliche Bedeutung in verschiedenen Kontexten haben können, während unterschiedliche Phänomene eine vergleichbare Bedeutung in verschiedenen Kontexten haben können. (vgl. Van Deth 2003) Die Übersetzung eines Erhebungsinstruments in verschiedene Sprachen ist eine ziemlich offensichtliche und häufig auftretende Fehlerquelle in der international vergleichenden Umfrageforschung, die eng mit der semantischen Bedeutung zusammenhängt. Es existieren allerdings auch diverse Kontexteffekte, die einen Einfluss auf die pragmatische Bedeutung eines Inhalts haben und einen *Instrument bias* verursachen können. Dies kann zu einer unterschiedlichen Wahrnehmung und Interpretation der in einer Umfrage präsentierten Stimuli und unterschiedlichen kognitiven Informationsverarbeitungsprozessen führen. Es kann zwischen drei verschiedenen Kontexteffekten unterschieden werden die oftmals gemeinsam auftreten und deren Wirkungsbereiche dementsprechend zum Teil nicht voneinander separiert werden können. (vgl. Braun/Harkness 2005) Die pragmatische Bedeutung eines Inhalts kann durch den Kontext des Erhebungsinstruments, die persönlichen Erfahrungen des Befragten und den kulturellen Kontext determiniert sein. Dies ist in engem Zusammenhang mit dem kognitiven Informationsverarbeitungsprozess erklärbar. (vgl. Sudman/Bradburn/Schwarz 1996; Braun 2003 a). So existieren abstrakte schematische Wissensstrukturen, die die Wahrnehmung und Interpretation neuer Informationen lenken. Diese Schemata werden automatisch als Reaktion auf einen Stimulus aktiviert. Diese Schemata unterscheiden sich in ihrer Zugänglichkeit, wobei auf kürzlich aktivierte Schemata bei der selektiven Informationsbeschaffung eher zurückgegriffen wird.

Wenn man vorhat ein Konstrukt über Kulturen beziehungsweise Ländergrenzen hinweg zu vergleichen, stellt sich demnach die Frage, ob und wie die Übersetzung der jeweiligen Items einer Skala die Resultate beeinflusst. Es geht dabei vornehmlich um die Herausforderung, die Semantik und die Intention einzelner Fragen und deren Antwortkategorien im Ländervergleich zu erhalten und somit sicherzustellen, dass auf

die Befragten aus verschiedenen Ländern ein gleichwertiger Stimulus einwirkt, auf den sie reagieren sollen. (vgl. Harkness/Schoua-Glusber 1998) Einzelne Begriffe oder ganze Fragen beziehungsweise deren Antwortkategorien können aber auch wenn sie oberflächlich betrachtet korrekt übersetzt sind, oftmals eine andere Bedeutungsdimension beinhalten, die bei der Übersetzung leicht übersehen werden können. Die existierenden strukturellen Unterschiede zwischen verschiedenen Sprachen stellen ein weiteres Problem dar. Diese beinhalten zum Beispiel grammatikalischer Unterschiede bezüglich der Existenz, einer unterschiedlicher Anzahl oder Verwendungsweisen von Geschlechtsreferenzen in verschiedenen Sprachen. (vgl. Harkness 2003) Das beeinflusst die Art und Weise wie Individuen auf sich selber, auf andere oder ihr Umfeld Bezug nehmen und stellt einen Teilbereich des zuvor beschrieben kulturellen Kontexts dar. Dies kann zu einem unterschiedlichen Frageverständnis von verschiedensprachigen Individuen führen.

Ähnliches ist für Antwortskalen festzustellen. So existieren Untersuchungsergebnisse, die vermuten lassen, dass identische Antwortskalen nicht in allen und Kulturen „gleich funktionieren“ und es interkulturell variierende Interaktionen zwischen Antwortverzerrungen und Antwortskalen gibt. (vgl. Harkness/Van de Vijver/Johnson 2003) So gibt es Untersuchungen, die darauf hinweisen, dass sich die wahrgenommene Intensität von sonst eigentlich korrekt übersetzten verbalisierten Antwortskalen zwischen verschiedenen Sprachen unterscheidet. (vgl. Mohler/Smith/Harkness 1998) Dies führt letztlich dazu, dass die so verzerrten Skalenwerte nicht mehr zwischen zwei Gruppen vergleichbar sind. Selbst wenn nonverbale Antwortskalen verwendet werden, ist ein *Instrument bias* nicht auszuschließen, da die Wahrnehmung und kulturell verankerte Verwendungsweisen von Zahlen und visuellen Antwortmöglichkeiten in verschiedenen kulturellen Kontexten ebenfalls variieren kann. (vgl. Smith 2003)

Ein Beispiel für einen Kontexteffekt des Erhebungsinstruments stellen die sogenannten Fragereiheneffekte oder Positionseffekte dar. Diese bezeichnen den Umstand, dass eine vorhergehende Frage den kognitiven Informationsverarbeitungsprozess einer nachfolgenden Frage beeinflusst und so eine Antwortverzerrung bedingt. Ein Positionseffekt würde dann auftreten, wenn ein durch eine vorangegangene Frage aktiviertes Schema für die Informationsbeschaffung im Rahmen des Antwortprozesses einer nachfolgenden verwendet wird. So haben Methodenexperimente gezeigt, dass auf

Grund unterschiedlich kultureller Normen die Bereitstellung von neuen Informationen in einer Konversation betreffend, Fragereiheneffekte unterschiedlichen Ausmaßes zwischen individualistischen und kollektivistischen Kulturen zur Folge haben können. (vgl. Schwarz 2003) Eine nicht intendierte Beeinflussung der Befragten kann allerdings auch zum Beispiel Einleitungstexte von Frageblöcken oder Antwortkategorien erfolgen. Sollten Fragen für den Befragten unklar sein oder Information bezüglich des Frageinhalts unzureichend kognitiv verankert sein, können Befragte versuchen diese Information dem Erhebungsinstrument zu entnehmen. (vgl. Braun 2003a) Es wurde gezeigt, dass Befragte Antwortskalen verwenden um Fragen zu interpretieren und vice versa. (vgl. Mohler/Smith/Harkness 1998) Zudem wurde festgestellt, dass der kulturelle Kontext einen Einfluss auf die Wahrnehmung der Antwortalternativen haben kann, was wiederum zu einer unterschiedlichen Informationsfunktion von diesen und zu Antwortverzerrungen führt.(vgl. Braun 2003a)[13] Aus der nationalen Umfrageforschung ist zum Beispiel bekannt, dass ein berichtetes Verhalten nicht unerheblich von den Antwortvorgaben beeinflusst ist. Dies ist auch in Zusammenhang mit dem später erläuterten Effekt der sozialen Erwünschtheit zu betrachten. Methodenexperimente haben gezeigt, dass Antwortkategorien mit verschiedenen auszuwählenden Häufigkeiten zwischen individualistischen und kollektivistischen Kulturen, eher die Antworten der Angehörigen individualistischer Kulturen verzerren. (vgl. Schwarz 2003)

Interviewer bias

Interviewer bias beinhalten Antwortverzerrungen, die durch Reaktionen der Befragten auf Merkmale der Interviewer oder deren Verhalten während des Interviews verursacht werden. Die Beeinflussung der Antworten durch Merkmale oder das Verhalten der Interviewer, ist in Zusammenhang mit dem Auftreten von sozialer Erwünschtheit zu betrachten. Sollten nun gewisse Merkmale des Interviewers mit dem Zweck einer Umfrage in irgendeiner Weise zusammenhängen, ist ein Auftreten von Antwortverzerrungen durch sozial erwünschte Antworten als Reaktion des Befragten auf eben diese Merkmale möglich. So wurde zum Beispiel gezeigt, dass das

[13] Wenn zum Beispiel die Einstellung zu vorehelichen Geschlechtsverkehr in einem modernen Land mit der in einem traditionellen Land verglichen werden, bewirkt ein unterschiedliches Durchschnittsalter bei der Erstheirat eine unterschiedliche Interpretation dieser Frage. (vgl. Braun 2003a) Während in traditionellen Ländern bei der Beantwortung der Frage eher an Erwachsene gedacht wird, müsste in modernen Ländern in der Fragestellung ein dementsprechendes Mindestalter angegeben werden, um die Ergebnisse vergleichbar zu gestalten.

Antwortverhalten von Befragten durch die empfundene kulturelle beziehungsweise soziale Distanz international variierende Antwortmuster verursachen kann (vgl. Johnson/Van de Vijver 2003; Kleiner/Pan 2006) Weitere Ursachen für *Interviewer bias* stellen Abweichungen vom vorgegebenen Fragentext, eine falsche Zuordnung der erhaltenen Antwort zu den Fragekategorien und bewusste Fälschungen dar. Diese Fehlerquellen verdeutlichen die Notwendigkeit einer guten Interviewschulung in den einzelnen Ländern. Nicht zuletzt auf Grund des bereits dargestellten Umstandes, dass in unterschiedlichen Länder verschiedene Umfragetraditionen existieren und dies auch einen Einfluss auf das Training und die Erfahrungswerte der Interviewer hat. Allerdings erscheinen gerade bezüglich des Interviewerverhaltens „best-practice" Anweisungen bezüglich deren Verhalten während des Interviews auf Grund von Normvariationen und darausfolgenden unterschiedlichen Verhaltenkodizes in verschiedenen Ländern ein Problem darzustellen. So verursacht zum Beispiel ein für Umfragen in westlichen Gesellschaften als ideal erachtetes Interviewerverhalten, das für den Interviewer die Rolle des neutralen Beobachters vorsieht, in Japan Antwortverzerrungen. (vgl. Johnson/Van de Vijver 2003)

Mode bias

Mode bias betreffen die unterschiedlichen Auswirkungen von diversen Erhebungsmodi. Die Wichtigkeit der Verwendung derselben Datenerhebungsmethode in jeder der nationalen Teilstudien ist hervorzuheben. So sind zum Beispiel die Unterschiede zwischen face-to-face Befragungen und telefonischen Befragungen groß.[14] Der verwendete Erhebungsmodus hat unterschiedliche Auswirkungen auf die Zusammensetzung der Stichprobe, auf die Auswahlprozedur und auf den *Coverage* und den *Nonresponse error*. (vgl. Saris, W.E./Kaase, M. 1997b) Zusätzlich wirkt sich der Erhebungsmodus auf die Möglichkeiten der Frageformulierungen, der verwendbaren Hilfsmittel aus und hat verschiedene Antworteffekte zur Folge die im internationalen Vergleich differieren können. (vgl. Skjåk/Harkness 2003) Zusammen mit der unterschiedlichen Interviewsituation und dem Umstand, dass auf den Befragten unterschiedliche Stimuli einwirken, kann die Entscheidung für einen der beiden Erhebungsmodi zu unterschiedlichen methodischen Verzerrungen und somit zu

[14] Für eine genauere Diskussion hierzu vergleiche Saris/Kaase 1997a. Dieser Sammelband behandelt ausschließlich die methodischen Effekte beziehungsweise Unterschiede zwischen face-to-face Befragungen und telefonischen Erhebungen.

unterschiedlichen Ergebnissen führen. (vgl. Saris/Kaase 1997b) Diese beiden in der Praxis üblicherweise verwendeten Erhebungmodi haben zwar jeweils ihre Vor- und Nachteile, allerdings erscheint eine face-to-face Befragung für die international vergleichende Umfrageforschung geeigneter zu sein. Denn der Anteil der Personen mit einem fixen Telefonanschluss schwankt zwischen den einzelnen Ländern relativ stark, was die Äquivalenz der Stichproben beeinträchtigt. (vgl. ebd.) In der internationalen Umfrageforschung können aber wie bereits erwähnt trotz einer identischen Methode länderspezifische Effekte auftreten, die die Messung verzerren würden. Eine andere Datenerhebungsmethode, die lediglich kurz erwähnt sein soll, weil sie gegenwärtig nicht mehr so häufig wie die anderen 2 erwähnten Methoden in Anspruch genommen wird, ist die schriftliche Befragung, die auch postalische Befragung genannt wird. Diese Unpopularität dürfte unter anderem auf die niedrigeren Ausschöpfungsraten und der mangelnden Kontrolle der Interviewsituation im Vergleich zu den anderen beiden Erhebungsmethoden zurückzuführen sein. (vgl. Schnell/Hill/Esser 2008:358-360) Allerdings haben verschiedene Untersuchungen gezeigt, dass gerade bei sensiblen Themen, die soziale Erwünschtheit bei dieser Datenerhebungsmethode auf Grund der größeren Privatsphäre bei der Beantwortung der Fragen geringer ist. (vgl. Johnson/Van de Vijver 2003) Es wurden allerdings auch hierbei kulturelle Unterschiede entdeckt.[15]

Respondent bias

Respondent bias betreffen Eigenschaften des Befragten, die zu Antwortverzerrungen führen. Die Akquieszenz oder Zustimmungstendenz, die soziale Erwünschtheit und die Meinungslosigkeit stellen hierbei die wichtigsten Typen der *Respondent bias* dar. Akquieszenz bezeichnet Zustimmung auf eine Frage ohne Bezug auf deren Inhalt. Sie kann als situational aktivierte Bewältigungsstrategie für unklar definierte Situationen von Personen mit geringer Ich-Stärke und von unterprivilegierten Personen aufgefasst werden. (vgl. Schnell/Hill/Esser 2008:355) Verschiedene Untersuchungen lassen auf kulturelle Unterschiede bezüglich des Auftretens von Akquieszenz schließen (vgl. Smith 2003)[16] Das Vorliegen von *Non-Attitudes* bezeichnet schließlich die Abgabe einer Antwort, auch wenn keine Meinung oder Wissen bezüglich einer Thematik vorliegt. Dies kann im Zusammenhang mit der Kosten-Nutzen-Perspektive betrachtet

[15] Bezüglich Online-Erhebungen wurden keine derartigen Erkenntnisse gefunden.
[16] So wurde zum Beispiel festgestellt, dass griechische Befragte häufiger zu Akquieszenz neigen als Befragte in anderen europäischen Ländern (vgl. Van Herk 2000)

werden und als eine Strategie zur Vermeidung von Kosten in Folge eines Eingeständnisses von Nichtwissen betrachtet werden. (vgl. Diekmann 2004:385-389) Eine weitere mögliche Antwortverzerrung stellt die Weigerung von Befragen dar, auf bestimmte Fragen eine Antwort zu geben beziehungsweise die Angabe, dass sie keine Antwort wissen, also meinungslos sind. Ein Vorliegen von *Item-Nonresponse* könnte dann einen systematischen Fehler darstellen, wenn die Verweigerung einer Antwort mit einer speziellen Ausprägung bezüglich der zu beantwortenden Thematik zusammenhängen sollte. Das Auftreten von *Item-Nonresponse* beziehungsweise deren mögliche indirekte Erscheinungsform über die Auswahl von neutralen oder „weiß-nicht"-Kategorien ist insbesonders in Verbindung mit sensiblen Themen zu erwarten. (vgl. Skjåk/Harkness 2003) Es herrscht allerdings kein Konsens darüber, ob es hinsichtlich der Vermeidung von *Item-Nonresponse* beziehungsweise der *Non-Attitude*-Problematik besser ist, bei Antwortkategorien beziehungsweise Antwortskalen „weiß-nicht"-Kategorien explizit auswählen zu lassen und bei Skalen eine neutrale Mittelkategorie anzubieten. (vgl. Smith 2003) Länderspezifische Unterschiede bezüglich des Vorliegens von Item-Nonresponse und der Inanspruchnahme von „weiß-nicht"-Kategorien und neutralen Mittelkategorien wurden bisher nicht festgestellt. Länderspezifische Unterschiede sind allerdings auch nicht auszuschließen, da einerseits Thematiken in den einzelnen Ländern eine variierende Sensitivität aufweisen und andererseits kulturelle Unterschiede bezüglich der sozialen Akzeptanz der Nichtbeantwortung einer Frage existieren. (vgl. Skjåk/Harkness 2003)

Es mag vielleicht überraschen, dass bei allen 4 Typen an Messfehlern die soziale Erwünschtheit dargestellt wurde. Dies ist unter anderem damit zu begründen, dass in der gängigen Literatur kein Konsens darüber existiert, was soziale Erwünschtheit letztlich ist und was ihr Auftreten begünstigt. Früher wurde davon ausgegangen, dass soziale Erwünschtheit eine Antworttendenz darstellt, die besonders bei sensiblen Fragen auftritt. Der Grund hierfür stellt eine Kosten- Nutzen- Abwägung zwischen dem Wunsch nach sozialer Anerkennung und der Vermeidung von Missbilligung dar, die die Befragten bei einer starken Differenz zwischen dem eigenen Wert und dem von ihnen wahrgenommenen Ort der sozialen Erwünschtheit dazu veranlasst, ihre Antwort in dessen Richtung zu editieren.(vgl. Diekmann 2004: 282-283) Neuere Literatur zu diesem Thema geht eher davon aus, dass soziale Erwünschtheit ein Persönlichkeitsmerkmal ist. So wurde festgestellt, dass die Wahrscheinlichkeit des

Auftretens von sozialer Erwünschtheit bei Minderheiten eines Landes, die für gewöhnlich über eine geringe soziale Macht verfügen und bei Personen mit einem Hang zu Konformität überdurchschnittlich hoch ist. (vgl. Johnson/Van de Vijver 2003)[17] Es erscheint letztlich so zu sein, dass soziale Erwünschtheit eine Kombination aus den beiden erwähnten Ansätzen ist. Sie dürfte ein Persönlichkeitsmerkmal sein, deren Wahrscheinlichkeit des Auftretens von verschiedenen Charakteristiken der besonderen sozialen Situation der Befragung wie von sensiblen Fragen, von Merkmalen und Verhalten des Interviewers und vom Erhebungsmodus beeinflusst werden kann.[18] In unterschiedlichen kulturellen Kontexten existieren auf Grund unterschiedlicher Normen unterschiedliche Auffassungen was akzeptable Verhaltensweisen und Einstellungen sind. (vgl. Johnson/Van de Vijver 2003) Zudem existiert eine unterschiedliche Motivation den eigenen eventuell hiervon abweichenden Wert auch preiszugeben. (vgl. ebd.)

3.1.2. Verzerrung

„Bias refers to the presence of nuisance factors that challenge the comparability of measurement across cultural groups." (Harkness/Vijver/Mohler 2003a:13) Von Verzerrungen wird also gesprochen, wenn Störfaktoren auftreten, die die Vergleichbarkeit der Messung über verschiedene kulturelle Gruppen beeinträchtigen. Das Konzept der Verzerrung steht in einem engen Zusammenhang zum Konzept der Validität. (vgl. Van de Vijver 1998)

> „An instrument is biased if its scores do not have the same psychological meaning across the cultural groups involved; more precisely, an instrument is biased if statements about (similarities and differences of) its scores do not apply in the psychological domain of the scores." (Van de Vijver 1998:43)

Ein Auftreten von Verzerrungen führt letztlich dazu, dass gefundene Unterschiede und Ähnlichkeiten zwischen zwei Gruppen messbedingt sind, also Artefakte darstellen. Van de Vijver unterscheidet zwischen folgenden drei Typen von Verzerrungen, die verschiedene Ebenen des noch vorzustellenden Konzepts der Äquivalenz betreffen und

[17] Die Untersuchungen hierzu beschränken sich zumeist auf die U.S.A. und verwenden für die Messung der sozialen Erwünschtheit zumeist die Marlowe-Crowne-Skala. So wurde zum Beispiel festgestellt, dass Amerikaner afrikanischer und mexikanischer Herkunft häufiger sozial erwünschte Antworten abgeben als die weisse Bevölkerung.(vgl. Ross/Mirowsky 1984, Warnecke u.a. 1997)

[18] So wird in der Literatur auch zwischen kultureller sozialer Erwünschtheit, wie zum Beispiel institutionalisierten Rollenerwartungen und situationaler sozialer Erwünschtheit, wie zum Beispiel die Reaktion des Befragten auf Interviewermerkmale unterschieden. (vgl. Schnell/Hill/Esser 2008:355)

so eine internationale Vergleichbarkeit der erhobenen Daten erschweren können(vgl. Van de Vijver/Leung 1997**;** Van de Vijver 1998; 2003a):

- *Construct bias*
- *Method bias*: Unterteilung in *Sample bias*, *Administration bias*, *Instrument bias*
- *Item bias*

Construct bias tritt auf, wenn ein gemessenes Konstrukt nicht identisch über verschiedene Gruppen ist, aber in diesen Gruppen mit denselben Indikatoren gemessen wird. (vgl. Van de Vijver 1998; 2003a) Anders ausgedrückt liegt ein *Construct bias* dann vor, wenn bei der Messung eines emischen Konstrukts kulturspezifische Dimensionen nicht erfasst werden. Die Identifikation dieses Typs der Verzerrung wird durch ein Vorwissen über die Beschaffenheit des Konstrukts über verschiedene Länder hinweg erleichtert. Dieses kann zum Beispiel aus einer vorhergehenden Messung stammen und dient dazu, das entsprechende Konstrukt um weitere Indikatoren aus bislang nicht beachteten Dimensionen zu erweitern oder nicht relevante Dimensionen zu entfernen.

Method bias bezeichnet die Quellen der Verzerrungen, die durch methodologische Aspekte einer Studie bestehen. (vgl. Van de Vijver 1998; 2003a) Diese können weiter in *Sample bias*, *Administration bias* und *Instrument bias* unterteilt werden. Die Auswirkungen und Variationen der Verzerrungen durch methodische Aspekte wurden bereits ausführlich in der Fehler-Typologie beschrieben. Die *Sample bias* umfassen Verzerrungen auf Grund der Nichtvergleichbarkeit von Stichproben. Die *Administration bias* behandeln Verzerrungen auf Grund von diversen Kommunikationsproblemen zwischen Interviewer und Befragten, oder Reaktionen auf Interviewermerkmale oder diverser Administrationsmodi, wie zum Beispiel Intervieweranweisungen und Erhebungsmodus. Die *Instrument bias* enthalten Verzerrungen der erhaltenen Ergebnisse durch das Instrument. Verzerrungen durch das Erhebungsinstrument entstehen, wenn Individuen systematisch unterschiedlich auf Charakteristiken eines Instruments reagieren. Das Instrument kann durch eine zwischen verschiedenen Gruppen unterschiedliche Vertrautheit mit Umfragen und den dargebotenen Stimuli und von unterschiedlichen Antwortverzerrungen, wie zum Beispiel der sozialen Erwünschtheit, oder der Akquieszenz eine Verzerrung der Ergebnisse verursachen.

Item bias bezeichnet Anomalien auf Ebene der Items. (vgl. Van de Vijver 1998; 2003a) Eine andere verbreitete Bezeichnung für diesen Verzerrungstypus ist *differential item functioning*. (vgl. ebd.) Ein *Item bias* liegt dann vor, wenn Befragte, die eigentlich realiter dieselbe Ausprägung bezüglich eines latenten Merkmals besitzen, einen anderen Skalenwert bei einem Item aufweisen, das eine relevante Dimension dieses Merkmals misst. Als Ausprägung eines Befragten bezüglich eines Konstrukts, wird zumeist der absolute Testwert betrachtet. Dieser entspricht der Summe der Werte, die ein Befragter auf alle Items erzielt, die die interessierenden Dimension(en) messen. Eine Ursache für das Auftreten von *Item bias* kann die Übersetzung der Items in eine andere Sprache darstellen. So können Items für Befragte mit verschiedenen Sprachen eine unterschiedliche semantische Bedeutung besitzen und demzufolge ein anderes Antwortverhalten zur Folge haben. Als eine weitere Ursache, dass Items eine gruppenspezifische Funktionalität aufweisen, können allerdings auch pragmatische Unterschiede, also Kontexteffekte identifiziert werden. Semantische und pragmatische Unterschiede führen letztlich dazu, dass Befragte aus verschiedenen Gruppen unterschiedlichen Stimuli, oder anders ausgedrückt Items mit einer unterschiedlichen Funktionalität, ausgesetzt sind. Die Verzerrungen der Antworten auf Items können nach dem Grad ihrer Verzerrung weiter in *uniform bias* und *non uniform bias* unterteilt werden. (vgl. Braun 2000; Van de Vijver 1998) Ein *uniform bias* liegt dann vor, wenn die Schwierigkeit der Items in den einzelnen Ländern unterschiedlich ist. Dies hat zur Folge dass gewissermaßen eine Verschiebung des Wertebereiches einer Skala stattfindet, aber alle deren Werte in vergleichbarer Weise betroffen sind. Dies hat im Falle einer Ratioskala eine Auswirkung auf den Ursprung beziehungsweise Nullpunkt einer Skala. Ein Beispiel für eine Ursache von *uniform bias* wären gruppenspezifische Differenzen bezüglich des Vorliegens von sozialer Erwünschtheit. Ein *non uniform bias* liegt vor, wenn ein Item in verschiedenen Gruppen eine unterschiedliche Bedeutung hat. Dies hat unterschiedliche Auswirkungen auf den gesamten Wertebereich. Somit wäre die Identität der Skalenwerte unterschiedlich.

Das Konzept der Verzerrung stammt zwar aus der vergleichenden Psychologie, wenn aber die Einteilung der drei verschiedenen Typen von Verzerrungen betrachtet wird, fällt auf, dass diese mit der zuvor beschriebenen Einteilung der Fehler korrespondiert. Es werden alle Typen der Verzerrung in der Fehler-Typologie unter *Measurement error* zusammengefasst. (vgl. Braun 2003b) So fällt auf, dass *Sample, Coverage* sowie

Nonresponse errors in dieser Typologie unter *Sample bias* fallen. Dafür existieren in der vorliegenden Typologie mehr Unterscheidungen bezüglich der durch das Instrument bedingten Verzerrungen als in der Typologie von Groves. Es wird zwischen Verzerrungen auf Ebene der Items und auf Ebene des Erhebungsinstruments unterschieden. Eine Form von *Construct bias* fehlt zum Beispiel in der Fehler-Typologie vollständig.

3.1.3. Das Konzept der Äquivalenz in der international vergleichenden Umfrageforschung

Die Existenz von Fehlern beziehungsweise Verzerrungen hat einen Einfluss auf die Vergleichbarkeit von Messwerten über verschiedene Gruppen hinweg. Der Grad der Vergleichbarkeit von Messwerten wird über das Konzept der Äquivalenz ausgedrückt. Die Reliabilität und somit auch die Validität einer jeden quantitativen Sozialforschung und nicht nur der internationalen Umfrageforschung beruhen auf dem Konzept der Äquivalenz. (vgl. Jowell/Kaase/Fitzgerald/Gillian 2007) Die Äquivalenz ist in verschiedenen Phasen des Forschungsprozesses gefährdet. Dies fängt schon bei der Formulierung der Forschungsfrage an, und reicht bis zur Kodierung der gegebenen Antworten. Dies könnte eine Ursache dafür sein, dass keine konsensuale Definition von Äquivalenz existiert, sondern vielmehr zahlreiche unterschiedliche Definitionen und Variationen zwischen diesen bestehen.[19] Es werden einige dieser Definitionen des Konzepts der Äquivalenz dargestellt, um schließlich zu einer Definition zu gelangen, die dem Erkenntnisinteresse dieser Arbeit entspricht.

„The measurement implications for comparability are addressed in the concept of equivalence. Equivalence refers to the comparability obtained in different cultural groups“. (Harkness/Vijver/Mohler 2003a:14)

„Broadly speaking, we refer to the equivalence of two phenomena if they have the same value, importance, use, function, or result. The important aspect of this concept is the restriction of similarity to one or more specifically defined properties.“ (Van Deth 2003:302)

[19] Eine Auflistung von über 50 verschiedenen Begriffen für die zwei grundlegenden Arten von Äquivalenz existiert bei Johnson 1998.

Unter den verschiedenen Verwendungsweisen des Konzepts der Äquivalenz lassen sich zwei Grunddimensionen erkennen, die diesen zugrunde liegen. Es handelt sich dabei um *interpretive* und *procedural equivalence*. (vgl. Johnson 1998) Interpretative Äquivalenz beschreibt die Ähnlichkeit der Interpretation von abstrakten latenten Konstrukten oder Konzepten in verschiedenen kulturellen beziehungsweise nationalen Gruppen. Sie kann dann als vorhanden betrachtet werden, wenn Konstrukte und theoretische Konzepte über verschiedene Gruppen hinweg eine äquivalente Bedeutung besitzen. Prozedurale Äquivalenz beschreibt hingegen Maßeinheiten und Messprozeduren, die notwendig erscheinen, um eine Vergleichbarkeit über verschiedene Gruppen hinweg zu ermöglichen. Dies beinhaltet die Vermeidung methodischer Verzerrungen auf diversen Ebenen, wie zum Beispiel auf sprachlicher oder administrativer Ebene[20]. Die Folgen der Absenz einer interpretativen oder prozeduralen Äquivalenz werden in folgendem Zitat formuliert.

> „Equivalence refers to the question whether there is any difference in measurement level of within- and between-group comparisons. If the measure is biased against some cultural group, individual differences within a cultural population and across cultural populations are not measured at the same scale." (Van de Vijver 1998:43)

Die Konzepte der Verzerrung und der Äquivalenz sind eng miteinander verknüpft, sie werden sogar häufig als Spiegelkonzepte angesehen. (vgl. Van de Vijver 2003a) So könnte also eine mögliche Definition von Äquivalenz lauten, dass eine vollständige Äquivalenz dann vorhanden wäre, wenn keine Verzerrungen existieren. Theoretisch wäre es auch möglich äquivalente Messergebnisse zu erhalten, wenn zwar Verzerrungen existieren, diese aber über verschiedene Gruppen identische Ausmaße hätten.

Es wird nun eine systematische Klassifikation der verschiedenen Arten der Äquivalenz dargestellt, die mit der Typologie der Verzerrung besser als die vorher dargestellte Einteilung der Äquivalenz in zwei Typen kombinierbar ist. In dieser systematischen Klassifikation der hierarchisch geordneten Typen der Äquivalenz sind die beiden grundlegenden Erscheinungsformen der Äquivalenz enthalten. Hierarchisch in diesem Sinne bedeutet, dass der zweite Typ der Äquivalenz die Existenz des ersten Typs, der dritte Typ die Existenz des zweiten Typs und der vierte Typ die Existenz des dritten

[20] Unter Verzerrungen auf administrativer Ebene werden in dieser Arbeit Verzerrungen auf Grund von diversen Kommunikationsproblemen zwischen Interviewer und Befragten, oder Reaktionen der Befragten auf Interviewermerkmale oder diverser Administrationsmodi, wie zum Beispiel Intervieweranweisungen und Erhebungsmodus verstanden.

Typs voraussetzen. (vgl. Van de Vijver 2003a) Höhere Typen der Äquivalenz sind anfälliger gegenüber Verzerrungen und demzufolge schwieriger herzustellen. (vgl. ebd.) Van de Vijver zufolge kann zwischen folgenden Typen von Äquivalenz unterschieden werden (vgl. Harkness/Vijver/Mohler 2003a; Van de Vijver/Leung 1997**;** Van de Vijver 1998; 2003a:):

- *Construct inequivalence*
- *Functional* beziehungsweise *stuctural* beziehungsweise *conceptual equivalence*
- *Measurement unit equivalence*
- *Scalar* oder *full score equivalence*

Construct inequivalence liegt dann vor, wenn ein gemessenes Konstrukt nicht identisch über verschiedene Gruppen ist und ein *Construct bias* vorliegt. (vgl. Van de Vijver 1998, 2003a) Sollte dies zutreffen, machen Vergleiche keinen Sinn, da sie einer jeden gemeinsamen Grundlage entbehren. Es ist fraglich, ob dieser Typus der Äquivalenz wirklich eine gesonderte Erwähnung finden sollte, da er letztlich lediglich die Negation der folgenden Erscheinungsformen der Äquivalenz darstellt.

Functional oder *Structural* oder *Conceptual equivalence* sind drei verschiedene Ausdrücke für dieselbe Erscheinungsform der Äquivalenz. Im weiteren Verlauf dieser Arbeit wird der Term funktionale Äquivalenz verwendet werden. Ein Erhebungsinstrument, das in verschiedenen kulturellen Gruppen angewendet wird, weist dann eine funktionale Äquivalenz auf, wenn das Konzept, das in jeder dieser Gruppen mittels operationalisierter Konstrukte gemessen werden soll eine etische beziehungsweise zumindest vergleichbare Bedeutung hat, also dieselben Dimensionen beinhaltet.(vgl.ebd.) Dieser Typ der Äquivalenz enthält Merkmale von sowohl der zuvor besprochenen interpretativen als auch der prozeduralen Äquivalenz. Werden mit einem Instrument mit denselben Indikatoren theoretische Konzepte über verschiedene Kulturen hinweg gemessen, die nicht identisch sind, liegt *Construct inequivalence* vor und es sind keine sinnvollen Vergleiche möglich. Sollte jedoch eine funktionale Äquivalenz vorhanden sein, ist die Mindestvorrausetzung für einen internationalen Vergleich erfüllt und es dürfen Korrelationskoeffizienten und Faktorenstrukturen verglichen werden. (vgl. Van de Vijver/Leung 1997)

Measurement unit equivalence ist dann vorhanden, wenn die Skalen eines Instruments zwar einen anderen Nullpunkt, aber dieselben Einheiten der Messung, aufweisen. (vgl. Van de Vijver 1998, 2003a) Dieser Typ der Äquivalenz erfordert mindestens ein Intervallskalenniveau und entspricht wie auch der nachfolgende Typ der *Scalar equivalence* der prozeduralen Erscheinungsform der Äquivalenz. Das Vorliegen von *Method bias* und *nonuniform Item bias* können die Erreichung einer *Measurement unit equivalence* verhindern. (vgl. ebd.) Hat zum Beispiel ein Item in verschiedenen Gruppen eine unterschiedliche Bedeutung, wäre die Identität der Skalenwerte unterschiedlich und es liegt keine *Measurement unit equivalence* vor. Das Vorhandensein von *uniform Item bias* ist hingegen kein Hindernisgrund für die Existenz einer *Measurement unit equivalence*. Ein Beispiel hierfür wäre die Verwendung einer Skala mit denselben Maßeinheiten über verschiedene kulturelle Gruppen hinweg, wobei das erhaltene Ergebnis durch einen unterschiedlichen Einfluss von sozialer Erwünschtheit in den einzelnen Gruppen verzerrt ist. Dies hat zur Folge dass gewissermaßen eine Verschiebung des Wertebereiches einer Skala stattfindet, aber alle Werte der Items in vergleichbarer Weise betroffen sind. Per Definition wäre in einem solchen Fall eine *Measurement unit equivalence* vorhanden. Ein Mittelwertvergleich zwischen verschiedenen Ländern bezüglich dieser Items dürfte streng statistisch betrachtet nicht durchgeführt werden, da eventuelle signifikante Unterschiede im Datensatz auf Grund der verschiedenen Nullpunkte der einzelnen Skalen entstanden sein könnten und nicht auf Grund wirklich existierender Unterschiede. Es müsste also eine statistische Kontrolle über etwaige Drittvariablen vorgenommen werden, um die Ergebnisse vergleichbar zu gestalten, oder anders formuliert, um aus einem etwaigen methodologischen Artefakt die „wahren“ interkulturellen Unterschiede herauszufiltern. (vgl. Van de Vijver 1998, 2003a)

Scalar oder *full score equivalence* existiert dann, wenn Skalen zusätzlich zu identischen Maßeinheiten auch denselben Nullpunkt über die Länder hinweg aufweisen. (vgl. ebd.)[21] Nur wenn eine *Scalar equivalence* vorhanden ist, können direkte Vergleiche durchgeführt und festgestellt werden ob Mittelwerte zwischen verschiedenen Nationen tatsächlich gleich sind oder Unterschiede aufweisen. Diese höchste Form der Äquivalenz kann nur erreicht werden, wenn keine Verzerrungen der Messergebnisse existieren. Hierdurch entsteht ein Problem. Denn gemäß dem Falsifikationsprinzip

[21] Um es anders zu formulieren: Wenn Befragte aus verschiedenen Ländern bei gleicher Einstellungsstärke denselben Skalenwert aufweisen, liegt *Scalar equivalence* vor.

Poppers ist die erreichte Ebene der Äquivalenz zwar „leicht“ widerlegbar, aber nur sehr schwer beweisbar. So ist es oftmals nicht möglich zu beweisen, dass es sich tatsächlich, um *Scalar equivalence* und nicht um *Measurement unit equivalence* handelt. (vgl. Van de Vijver 2003a) Es müssen nämlich alle erdenkbaren Arten von Verzerrungen ausgeschlossen werden, bevor *Scalar equivalence* vorläufig angenommen werden kann. Dies ist durch den Umstand, dass zumeist eine hohe Anzahl möglicher Drittvariablen existiert, die sich mit zunehmender Anzahl und Heterogenität der Länder, die in eine Analyse aufgenommen werden vervielfacht, zumeist kein leichtes Unterfangen.[22]

Generell kann festgestellt werden, dass es bei dem Konzept der Äquivalenz, genauso wie beim Konzept der Verzerrung, hauptsächlich darum geht, die Problematik der Anwendung eines Erhebungsinstruments in zwei oder mehr Kulturen in den Mittelpunkt zu stellen. Verzerrung und Äquivalenz sind also keine inhärenten Charakteristiken eines Erhebungsinstruments, sondern entstehen erst in dessen Anwendung. (vgl. Van de Vijver 2003a) Das Ziel einer jeden international vergleichenden Umfragestudie sollte die Erreichung einer möglichst hohen Ebene der Äquivalenz sein. Dies ist jedoch kein leichtes und vor allem ein kostspieliges Unterfangen, da wie in den beiden vorhergehenden Kapiteln dargestellt wurde, eine Vielzahl an Fehlerquellen existieren, die die Ergebnisse eines internationalen Vergleichs verzerren können.

Es erscheint noch wichtig zu erwähnen, dass die Voraussetzung für die Erreichung einer Äquivalenz auf Messebene keineswegs die Implementierung identischer Erhebungsinstrumente beziehungsweise Prozeduren darstellt. Es darf nämlich nicht vergessen werden, dass selbst die Effekte einer identischen Methode zwischen den einzelnen Ländern durch die Existenz von Kontexteffekten variieren können. So können Unterschiede bezüglich des Ausmaßes diverser Verzerrungen der nationalen Teilergebnisse einer Studie eine internationale Vergleichbarkeit verhindern. (vgl. Saris 1998) Eine identische Methode stellt also nicht notwendigerweise eine äquivalente Methode dar. Genereller ausgedrückt, kann gesagt werden, dass selbst eine noch so große Ähnlichkeit von verwendeten Prozeduren die Vergleichbarkeit der Daten nicht garantieren kann. So kann zum Beispiel ein Konstrukt in verschiedenen Ländern mittels unterschiedlicher Indikatoren gemessen werden und trotzdem oder vielleicht sogar gerade deswegen funktional äquivalent sein. Ein Problem hierbei stellt der Umstand dar,

[22] In dieser Diplomarbeit kann der Beweis der Existenz einer *Scalar equivalence* nicht erbracht werden, da relevante Drittvariablen nicht im Fragebogen enthalten sind.

dass eine äquivalente Methode erst rückwirkend nach der Durchführung einer Studie als eine solche vorläufig bestätigt werden kann. Deswegen erscheint eine genaue Dokumentation des Forschungsvorgangs als unentbehrlich, um letztlich die Quellen von Verzerrungen und der dadurch mangelnden Äquivalenz zu identifizieren und womöglich zu neutralisieren.[23] Eine Äquivalenz von Messprozeduren und Ergebnissen darf also weder ohne weiteres angenommen werden, noch kann deren Existenz im Rahmen der Studiendurchführung sichergestellt werden. (vgl. Van Deth 2003) Vielmehr muss der hergestellte Typ der Äquivalenz für einen bestimmten Vergleich nachträglich überprüft werden.[24]

[23] Für eine detailliertere Darstellung bezüglich der Notwendigkeit einer genauen Dokumentation der Forschungsprozesse und den Anforderungen, die diese Dokumentationen erfüllen sollten. (vgl Mohler/Uher 2003; Van Deth 2003)

[24] In dieser Arbeit liegt der Fokus auf der internationalen Vergleichbarkeit von subjektiven Variablen wie zum Beispiel Einstellungen bezüglich eines bestimmten Themas. Es existieren allerdings auch große Unterschiede bezüglich der sozio-ökonomischen und sozio-demografischen Klassifikationsschemata zwischen den einzelnen Ländern. Diese Unterschiede machen internationale Vergleiche zwischen sozio-ökonomischen und sozio-demografischen Variablen und zwischen vermuteten Zusammenhängen bei denen diese in der Soziologie oftmals Erklärungsfaktoren darstellen, häufig unmöglich. (vgl. Braun/Mohler 2003) Verzerrungen wie zum Beispiel *Construct bias*, können bei objektiven Variablen genauso auftreten wie bei subjektiven Variablen und sind demnach auch daraufhin zu kontrollieren. (vgl. ebd.)

4. Der European Social Survey

In dieser Diplomarbeit wird der European Social Survey (ESS), eine periodisch durchgeführte international vergleichende Umfrage, auf seine Güte für den internationalen Vergleich getestet. Es wird also gemäß des eben vorgestellten Konzeptes der Äquivalenz exemplarisch anhand zweier Skalen überprüft, inwieweit Vergleiche der erhaltenen Ergebnisse möglich sind. Der ESS wurde 2002 das erste Mal durchgeführt. Es folgten drei weitere Erhebungswellen 2004, 2006 und 2008. An der ersten Runde des ESS nahmen 22 Länder teil, in der zweiten 26, in der dritten 25 und in der vierten Runde 31 Länder. Der ESS stellt ein Messinstrument dar, das erklären soll, wie soziale Werte, kulturelle Normen und Verhaltensmuster in verschiedenen Populationen Europas verteilt sind, in welcher Weise sich diese innerhalb und zwischen verschiedenen Nationen unterscheiden und in welche Richtung und Geschwindigkeit sie dies tun. (vgl. Jowell/Kaase/Fitzgerald/Gillian 2007) Das heißt, es ist nicht nur die Validität und Reliabilität der Items relevant, sondern auch deren Vergleichbarkeit über Zeit und Raum hinweg. Um interindividuelle Veränderungen von Merkmalen im Zeitvergleich feststellen zu können, existiert beim ESS ein Kernmodul, das alle zwei Jahre ziemlich unverändert abgefragt wird.

Die Entscheidung für den ESS ist unter anderem deswegen gefallen, da der gesamte Forschungsablauf im Vergleich zu den anderen „großen“ regelmäßig durchgeführten internationalen Umfragen am besten dokumentiert ist.[25] Beim ESS existiert praktisch kein Bereich des Forschungsablaufs, für den keine spezifische Vorgaben für die nationalen Forschungsteams gemacht werden und der nicht durch eine zentrale Koordinierungsstelle kontrolliert und dokumentiert wird. (vgl. Lynn/Japec/Lyberg 2006) Neben der zentralen Koordinierungsstelle wurden weitere Arbeitsgruppen mit unterschiedlichen thematischen Schwerpunkten gebildet. Diese Arbeitsgruppen sind in ständigen Kontakt mit der zentralen Koordinierungsstelle und den nationalen

[25] Es existieren auf der Homepage des ESS (http://www.europeansocialsurvey.org/) und in den Datenarchiven des ESS (http://ess.nsd.uib.no/) neben den frei zugänglichen Datensätzen aus den einzelnen Umfragewellen des ESS zahlreiche Dokumentationen der Methodologie, der Forschungsvorgänge und Ergebnisse dieser, die kostenlos heruntergeladen werden können. Zudem wurde ein Sammelband (Jowell/Roberts/Fitzgerald/Gillian 2007) publiziert bei dem die Entwicklung, das Studiendesign und die Durchführung bezüglich der ersten zwei Wellen des ESS beschrieben werden.

Koordinatoren und überprüfen die Einhaltung der Vorgaben und leisten Beistand bei auftauchenden Problemen. (vgl. Jowell/Kaase/Fitzgerald/Gillian 2007) Die Vorgaben für die nationalen Forschungsteams sind im Vergleich zu anderen international vergleichenden Umfragen ziemlich restriktiv. Die Vorgaben umfassen zum Beispiel die Stichprobengröße oder die Beschränkung auf face-to-face Interviews. (vgl. ebd.) Auf diese Weise wird versucht, die Wahrscheinlichkeit des Auftretens von Variationen durch Verzerrungen zwischen den einzelnen Ländern zu minimieren, die auf Grund des Einsatzes von nicht-äquivalenten Methoden entstehen könnten. Ein Novum in der international vergleichenden Umfrageforschung stellt zum Beispiel die Dokumentation von länderspezifischen Ereignissen im Rahmen des ESS dar.[26] Es werden in den meisten teilnehmenden Ländern nationale Pretests durchgeführt und verschiedene Methodenexperimente durchgeführt. (vgl. Saris/Gallhofer 2007a) Zudem wird pro Erhebungswelle eine Pilotstudie in zwei Ländern durchgeführt, bei dem neue, noch nicht erprobte Fragen auf ihre Angemessenheit für einen internationalen Vergleich getestet werden. (vgl. ebd.) Es werden also sehr viele Ressourcen dafür verwendet, um eine bestmögliche Vergleichbarkeit der Daten sicherzustellen.

Es geht in dieser Arbeit nicht nur darum Fehlerquellen aufzudecken und deren Entstehung zu bemängeln. So werden ebenfalls die zum Teil vorbildhaften Eigenschaften des ESS-Projektes hervorgehoben. Denn wäre alleinig ersteres die Intention, wäre gewiss nicht diese internationale Studie herangezogen worden. Letztlich ist aber keine Wissenschaftsdisziplin vor Fehlern gefeit. Dies gilt insbesondere für eine komplexe Wissenschaft, wie sie die Sozialwissenschaften darstellen und hierbei besonders auf der Ebene der international vergleichenden Forschung mit ihren großen heterogenen Untersuchungseinheiten.

[26] Dies stellt einen Fortschritt gegenüber anderen international vergleichenden Umfragestudien dar. Denn nationale Ereignisse können einen Einfluss auf die Einstellungen und Meinungen von Individuen haben, indem sie den kognitiven Informationsverarbeitungsprozess beziehungsweise die kognitiven Verankerung einer zu erfragenden Information länderspezifisch beeinflussen. (vgl. Stoop 2007) So entstehen Unterschiede zwischen Ländern, die zwar einstellungsbedingt sind, aber dennoch Verzerrungen eines Vergleichs darstellen, da diese Unterschiede lediglich temporär sind. Es sind im ESS allerdings nicht für alle teilnehmenden Länder Dokumentationen verfügbar. So fehlt zum Beispiel eine Dokumentation der länderspezifischen Ereignisse für Österreich.

4.1. Begründung der Auswahl der Länder für die Sekundäranalyse

In der vorliegenden Arbeit soll die Äquivalenz von zwei Skalen, die die Einstellung von Migration messen, für einen Vergleich zwischen den Untersuchungseinheiten Deutschland, England und Österreich überprüft werden. Deutschland wurde auf Grund der unterschiedlichen historischen Entwicklung beim ESS für die Stichprobenziehung in Westdeutschland und Ostdeutschland aufgeteilt wurde.[27] Diese Aufteilung von West- und Ostdeutschland wird im Laufe der Analyse beibehalten, um möglichst homogene Untersuchungseinheiten für die Sekundäranalyse zu verwenden. Zudem kann mittels dieser regionalen Trennung eines Nationalstaats, sehr gut aufgezeigt werden, dass Nationalstaaten nicht per se kulturell homogenen Einheiten darstellen. Es existieren also vier Analyseeinheiten, wobei jeweils zumindest zwei der vier Analyseeinheiten gewisse Gemeinsamkeiten bezüglich eines relevanten Merkmals aufweisen. Diese relevanten Charakteristiken der vier Untersuchungseinheiten bezüglich der zu bearbeitenden Thematik stellen die Geschichte der Migration und kulturelle Gemeinsamkeiten beziehungsweise Unterschiede dar.

Die Betrachtung Großbritanniens ist insofern interessant, da der Master-Fragebogen in Englisch verfasst wurde und die anderen Fragebögen aus diesem übersetzt wurden. (vgl. Harkness 2007) Österreich und Deutschland wurden deshalb gewählt, dass selbst die Fragebögen in der eigentlich selben Sprache unterschiedliche Formulierungen aufweisen, wobei noch erörtert werden wird, ob dies dem Prinzip der Äquivalenz entspricht oder nicht. Denn die Absenz einer Übersetzungsnotwendigkeit des Fragebogens durch die Existenz einer gemeinsamen Sprache schließt noch keine semantischen oder pragmatischen Bedeutungsunterschiede zwischen den 3 Analyseeinheiten aus.

Es wird nun die Geschichte der Migration und die gegenwärtige Situation in den einzelnen Ländern miteinander überblicksmäßig verglichen werden.[28] Dieses Vorgehen

[27] Der Begriff Ostdeutschland umfasst in diesem Sinne die Gebiete der ehemaligen DDR, während der Begriff Westdeutschland die Gebiete der ehemaligen BRD umfasst. Für eine Darstellung der zahlreichen Unterschiede zwischen Ost- und Westdeutschland aus sozialwissenschaftlicher Perspektive sei an dieser Stelle auf den Sammelband von Jan van Deth verwiesen. (2004)

[28] Da diese Arbeit hauptsächlich die Möglichkeit eines internationalen Vergleichs behandelt, wird die Migrationsgeschichte der einzelnen Länder nur rudimentär behandelt. Für eine vertiefende Analyse der

erscheint sinnvoll, da anzunehmen ist, dass die Einstellungen und Meinungen zu Migranten neben Faktoren, wie zum Beispiel der wahrgenommenen Ressourcenknappheit eines Landes, vor allem durch die Sichtbarkeit des Phänomens der Immigration in der Öffentlichkeit beeinflusst werden. (vgl. Rosar 2004) Zuerst muss allerdings der Begriff Migration, der in der Soziologie nicht nur räumliche Wanderungen sondern auch soziale Mobilität umfasst, konkretisiert werden. Es existiert nämlich eine Vielzahl an Definitionen was Migration darstellt. Ein Vergleich von Ausländern, Migranten beziehungsweise Individuen mit Migrationshintergrund über mehrere Länder erweist sich auf Grund der Existenz von zum Teil unterschiedlichen Definitionen dieser Begrifflichkeiten in den einzelnen Ländern als nicht unproblematisch.[29] Wenn in dieser Arbeit von Migration gesprochen wird, ist damit eine internationale Migration gemeint, bei der Personen auf Grund diverser Motivationslagen, freiwillig oder unfreiwillig dauerhaft ihren Wohnort in ein anderes Land verlegen. Die folgende Definition fasst diese Bedeutungsebene relativ gut. "Migration ist der auf Dauer angelegte bzw. dauerhaft werdende Wechsel in eine andere Gesellschaft bzw. in eine andere Region von einzelnen oder mehreren Menschen." (Treibel 1990:21)

Nach dem Zweiten Weltkrieg waren Deutschland und Österreich ethnisch beinahe homogene Länder. (vgl. Weiss/Reinprecht 2004:46) Westdeutschland und Österreich weisen zudem eine nach 1945 vergleichbare Migrationsgeschichte auf. Diese ist hauptsächlich durch die Anwerbung von Gastarbeitern in den 60er und frühen 70er Jahren, sowie durch kurdische Migranten aus der Türkei und Flüchtlingen auf Grund des Balkan-Konflikts in den 90er Jahren gekennzeichnet. Davor kann man diese Länder als klassische Auswanderungsländer bezeichnen.

Etwas anders gestaltete sich die Situation in Ostdeutschland. So existierte zwar auch in der ehemaligen DDR auf Grund der zahlreichen Emigranten, die vor allem nach Westdeutschland auswanderten, eine Arbeitskräfteknappheit und die Notwendigkeit für

---

Entwicklung der Migration und die Entwicklung der diesbezüglichen rechtlichen Rahmenbedingungen in den betrachteten Ländern sei auf Edda Currle verwiesen. (2004)

[29] Im Fragebogen des ESS wurde die Problematik des Migrationsbegriffes umgangen, da im Einleitungstext klargestellt wird, dass es sich beim interessierenden Sachverhalt um internationale Migration handelt („Menschen, die aus anderen Ländern ins Land kommen, um hier zu leben"). (vgl. Kap. 5.2) In weiterer Folge werden Migranten dann als Zuwanderer bezeichnet. Es wird der in den verschiedenen Ländern möglicherweise unterschiedlich konnotierte Begriff Migration beziehungsweise Migranten bewusst vermieden.

Arbeitsmigranten anzuwerben. Der Migrationsstrom hatte aber einen geringeren Umfang als in Österreich beziehungsweise Deutschland. Zudem lebten die Arbeitsmigranten, die aus den anderen sozialistischen Nationalstaaten angeworben wurden und die Bevölkerung separiert voneinander. (vgl. Reißlandt 2005)[30] Ab 1990 existierte in Ostdeutschland ein mit Westdeutschland vergleichbarer Migrationsstrom. (vgl. Rosar 2004) Hinzu kommt, dass die Gebiete Ostdeutschlands nach der deutschen Wiedervereinigung einem beschleunigten gesellschaftlichen Wandel unterworfen waren. Dies wird in der soziologischen Literatur auch als Transformation bezeichnet und kann die Einnahme einer negativen Haltung bezüglich Migranten begünstigen. (vgl. Rosar 2004)

Im Gegensatz zu den bereits betrachteten Ländern ist die Entwicklung der Immigration in Großbritannien durch die Kolonialzeit geprägt. Dies hat zur Folge, dass die Migration nach dem 2. Weltkrieg früher einsetzte als in den anderen betrachteten Ländern und die Migranten eben vorwiegend aus dem Commonwealth of Nations beziehungsweise aus Irland stammen.(vgl. Currle 2004)

Nun sollen noch die Anteile der Bevölkerung mit einer ausländischen Staatszugehörigkeit, gemessen an den Gesamtbevölkerungen, in den einzelnen Ländern verglichen werden. So betrug der Anteil der Bevölkerung mit einer ausländischen Staatszugehörigkeit an der Gesamtbevölkerung im Jahr 2004 in Deutschland 8,9 %, in Österreich 9,4 % und in Großbritannien 4,7%. (vgl. Eurostat 2006)[31] In Deutschland existieren diesbezüglich große Unterschiede zwischen Ost- und Westdeutschland. So hat der Ausländeranteil 2006 in Ostdeutschland lediglich 4,8 % betragen, während in Westdeutschland rund 9,6 % der dort lebenden Bevölkerung eine fremde Staatsbürgerschaft besaßen. (vgl. Bundeszentrale für politische Bildung o.J:7-8)[32]

[30] http://www.bpb.de/themen/8Q83M7,0,0,Migration_in_Ost_und_Westdeutschland_von_1955_bis_2004.html

[31] http://www.eds-destatis.de/de/downloads/sif/nk_06_08.pdf

[32] http://www.bpb.de/files/HXVH9I.pdf, Die Prozentzahlen wurden auf Basis der dargestellten Daten errechnet. Ostdeutschland umfasst hierbei den Stadtstaat Berlin. Ohne den Ausländeranteil Berlins, weist Ostdeutschland einen Ausländeranteil von lediglich 2,4 % auf.

4.2. Vorgangsweise bei der Überprüfung der erreichten Ebene der Äquivalenz im Rahmen einer Sekundäranalyse

Es können generell zwei Ansätze unterschieden werden, wie mit Verzerrungen und den Konsequenzen für die erreichte Form der Äquivalenz einer international vergleichenden Umfrage umgegangen wird. (Van de Vijver 2003a) Sogenannte *a priori* Techniken werden vor der eigentlichen Datenerhebung eingesetzt und haben den Zweck die Wahrscheinlichkeit des Auftretens von Verzerrungen zu minimieren. Der Fokus liegt bei diesen Techniken auf der bestmöglichen Entwicklung des Fragebogen- und Stichprobendesigns. Zum Einsatz kommen hierfür zum Beispiel Multitrait-Multimethod- oder Split Ballot Verfahren.[33] Der zweite und für diese Arbeit relevante Ansatz, da eine Sekundäranalyse eines bereits bestehenden Datensatzes vorgenommen wird, stellt der sogenannte *post hoc* Ansatz dar. Dieser beinhaltet die Anwendung diverser statistischer Verfahren, um die mögliche Existenz von Verzerrungen bei einem bereits bestehenden Datensatz nachzuweisen, die Quellen der Verzerrung zu identifizieren und wenn möglich Korrekturen vorzunehmen, um diese Verzerrungen zu neutralisieren. Es müssten beide Techniken angewendet werden, um eine Minimierung der Verzerrungen beziehungsweise eine Maximierung der Äquivalenz der Daten sicherzustellen. Die Anwendung von *a priori* Techniken ist einer kleinen Anzahl von Personen vorbehalten, die diese Studie entwickeln. Nur diese können diverse Techniken zur Qualitätssicherung einsetzen, wie zum Beispiel Pretests und Methodenexperimente. Diese Vorgänge sind im Idealfall dokumentiert und somit nachvollziehbar gestaltet. So können aus den Ergebnissen dieser *a priori* Techniken, insofern diese veröffentlicht werden, eventuell Rückschlüsse auf vorhandene Verzerrungen der Daten gezogen werden.

Es existiert mittlerweile eine Fülle an Literatur, die die Überprüfung der erreichten Ebene der Äquivalenz durch den Einsatz verschiedener statistischer Verfahren im Rahmen eines *post hoc* Ansatzes thematisiert und gewissermaßen eine Vorbildfunktion

[33] Diverse Techniken hierfür werden zum Beispiel im Sammelband von Harkness/Vijver/Mohler 2003b dargestellt, eine Übersicht dieser Techniken findet sich zum Beispiel bei Van de Vijver 1998 und 2003b, Johnson 1998 und Blair/Piccinino 2005. Für einen detaillierten Einblick in den Multitratit-Multimethod Ansatz um Messinstrumente zu evaluieren, sei auf den Sammelband von Saris/Münnich verwiesen. (1995) Für eine Darstellung von Split Ballot Verfahren sowie deren Zweck siehe Fowler. (2004)

für die Vorgehensweise in dieser Arbeit einnehmen.[34] „The best advice to researchers is probably to employ as many of these techniques as possible and within reason, given that various methodologies may be more appropriate to one specific form of equivalence or another." (Johnson 1998: 31) Es werden in dieser Arbeit mehrere statistische Verfahren angewandt, da diese unterschiedliche Vor- und Nachteile bezüglich der Entdeckung von Verzerrungen auf Ebene der Konstrukte und der einzelnen Items aufweisen. Das Ziel besteht hierbei stets in der Feststellung, ob Verzerrungen existieren, die unterschiedliche Auswirkungen in den untersuchten Ländern offenbaren. Die Vorgangsweise bei der Überprüfung der Äquivalenz kann in drei Abschnitte unterteilt werden. Eine explorative Phase, die weiter in eine Durchsicht der Dokumentierung der einzelnen Forschungsvorgänge und eine explorative Datenauswertungsphase unterteilt werden kann. In diesem Teil der Arbeit geht es primär darum, diverse Auffälligkeiten zu entdecken, die Verzerrungen auf der methodischen Ebene und auf der Ebene der Items zu darstellen könnten**.** Der nächste Abschnitt besteht schließlich aus der Feststellung des vorliegenden Ausmaßes der Äquivalenz der zwei Skalen mittels mehrerer statistischer Verfahren. So wird überprüft, ob Verzerrungen vorliegen, die valide Vergleiche der Einstellungen zur Migration und Meinungen zu den Auswirkungen der Migration zwischen Großbritannien, Österreich, Ost- und Westdeutschland erschweren.

Da zur Feststellung der erreichten Ebene der Äquivalenz mehrere statistische Verfahren zur Anwendung kommen, muss eine Problematik bezüglich des Skalenniveaus der zu analysierenden Daten dargestellt werden. Es wurden im ESS fast ausschließlich Likert-ähnliche Skalen verwendet, um latente Einstellungen und Meinungen zu messen. Durch die Verwendung von sogenannten Rating-Skalen werden Daten gewonnen, die zwar intervallskalierten Charakter aufweisen, aber streng statistisch gesehen nur Ordinalskalenniveau aufweisen. Wie mit diesen Daten umzugehen ist und welche Auswirkungen dies auf die Anwendung von statistischen Verfahren hat, die ein Intervallskalenniveaus voraussetzen, ist in den Sozialwissenschaften eine schon lang existierende Streitfrage.[35] Es wäre letztlich ein Nachweis notwendig, dass die

[34] Diverse Ansätze finden sich bei Johnson (1998) Braun/Scott (1998); Braun (2000), Van de Vijver/Leung (1997), Van de Vijver (1998; 2003b), Zucha (2002; 2004; 2005), Rother (2005) und Reeskens/Hooghe(2008).

[35] Dies ist in einem engen Zusammenhang mit der axiomatischen Messtheorie zu betrachten, aus der das Repräsentations- und Eindeutigkeitstheorem abgeleitet werden können und das ebenfalls daraus ableitbare und diskutierte Bedeutsamkeitsproblem behandelt. Eine diesbezügliche Diskussion dieser

verwendeten Rating-Skalen ein Intervallskalenniveau aufweisen, um dieses auch annehmen zu können. Dieser Nachweis wird auf Grund verschiedener Umstände in der Praxis allerdings so gut wie nie erbracht. Dies trifft auch auf die Skalen des ESS zu. Deswegen wird in dieser Arbeit davon ausgegangen, dass es sich bei den mit Rating-Skalen gemessenen Einstellungen und Meinungen, um ordinalskalierte Daten mit intervallskalierten Charakter handelt. Da Skalenitems von Rating-Skalen einen intervallskalierten Charakter aufweisen, werden trotzdem deren Mittelwerte und Varianzen berechnet und auch multivariate Verfahren angewandt, die intervallskalierte Daten voraussetzen. Jedoch werden diese Ergebnisse nicht inhaltlich interpretiert werden, sondern lediglich zur Überprüfung der erreichten Ebene der Äquivalenz eingesetzt werden.

4.2.1. Der explorative Teil der Überprüfung der erreichten Ebene der Äquivalenz

Zunächst wird für die Überprüfung der Äquivalenz der zwei Skalen für einen Vergleich zwischen Deutschland, Großbritannien und Österreich explorativ vorgegangen. Dies dient der Überprüfung bezüglich der Existenz von Verzerrungen, die das Ausmaß der Äquivalenz beeinträchtigen könnten.[36] Dieser Vorgang unterscheidet sich bis auf den Umstand, dass auch über die einzelnen Länder verglichen wird, im Prinzip nur unwesentlich von der üblichen Vorgehensweise in der nationalen Umfrageforschung. Es werden zuerst die Dokumentationsprotokolle des gesamten Studiendesigns des ESS durchgesehen, um eventuelle Unregelmäßigkeiten zu entdecken. Dies betrifft die Kontrolle, ob ein *Administration* oder *Sample bias* vorliegen könnte. (vgl. Kap. 3.1.2) Der erste wichtige Punkt stellt hierbei die Überprüfung des Ausmaßes der Äquivalenz der Stichproben der einzelnen Länder dar. Es werden die Stichprobendesigns, der *Coverage error* und der *Unit-Nonresponse error* der einzelnen Länder miteinander verglichen. Zusätzlich wird überprüft, ob und wie eventuelle Unregelmäßigkeiten durch nachträgliche durchgeführte Korrekturverfahren, wie zum Beispiel eine Gewichtung

---

Axiome und der Problematik des Nachweises der Existenz eines Intervallskalenniveaus in den Sozialwissenschaften und der darausfolgenden Diskussion findet sich bei Schnell/Hill/Esser 2008:140-149 und Diekmann 2004: 255-258.

[36] Im Rahmen dieser Sekundäranalyse werden lediglich Daten verwendet, die im Rahmen der dritten Welle des ESS erhoben wurden. Dementsprechend stellt die Stabilität des verwendeten Messinstruments beziehungsweise die Stabilität der damit gemessenen Einstellungen, Werte, und Verhaltendispositionen im Zeitverlauf keine relevante Fehlerquelle dar. Dies bedeutet allerdings auch, dass es sich bei eventuell vorhandenen Verzerrungen, um ein einmaliges Ereignis handeln kann.

bereinigt werden konnten. Anschließend hierzu werden die zwei zu überprüfenden Skalen, die im ESS zur Einstellungs- und Meinungsmessung zu Migranten verwendet wurden, präsentiert und deren Konstruktionsweise betrachtet. Zudem wird die Vorgangsweise der Übersetzung im ESS und die unterschiedlichen Übersetzungen der Items und Antwortskalen in verschiedene Sprachen betrachtet. Dies dient dazu Hinweise zu entdecken, ob die einzelnen Skalenitems beziehungsweise die Antwortskalen eine vergleichbare semantische und pragmatische Bedeutung in den einzelnen Ländern besitzen.

Nachdem diese Punkte abgehandelt wurden, beginnt die explorative Auswertungsphase mit dem Datensatz des ESS. Diese beginnt mit einer Überprüfung des Datensatzes auf etwaige Diskrepanzen, wie zum Beispiel länderspezifische Unterschiede bezüglich der Kodierung oder unterschiedliche Definitionen der fehlenden Werte der relevanten Skalenitems. Es werden die deskriptiven Statistiken und diverse Verteilungsmaße der einzelnen Skalenitems betrachtet und eine Missing-Value Analyse durchgeführt, um etwaige Verzerrungen auf Ebene der Items zu entdecken. Die Angabe der univariaten Verteilungen hat den Zweck, etwaige relationale Unterschiede bezüglich der anteilsmäßigen Zellenbesetzungen eines Items im Ländervergleich, sowie mit den anderen Items der jeweiligen Skala innerhalb eines Landes zu entdecken. Die Betrachtung der Lage-, Streuungs- und Formmaße der einzelnen Items erfolgen nach demselben Prinzip und erfüllen einen ähnlichen Zweck. Es werden die einzelnen arithmetischen Mittelwerte und Standardabweichungen betrachtet und festgestellt, ob eine Normalverteilung der Daten gegeben ist. Die Betrachtung der Mittelwerte und Standardabweichungen kann erste Hinweise auf mögliche länderspezifische Unterschiede bezüglich der skaleninternen Struktur der Items geben. So wird eine erste Einsicht dahingehend gewonnen, ob die Reihenfolge der Items bezüglich der Höhe der Mittelwerte, welche auch als in standardisierter Form als Itemschwierigkeiten betrachtet werden, in den einzelnen Ländern vergleichbar ist. Zudem können Rückschlüsse bezüglich der Ähnlichkeit der einzelnen Items innerhalb der Skalen gezogen werden, die ja ein eindimensionales Konzept messen sollen. Sollte bei einem einzelnen Item ein höherer Mittelwert als bei den restlichen Items einer Skala vorliegen und dies ist in den anderen Ländern für dieses Item so nicht beobachtbar, könnte dies auf die Existenz eines *Item bias* hinweisen. Falls jedoch alle Items der Skala in einem Land einen höheren Mittelwert aufweisen als in den anderen betrachteten Ländern ist dies noch

nicht per se problematisch, da dies ein Ausdruck eines tatsächlich vorhanden Unterschiedes bezüglich des zu messenden Konstrukts sein könnte. Jedoch ist ein *uniform bias* in dieser Forschungsphase ebenfalls noch nicht auszuschließen.

Die Überprüfung, ob eine Normalverteilung vorhanden ist, ist durch den Umstand notwendig, dass die konfirmatorische Faktorenanalyse, die im weiteren Verlauf noch zur Anwendung kommt, eine Normalverteilung der Items voraussetzt. Eine Normalverteilung ist zwar gemäß des Ordinalskalenniveaus der Daten grundsätzlich nicht möglich. Es wird allerdings geprüft, ob zumindest eine näherungsweise Normalverteilung gegeben ist.

Die Analyse der fehlenden Werte der Items, dient einerseits der Überprüfung, ob gewisse Items eine auffällig hohe Anzahl an fehlenden Werten aufweisen. Die fehlenden Werte könnten durch die Formulierung oder durch eine relativ zu den anderen Skalenitems höheren Schwierigkeitsgrad des jeweiligen Items verursacht sein. Eine hohe Anzahl an fehlenden Werten würde in weiterer Folge bei der Anwendung von multivariaten Verfahren eine Reduktion der Stichprobengröße bedeuten. Denn gewisse Verfahren, wie zum Beispiel die konfirmatorische Faktorenanalyse, nehmen einen listenweisen Fallausschluss bei Vorliegen eines fehlenden Wertes bei den relevanten Items vor. Dies wird jedoch noch bei den betreffenden Verfahren erläutert werden. Viel wichtiger ist allerdings der Ausschluss eines möglichen *Item-Nonresponse errors*. Für diesen Zweck wird mittels Korrelationen überprüft, ob die Verweigerung einer Antwortabgabe mit anderen Persönlichkeitsmerkmalen der Befragten assoziiert ist. Sollte dies der Fall sein, wäre die Validität des betreffenden Items äußerst fragwürdig und hätte auch Auswirkungen auf die Genauigkeit der Ergebnisse von multivariaten Verfahren.

4.2.2. Der Einsatz von korrelationsbasierten Methoden zur Überprüfung der funktionalen Äquivalenz

Schließlich beginnt die Überprüfung, ob und in welchem Umfang eine funktionale Äquivalenz der gemessenen Konstrukte vorhanden ist. Hierfür eignen sich besonders Verfahren, die auf Korrelationen oder Kovarianzen der Items basieren, da diese sich sehr gut zur Überprüfung der Struktur der einzelnen Items zueinander und der zugrundeliegenden Faktoren beziehungsweise der Dimensionen eignen. (vgl. Van de Vijver/Leung 1997, Zucha 2004:65) Dies beinhaltet einfache Korrelationsanalysen, Verfahren zur Überprüfung der internen und externen Konsistenz der Skalen, eine explorative und eine konfirmatorischen Faktorenanalyse. (vgl. Van de Vijver/Leung 1997, Van de Vijver 2003b; Braun 2000, Johnson 1998, Zucha 2004) Die Durchführung von Korrelationsanalysen, von Konsistenzanalysen und von explorativen Faktorenanalysen sind Vorgangsweisen, die auch in der nationalen Umfrageforschung für gewöhnlich zur Bestimmung der Reliabilität und Validität von Skalen zum Einsatz kommen. Für die Festellung, ob sogar ein höheres Ausmaß der Äquivalenz der Skalen in den betrachteten Analyseeinheiten vorliegt, eignen sich nur Verfahren die auf den Kovarianzen der Items basieren. (vgl. Van de Vijver 2003b) Die Überprüfung der beiden Skalen bezüglich der Existenz einer *Measurement unit equivalence* wird in der vorliegenden Arbeit lediglich mittels der Durchführung der konfirmatorischen Faktorenanalyse vorgenommen, da diese auf den Kovarianzen der Items basiert.[37]

Für eine erste Überprüfung der Ähnlichkeit der skaleninternen Items eignet sich die Betrachtung der Interkorrelationen der Items. Dies ist gewissermaßen eine explorative Version der Prüfung der internen Konsistenz einer Skala, wobei noch jedes Item mit jedem anderen einzelnen Item einer Skala korreliert wird. Im Gegensatz beruhen die Konsistenzanalysen und die explorative Faktorenanalyse auf der Korrelation eines einzelnen Items mit den restlichen Items einer Skala. Mit diesen Verfahren kann festgestellt werden, bis zu welchem Grad die einzelnen Items zu einer Skala gehören und die Homogenität der Skala gewährleistet ist. Der Unterschied zur üblichen

[37] Es werden zwar, wie im vorigen Kapitel dargestellt, auch die Mittelwerte der Items zwischen den Analyseeinheiten verglichen um diesbezügliche Auffälligkeiten zu entdecken. Allerdings ist eine statistische Überprüfung der Mittelwertsunterschiede auf Grund des vorliegenden Skalenniveaus der Daten nicht möglich.

Verwendungsweise im Rahmen einer Itemanalyse im nationalen Kontext besteht bei der Überprüfung für die Eignung für einen internationalen Vergleich darin, dass zusätzlich ein Gruppenvergleich durchgeführt wird. Das heißt, es sind zwar teilweise durchaus auch die absoluten Werte, die mittels dieser Verfahren gewonnen werden, von Bedeutung. Es geht allerdings eigentlich vielmehr darum, die relationalen Muster dieser Werte zwischen den einzelnen Ländern zu vergleichen. Die Items wären dann als äquivalent zu betrachten, wenn deren Struktur zueinander zwischen den Analyseeinheiten gleich wäre. (vgl. Braun 2000:3)

Nachdem die Frage geklärt wurde, inwieweit die Skalen bezüglich ihrer internen und externen Konsistenz im Ländervergleich übereinstimmen, wird eine explorative und eine konfirmatorische Faktorenanalyse durchgeführt**.** Es werden nun diese zwei Verfahren detailliert dargestellt, um die Logik dieser Verfahren und die Zweckmäßigkeit für die Überprüfung der erreichten Ebene der Äquivalenz detaillierter zu erläutern.

Explorative Faktorenanalyse

Die Faktorenanalyse basiert auf den Interkorrelationen der Items und zählt zu den Strukturen-entdeckenden Verfahren. Das Ziel einer Faktorenanalyse stellt die Entdeckung latenter Einflussfaktoren dar, die die Existenz von unterschiedlichen Ausprägungen von individuellen Einstellungen, Meinungen und Verhaltensweisen erklären können. Die Faktorenanalyse hat den Zweck aus einer Vielzahl von manifesten Variablen einige wenige wichtige latente Dimensionen herauszufiltern, die den unterschiedlichen Ausprägungen von Individuen bezüglich dieser manifesten Variablen zugrunde liegen. Die Faktorenanalyse dient also dem Zweck der Dimensionsreduktion. Es werden Variablen, die miteinander korrelieren, zu einem Faktor zusammengefasst. Die explorative Faktorenanalyse unterscheidet sich von der herkömmlichen Faktorenanalyse dadurch, dass die Anzahl der zu extrahierenden Faktoren nicht festgelegt ist, sondern auf Grundlage der Interkorrelationen der Items bestimmt wird.

Die Anwendung einer explorativen Faktorenanalyse hat in dieser Arbeit den Zweck zu überprüfen, ob die Skalen eine funktionale Äquivalenz aufweisen. Es wird also geprüft, ob und in welchem Ausmaß eine Vergleichbarkeit der Faktorenstruktur gegeben ist.

Sollte zumindest die Faktorenstruktur der mit den Skalen gemessenen Konstrukte zwischen den vier Analyseeinheiten ähnlich sein, kann davon ausgegangen werden, dass diese Skalen eine funktionale Äquivalenz aufweisen. (vgl. Van de Vijver/Leung 1997) Zuerst gilt es festzustellen, ob die Skalen wirklich jeweils ein eindimensionales Konstrukt messen. Anders formuliert, es wird geprüft, ob lediglich ein Einflussfaktor existiert, der der Beantwortung der Items der jeweiligen Skala zu Grunde liegt. Sollte in einer der betrachteten Analyseeinheiten für eine Skala ein zweiter Einflussfaktor existieren, wäre eine weitere Überprüfung der funktionalen Äquivalenz für diese hinfällig. Für diesen Zweck werden für die vier Analyseeinheiten separate Faktorenanalysen durchgeführt. Zudem wird auch eine gemeinsame Faktorenanalyse der vier Analyseeinheiten durchgeführt, da diese Faktorenlösung den Durchschnitt repräsentiert und sich somit gut für einen Vergleich von etwaigen Abweichungen eignet. In weiterer Folge wird die Faktorenstruktur, repräsentiert durch die Faktorenladungen der einzelnen Items betrachtet, um festzustellen inwieweit diese in den einzelnen Ländern ähnlich ist. Die Höhe der einzelnen Faktorladungen ist zwar nicht gänzlich unwichtig, ist aber für diesen Zweck nicht ausschlaggebend. Die Betrachtung der Relation der Faktorladungen zueinander in den einzelnen Ländern entscheidet das Ausmaß der Ähnlichkeit der Faktorenstrukturen. (vgl. Vijver 2003b) Die Reliabilität der Faktoren, also die Eigenwerte dieser darf sich unterscheiden, aber die Bedeutung des Faktors für die Itemausprägungen im Vergleich zu den anderen Items muss zwischen den Ländern ähnlich sein. (vgl. ebd.) Die Faktorenstrukturen werden zuerst paarweise zwischen den Analyseeinheiten verglichen und schließlich ebenfalls paarweise mit der Faktorenstruktur des Gesamtdatensatzes.

Es muss jedoch angemerkt sein, dass verschiedene Vorrausetzungen existieren, um eine Faktorenanalyse durchführen zu können und verschiedene Gütemaße, die die Qualität der endgültigen Faktorenlösung angeben. So sollten die Daten mindestens intervallskaliert, die Stichprobe ausreichend groß und eine bestimmte Mindestanzahl an Variablen vorhanden sein. (vgl. Backhaus/Erichson/Plinke/Weiber 2006 325-331) Es sollten bei den betreffenden Variablen nicht zu viele fehlende Werte beziehungsweise kein *Item-Nonresponse error* vorliegen. Zudem ist ermöglicht eine geringe Streuung der Antworten zwischen den einzelnen Komponenten genauere Parameterschätzungen. Durch den Umstand, dass diese Faktorenanalyse mit Daten durchgeführt wird die lediglich den Charakter von intervallskalierten Daten aufweisen, werden lediglich die

Faktorenstrukturen betrachtet und diese letztlich nicht inhaltlich interpretiert werden. Es zeigt sich zudem, dass eine Faktorenanalyse in der Praxis auch mit Likert-ähnlichen Skalen robuste Ergebnisse liefert, wenn diese Skalen zumindest vier Ausprägungen aufweisen. (vgl. Kim/Müller 1978:74)

Es existieren mehrere Varianten der Faktorenanalyse, die abweichende Zielsetzungen aufweisen und denen unterschiedliche Annahmen und Methoden zur Faktorenextraktion und Faktoreninterpretation zu Grunde liegen.[38] In dieser Arbeit wird eine Hauptachsenanalyse durchgeführt und demzufolge ausschließlich auf deren Eigenschaften eingegangen. Die Hauptachsenanalyse hat den Zweck, durch Faktoren, die als Beurteilungsdimensionen aufgefasst werden können, die Varianz der Faktoren zu erklären. (vgl. Backhaus/Erichson/Plinke/Weiber 2006:291-295) Das Verfahren der Hauptachsenanalyse eignet sich deshalb für die Dimensionalitätsüberprüfung der Skalen und der Überprüfung der Ähnlichkeit der Faktorenstrukturen, da zwischen Kommunalitäten und Einzelrestvarianzen unterschieden wird und so die Faktoren kausal interpretiert werden. (vgl. ebd.) Ob eine Rotation der anfänglichen Faktorenlösung durchgeführt wird, wird erst nach Kenntnis der Interkorrelationen der Items beziehungsweise bei der Durchführung der eigentlichen Faktorenanalyse unter Berücksichtigung der Güte der anfänglichen Faktorenlösung und der Beschaffenheit der Skalen entschieden. (vgl. Kap. 6)

Konfirmatorische Faktorenanalyse

Es existiert aber letztlich bei der explorativen Faktorenanalyse keine Maßzahl, um die erreichte funktionale Äquivalenz im Sinne einer vergleichbaren Faktorenstruktur und einer vergleichbaren Konstruktvalidität zwischen den einzelnen Analyseeinheiten definitiv feststellen zu können. Deswegen wird als letzter Schritt eine konfirmatorische Faktorenanalyse durchgeführt, in der Gütemaße hierfür berechnet werden können. Diese ist ein Strukturen-prüfendes Verfahren, da deren Zweck die Überprüfung von vermuteten Zusammenhängen darstellt. Die Struktur der Beziehung zwischen den Items und der dem Konstrukt werden also bereits zu Beginn der konfirmatorischen Faktorenanalyse festgelegt. Die konfirmatorische Faktorenanalyse basiert genauso wie die Faktorenanalyse auf Korrelationskoeffizienten oder auf der Kovarianz zwischen

[38] Für die Darstellung der verschiedenen Arten der Faktorenextraktion sei auf Brosius 1998 und Backhaus/Erichson/Plinke/Weiber 2006 verwiesen.

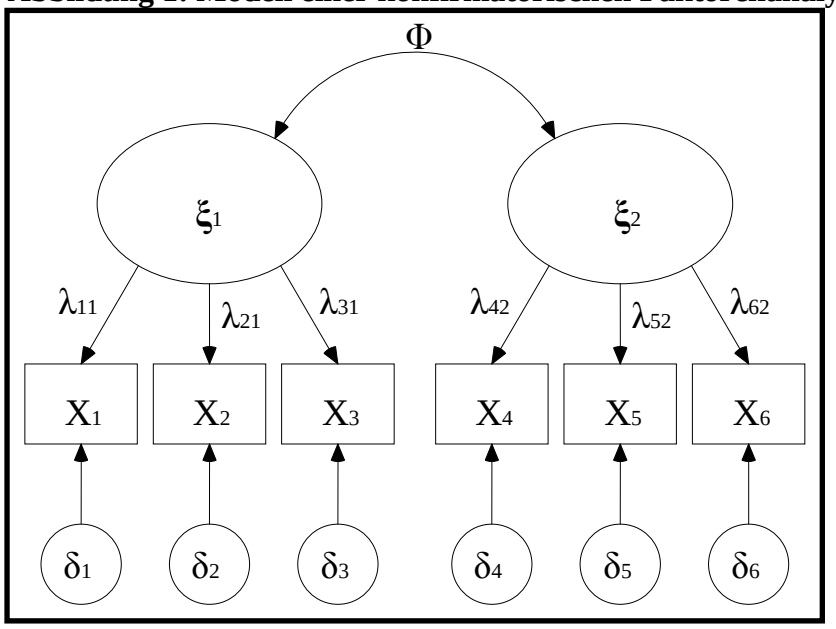
manifesten Variablen. In der vorliegenden Arbeit basiert die konfirmatorischen Faktorenanalyse auf einer Kovarianzmatrix, da nur so die Existenz einer *Measurement unit*- und *Scalar equivalence* der zwei Skalen festgestellt werden kann.

Die Hypothesen, die mittels der Durchführung der konfirmatorischen Faktorenanalyse geprüft werden sollen:

- Das mit den Skalen gemessene theoretische Modell, welches die Einstellung bezüglich Migration beziehungsweise der Meinung der Auswirkungen dieser enthält, stimmt mit den Daten aus den einzelnen Analyseeinheiten überein.
- Der durchgeführte Gruppenvergleich zwischen den Analyseeinheiten zeigt, dass die Skalen funktional äquivalent sind.
- Der durchgeführte Gruppenvergleich zwischen den Analyseeinheiten zeigt, dass auch eine höhere Ebene der Äquivalenz erreicht wurde.

Eine Besonderheit der konfirmatorischen Faktorenanalyse stellt die Möglichkeit dar, die Güte von Struktur- und Messmodellen für verschiedene Gruppen und für Veränderungen über mehrere Messzeitpunkte hinweg, in Form inter- und intraindividueller Veränderungen, festzustellen. (vgl. Reinecke 2005:15-16) Neben dem Umstand, dass bestimmt werden kann, wie zentral einzelne Indikatoren für die Messung eines Konstruktes sind, ist die Möglichkeit der Bestimmung der Güte von Messmodellen für mehrere Populationen, der Grund für die Durchführung der konfirmatorischen Faktorenanalyse. Während mit der explorativen Faktorenanalyse lediglich die Eindimensionalität der Skalen und die augenscheinliche Ähnlichkeit der Faktorenstrukturen zwischen den Analyseeinheiten festgestellt werden können, ermöglicht die Durchführung einer konfirmatorischen Faktorenanalyse eine detailliertere Analyse der erreichten Ebene der Äquivalenz. (vgl. Van de Vijver 2003b) So kann anhand von Gütemaßen nicht nur festgestellt werden, ob und bis zu welchem Grad die Modelle auf Mess- und Strukturebene zwischen den Analyseeinheiten ähnlich sind. Es kann auch festgestellt werden, ob dies auch für die Faktorladungen, die Faktorkovarianzen und die Einzelrestvarianzen zutreffend ist. In Abbildung 1 ist ein Modell einer konfirmatorischen Faktorenanalyse dargestellt, dass aus zwei Messmodellen und einem Strukturmodell besteht.

Abbildung 1: Modell einer konfirmatorischen Faktorenanalyse



Anmerkung: δ=Residualvariable der Indikatoren; λ= Faktorladungen, ξ=latente Variable, Φ= Korrelation zwischen den latenten Faktoren

Würde nun für dieses Modell auf einer empirischen Datengrundlage basieren, ließen sich Parameterwerte hierfür schätzen. Es wird hierbei jeweils eine unstandardisierte und eine standardisierte Lösung ausgegeben. Bei der unstandardisierten Lösung werden Regressionskoeffizienten zwischen den Indikatoren und dem Konstrukt, die Kovarianz zwischen den Konstrukten und die Varianz der Konstrukte und der uniquen Faktoren ausgegeben. (vgl. Backhaus/Erichson/Plinke/Weiber 2006:404-406) Bei der standardisierten Lösung werden die Varianzen der Indikatoren und Konstrukte hingegen auf den Wert Eins fixiert. So werden die Korrelationen zwischen den Indikatoren und dem Konstrukt, welche den Faktorladungen entsprechen, ausgegeben. Zudem wird die Korrelation zwischen den Konstrukten und die Indikatorreliabilität, welche der quadrierten multiplen Korrelationskoeffizienten entspricht, ausgegeben.

Dann wird mittels diverser *Goodness-of-Fit*-Indizes und Teststatistiken geprüft, wie gut das Modell mit den Daten übereinstimmt. Die Anpassungsgüte eines konfirmatorischen Modells wird auch als Modellfit bezeichnet. Sollten diese Gütemaße zeigen, dass sie nicht übereinstimmen, können Modellanpassungen vorgenommen werden, die weiterhin theoriegeleitet erfolgen sollten. Dieses veränderte Modell wird anschließend einer erneuten Überprüfung unterzogen. Sollten schließlich die Gütemaße des Modells auf einen guten Modellfit hinweisen, wird dieses Modell nach derselben Vorgehensweise auch in jeder der anderen Analyseeinheiten auf Übereinstimmung mit den Daten

überprüft. Sollte dieses Modell in allen Analyseeinheiten mit den Daten übereinstimmen, wäre dies ein Anzeichen für die Existenz einer funktionalen Äquivalenz der Skalen. (vgl. Van de Vijver 2003b) Jedoch darf dieses Ergebnis nicht überbewertet werden, da dies noch nichts über die Ähnlichkeit der einzelnen Parameterschätzungen, wie zum Beispiel die Faktorladungen aussagt, sondern lediglich über die Ähnlichkeit der Gesamtstruktur des Modells. Es können die Parameter der einzelnen Modelle allerdings auch verglichen werden, um festzustellen, ob diese vergleichbar sind. Für gewöhnlich werden hierfür die unstandardisierten Parameter der einzelnen Gruppen miteinander verglichen, weil empirische Varianzen beziehungsweise Standardabweichungen in diesen zumeist verschieden sind. (vgl. Reinecke 2005:69)

Der letzte und entscheidende Schritt bei der Überprüfung, welches Ausmaß der Äquivalenz der Skalen für einen Vergleich in den betrachteten Analyseeinheiten vorliegt, stellt die Durchführung eines multiplen Gruppenvergleichs dar. Dieser simultane Gruppenvergleich hat den Zweck festzustellen, ob das Modell in den einzelnen Analyseeinheiten in vergleichbarer Weise mit den Daten übereinstimmt. Es wird getestet, inwieweit eine Invarianz des Modells zwischen den verschiedenen Analyseeinheiten existent ist. Die Invarianz des Modells wird durch die schrittweise Gleichsetzung der einzelnen Parameter zwischen den Gruppen geprüft. Sollte sich die Güte des Models, sichtbar anhand einer Teststatistik, zwischen dem restriktiveren und dem weniger restriktiven Modell verschlechtern, wird davon ausgegangen, dass die betreffenden Parameter des Modells zwischen den Gruppen nicht gleich sind.

Es kann zwischen einer *bottom-up* und einer *top-down* Prozedur unterschieden werden, anhand derer der multiple Gruppenvergleich durchgeführt werden kann. (vgl. Reinecke 2005:151-154) Bei der *bottom-up* Vorgangsweise wird vom Basismodell ausgegangen und schrittweise die einzelnen Parameter gleichgesetzt während bei der *top-down* Prozedur vom restriktivsten Modell ausgegangen wird und schrittweise die Gleichsetzung der Parameter aufgehoben wird. In dieser Arbeit wird die Überprüfung der erreichten Ebene der Äquivalenz anhand der *bottom-up* Prozedur durchgeführt. Der erste Schritt stellt die Überprüfung der Güte des Modells für die jeweiligen Länder ohne Festsetzung irgendwelcher Parameter dar. Ist der Modellfit in allen Ländern angemessen, kann davon ausgegangen werden, dass die Struktur der Beziehung der Indikatoren zum Faktor in allen Ländern gleich ist und eine funktionale Äquivalenz

gegeben ist. (vgl. Van de Vijver 2003b) Als nächster Schritt werden die einzelnen Faktorladungen zwischen den Gruppen gleichgesetzt. Zeigt sich, dass auch die Faktorladungen zwischen den Analyseeinheiten invariant sind, liegt zumindest eine *Measurement unit equivalence* vor. Danach folgt eine zusätzliche Gleichsetzung der Kovarianzen der latenten Variablen. Als letzter Schritt erfolgt eine zusätzliche Gleichsetzung der Fehlervarianzen.

Je ähnlicher die Parameterschätzungen zwischen den Modellen der einzelnen Analyseeinheiten sind, desto höher ist das Ausmaß der vorliegenden Äquivalenz. (vgl. Van de Vijver/Leung 1997) Wobei eine Varianz der Parameter der uniquen Faktoren für das vorliegende Ausmaß der Äquivalenz am unbedeutsamsten ist. Dies entspricht in etwa der Billigung der Variation der Eigenwerte der Faktoren zwischen den Analyseeinheiten bei der explorativen Faktorenanalyse. (vgl. Van de Vijver 2003b) Sollte sich jedoch herausstellen, dass nach der Gleichsetzung der übrigen Parameter sich die Güte des Gesamtmodells verschlechtert, werden paarweise Vergleiche zwischen den einzelnen Gruppen durchgeführt. Damit soll festgestellt werden, zwischen welchen Ländern welches Ausmaß an Äquivalenz vorliegt. Sollte zwischen zwei Gruppen keine Parameterähnlichkeit vorherrschen, werden zudem die betreffenden Parameter der beiden Messmodelle der Skalen abwechselnd gleichgesetzt, um zu überprüfen, ob zumindest eine Ähnlichkeit bei einem der beiden Messmodelle gegeben ist.

5. Explorative Vorgehensweise der Überprüfung der erreichten Ebene der Äquivalenz

5.1. Details zur Erhebungsphase und des Stichprobendesigns des ESS

In der Tabelle 2 sind die wichtigsten Details zur administrativen Vorgangsweise bei der dritten Erhebungswelle des ESS, wie zum Beispiel die gewählte Datenerhebungsmethode und das jeweilige Design der Stichproben dargestellt. Dies hat den Zweck zu überprüfen, ob vielleicht offensichtliche *Administration* oder *Sample bias* vorliegen könnten, wobei der Möglichkeit der Existenz einer Verzerrung der Stichproben eine besondere Aufmerksamkeit gewidmet wird.

Es existieren zwar leichte Variationen zwischen den einzelnen Ländern bei den gewählten Datenerhebungsmethoden, die allerdings so gering ausfallen, dass Verzerrungen durch diese auszuschließen sind. Als problematisch könnte es sich jedoch erweisen, dass die Befragungen in Österreich um einiges später als in Deutschland und Großbritannien durchgeführt wurden. Es könnten nämlich länderspezifische Ereignisse aufgetreten sein, die einen Einfluss auf Einstellungen und Meinungen der befragten Personen haben könnten. Dies könnte eine Vergleichbarkeit der erhobenen Daten gefährden.[39] Weitere Unterschiede bezüglich der administrativen Vorgehensweise betreffen die Durchführung von nationalen Pretests, die Versendung von zusätzlichen Informationsbroschüren und Ankündigungsbriefen, der Verwendung von Incentives und der Anzahl der Kontaktversuche.[40] Ob diese administrativen Unterschiede Auswirkungen auf die Ausschöpfungsrate haben und vielleicht sogar Verzerrungen bedingen, wird im Anschluss an die Diskussion der Stichprobendesigns erörtert.

[39] Für eine Darstellung des Einflusses von nationalen sowie internationalen Ereignissen auf die Einstellungen und Meinungen von Individuen siehe Stoop 2007.

[40] Für eine Diskussion der Vor- und Nachteile von Maßnahmen zur Erhöhung der Ausschöpfungsrate siehe Neller 2005.

Tabelle 2: administrative Vorgangsweise bei der 3. Erhebungswelle des ESS

	Österreich	Deutschland	Großbritannien
Grundgesamtheit	älter als 15 Jahre	älter als 15 Jahre	älter als 15 Jahre
Datenerhebungsmethode	Face-to-Face (PAPI)	Face-to-Face (CAPI)	Face-to-Face (CAPI)
Erhebungszeitraum	18.07.07-15.11.07	01.09.06-15.01.07	05.09.06-14.01.07
nationaler Pretest	Nein	Ja	Ja
Informationsbroschüren und Ankündigungsbrief	Nein	Ja	Ja
Incentives	Nein	Ja	Ja
Minimale Anzahl an Kontaktversuchen	3, davon 1 am Wochenende und 1 am Abend	4, keine weiteren Angaben	5, davon 1 am Wochenende und 1 am Abend
Stichprobendesign	geschichtete dreistufige Zufallsauswahl	geschichtete zweistufige Zufallsauswahl	geschichtete dreistufige Zufallsauswahl
Anzahl der Schichten	363	1100 für West-, 439 für Ostdeutschland	Keine Angabe
1. Stufe: Auswahl der Primäreinheiten	380 Cluster in 300 Gemeinden; die Nummer der Cluster ist größenproportional zur Bevölkerungszahl einer Schicht, innerhalb einer Schicht erfolgt die Auswahl durch systematische größenproportionale Zufallsauswahl	Auswahl von 104 Gemeinden oder anders ausgedrückt Cluster für West- und 52 für Ostdeutschland mittels systematischer größenproportionaler Zufallsauswahl. Dies ergibt 110 Sampling Points im Westen und 57 im Osten, weil manche ausgewählte größere Gemeinden mehr als einen Sampling Point haben	Auswahl von 192 Schichten bestehend aus Postleitzahlen aus GB, die je weniger als 500 Adressen umfassen und dem Postcode Adress File (PAF) entnommen werden und 6 aus Nordirland; innerhalb einer Schicht erfolgt die Auswahl durch systematische größenproportionale Zufallsauswahl
2. Stufe: Auswahl der Sekundäreinheiten	Bei jedem Cluster werden 10 Haushalte für die Bruttostichprobe ausgewählt, von denen 5 aus einer Telefonbuch-CD gezogen werden und 5 per Random Route Verfahren ausgewählt werden. Startpunkt: der aus dem Telefonbuch-CD gezogene Haushalt	In jedem der 167 Sampling Points systematische Zufallsauswahl von 29 Personen mittels lokaler Melderegister. Zusätzliche Ziehung von 21 Individuen pro Sampling Point. Falls die Non-Response-Rate zu hoch ist, Ziehung nach selber Vorgangsweise, allerdings von jedem Sample Point gleich viele	Zufällige Auswahl von jeweils 24 Adressen
3. Stufe: Auswahl der Tertiäreinheiten	Innerhalb eines Haushalts erfolgt die Ziehung der Zielperson mittels der Geburtstagsmethode		Wenn mehrere Wohneinheiten bei einer Adresse: Zufallsauswahl mittels Schwedenschlüssel; innerhalb einer Wohneinheit: Auswahl der Person ebenfalls mittels Schwedenschlüssel

Quelle: Jowell and the Central Coordinating Team (2007)[41]
Lesebeispiel: Für die deutsche und britische Fragebogenversionen wurde in der dritten Erhebungswelle des ESS ein Pretest durchgeführt. Für die österreichische Version wurde hingegen kein Pretest durchgeführt.

[41] http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation

Vergleiche der einzelnen Stichprobendesigns[42]

Der Vorteil einer Ziehung aus einer Telefonbuch-CD wie beim österreichischen Stichprobendesign des ESS ist, dass sie billig und schnell zu realisieren ist. Es ist allerdings rein durch eine Ziehung aus dem Telefonbuch keine Erfassung von Haushalten ohne Telefonanschluss und mit Geheimnummer möglich. Hierbei muss auf die Möglichkeit einer Stichprobenverzerrung auf Grund unterschiedlicher Handynutzungsraten in den verschiedenen Ländern hingewiesen werden, die letztlich die Äquivalenz gefährdet. Die Entwickler des Designs für die österreichische Stichprobe wollten dies durch ein Random-Route-Verfahren kompensieren.

Das Stichprobendesign Großbritanniens weist gewisse gemeinsame Merkmale mit dem österreichischen Design auf. So weist dieses Vorteile bezüglich der Kosten und der dafür verwendeten Zeitressourcen auf, muss allerdings ebenfalls bereinigt werden. Dadurch, dass die Stichprobe aus einem Adressenregister gezogen wird wo alle Adressen enthalten sind, entfällt die Notwendigkeit den *Coverage error* zu kompensieren. Dafür ist es notwendig, dass für die endgültige Auswahl der Stichprobeneinheiten eine zweimalige Anwendung eines Schwedenschlüssels notwendig werden kann.

Eine Ziehung aus Melderegistern wie im Fall der deutschen Stichprobe hat den entscheidenden Vorteil, dass der Auswahlrahmen für die Stichprobe laufend aktualisiert wird und alle mit dem Hauptwohnsitz im jeweiligen Land gemeldeten Personen enthalten sind. Dies hat zur Folge, dass Inländer und Ausländer, dieselbe Auswahlwahrscheinlichkeit haben. Ein weiterer Vorteil der Zufallsauswahl aus Melderegistern stellt der Umstand dar, dass im Falle einer Interviewverweigerung zumindest minimale Information bezüglich des Verweigerers, wie zum Beispiel das Geschlecht und das Alter vorhanden sind, welche verwendet werden können, um etwaige systematische Verzerrungen der Stichprobe zu kontrollieren und bei Bedarf fehlende Daten zu imputieren oder nachträglich eine Nonresponse-Gewichtung vorzunehmen. Die Nachteile bestehen darin, dass die Ziehung aus lokalen Melderegistern mit hohen Kosten verbunden ist und in der Praxis Probleme bereiten

[42] Für eine Darstellung der spezifischen Probleme bei international vergleichenden Umfragestudien bezüglich stichprobenbedingter Fehler siehe Kapitel 3.1.1.

kann, falls zum Beispiel Gemeinden nicht bereit oder in der Lage sind Adressziehungen durchzuführen. (vgl. Schnell/Hill/Esser 2008:288)

Der wichtigste Unterschied zwischen den einzelnen Stichprobendesigns besteht in den Auswahlwahrscheinlichkeiten der Befragten. Die aus Melderegistern gezogene Personenstichprobe ist auf Personenebene selbstgewichtend, weil alle Zielpersonen die gleiche Auswahlwahrscheinlichkeit haben. Bei den anderen beiden Stichprobendesigns dagegen werden lediglich die für Interviews vorgesehenen Haushalte beziehungsweise Adressen gleichwahrscheinlich ausgewählt. Durch die Verwendung der Geburtstags- beziehungsweise Schwedenschlüsselmethode ist es auf Grund der unterschiedlichen Auswahlwahrscheinlichkeiten der Befragten, notwendig Transformationsgewichtungen durchzuführen. Anders ausgedrückt, es muss proportional zu der Anzahl der Wohneinheiten bei einer Adresse und den in einem Haushalt befindlichen Personen, die der Grundgesamtheit angehören, gewichtet werden, um verzerrungsfreie Aussagen über die interessierende Population machen zu können. Dieses so errechnete Gewicht wird als Designgewicht bezeichnet, weil es durch das Design einer Stichprobe determiniert ist. (vgl. Schnell/Hill/Esser 2008:280) So korrigieren die Designgewichte im ESS nicht nur die Designeffekte einer Stichprobe bezüglich der unterschiedlichen Auswahlwahrscheinlichkeiten der Stichprobenelemente, sondern auch jene die durch die Zusammenfassung der Stichprobenelemente zu Klumpen entstehen. (vgl. Häder/Lynn 2007) [43] Dies ist notwendig, da Elemente eines Klumpens für gewöhnlich eine größere Homogenität aufweisen als Personen in der Grundgesamtheit und diese Homogenität auch zwischen den Ländern variiert. (vgl. ebd.)

Wie bereits erläutert, sind idente Stichprobendesigns nicht notwendig, um eine Äquivalenz der Stichproben herzustellen beziehungsweise auf Grund länderspezifischer rechtlicher Unterschiede auch nicht realistisch.(vgl. Kap. 3.1.1) Wichtig ist hingegen, dass diese auf einer einheitlichen Definition der interessierenden Grundgesamtheit und

[43] Neben der Designgewichtung existiert im ESS noch die Möglichkeit eine Gewichtung der Bevölkerungsgröße vorzunehmen. Es werden Stichproben gezogen, die eine vergleichbare Anzahl an Elementen enthalten, egal wie hoch die tatsächliche Bevölkerungszahl eines Landes ist. Das sogenannte Population size weight ist ein Korrekturfaktor dieses Umstandes und stellt sicher, dass ein jedes Land gemäß seiner tatsächlichen Bevölkerungsstärke in der Umfrage repräsentiert wird. (ESS o.J.d: http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation) Während das Designgewicht eigentlich immer verwendet werden muss, um „repräsentative“ Ergebnisse zu erhalten, muss das Populationsgewicht lediglich bei simultanen Vergleichen zwischen verschiedenen Ländern angewandt werden. (vgl. ebd.) Dies beinhaltet zum Beispiel den Vergleich absoluter Häufigkeiten oder Mittelwertvergleiche.

auf einer wahrscheinlichkeitsbasierten Prozedur basieren, wobei die Auswahlwahrscheinlichkeit der Stichprobenelemente bekannt sein muss. So kann mittels einer nachträglichen Imputation von Designgewichten eine idente Auswahlwahrscheinlichkeit aller Befragten und somit die Basis für eine Äquivalenz der Stichproben zwischen den einzelnen Ländern hergestellt werden. Basis deswegen, weil noch immer ein *Nonresponse error* vorliegen könnte, der eine Vergleichbarkeit der Stichproben verhindern könnte.

Vergleich der Ergebnisse der Stichprobenziehungen

Den Abschluss der Betrachtung der Stichproben der dritten Erhebungswelle des ESS bildet ein Vergleich der Ergebnisse der Stichprobenziehungen in den einzelnen Ländern. Damit soll festgestellt werden, ob vielleicht ein *Nonresponse error* in einem der Länder vorliegen könnte, der die Äquivalenz der Stichproben beeinträchtigen würde.[44] Die Unterscheidung zwischen verschiedenen Gründen für Nonresponse ist möglich, da deren Erhebung im Rahmen der Feldarbeit zwingend vorgeschrieben war. Sie ist deshalb notwendig, da im ESS keine Nonresponse-Gewichtung durchgeführt wurde, um Unterschiede zwischen der Stichprobe und der interessierenden Grundgesamtheit zu reduzieren. (ESS o.J.d)[45] Eine Besonderheit des ESS stellt der Umstand dar, dass für die Bestimmung der Äquivalenz der Stichproben im Sinne einer vergleichbaren Präzision weniger die Nettostichprobengröße wichtig ist, sondern die effektive Stichprobengröße. (vgl. Häder/Lynn 2007) Die effektive Stichprobengröße ist der Quotient aus der Nettostichprobengröße und dem absoluten Designeffekt einer Stichprobe. Das Ziel beim ESS war es für Länder, die mehr als 2 Millionen Einwohner haben, eine effektive Stichprobengröße von mindestens 1500 Personen zu erreichen, um eine vergleichbare Präzision zwischen den Stichproben herzustellen. (vgl. Häder/Lynn 2007) Da die Designeffekte einer Stichprobe vor der Ziehung der Elemente schon geschätzt werden können, war es durch diese Festlegung und der erwarteten Ausschöpfungsrate in einem Land auch möglich die Bruttostichprobengröße zu bestimmen, die eine vergleichbare Präzision der einzelnen Substichproben der Länder ermöglichen sollte. (vgl. Häder/Lynn 2007)

[44] Da in dieser Arbeit lediglich ein Befragungszeitpunkt behandelt wird, kann es sich um eine einmalige Verzerrung der Stichproben handeln.

[45] http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation

Tabelle 3: Kennzahlen der Stichproben der Analyseeinheiten

	Österreich	Deutschland	Großbritannien
Bruttostichprobengröße	3800	5712	4752
stichprobenneutrale Ausfälle	40	359	365
Bereinigte Stichprobe	3760 (100 %)	5353 (100 %)	4387 (100 %)
explizite Verweigerung	920 (24,5 %)	1267 (23,7 %)	1192 (27,2 %)
Unmöglichkeit des Kontakts	378 (10 %)	387 (7,2 %)	408 (9,3 %)
Verhinderung	39 (1 %)	633 (11,8 %)	378 (8,6 %)
ungültige Interviews	18 (0,5 %)	150 (2,8 %)	15 (0,3 %)
Nettostichprobengröße	2405 (64 %)	2916 (W: 1901,O: 1015) (54, 5 %)	2394 (54,6 %)
Effektive Stichprobengröße	1509	1202	1386
Ausschöpfungsrate	64 %	54,5 %	54,6 %

Quelle: Jowell and the Central Coordinating Team (2007)[46]; ESS Central Coordinating Team (2008)[47]
Anmerkung: eigenständige Berechnung der Kennzahlen auf Basis der zur Verfügung gestellten Werte auf folgende Weise: Bereinigte Stichprobengröße = Bruttostichprobengröße- stichprobenneutrale Ausfälle (=Adresse unbewohnt beziehungsweise zum Beispiel Bürogebäude, Befragter umgezogen od. verstorben); Nettostichprobengröße= Bereinigte Stichprobe - Explizite Verweigerung - Unmöglichkeit des Kontakts – Verhinderung (=Sprachbarriere, mentale oder physische Beeinträchtigung des Befragten und andere Gründe)- ungültige Interviews; Ausschöpfungsrate= (Nettostichprobengröße/ Bereinigte Stichprobe)*100. Die angeführten Anteilswerte beziehen sich jeweils auf die Bereinigte Stichprobe. Die angegeben Prozentzahlen wurden nach mathematischer Konvention gerundet und die Summe der Anteilswerte kann deswegen geringfügig vom Sollwert (=100 %) abweichen.
Lesebeispiel: Die Ausschöpfungsrate beträgt in Österreich 64 %, in Deutschland 54,5 % und in Großbritannien 54,6 %.

Wenn die Tabelle 3 betrachtet wird, fällt auf dass die Erreichung einer vergleichbaren Präzision der Substichproben der einzelnen Länder nicht vollends geglückt ist**.** So erreicht die österreichische Stichprobe, bei einer Ausschöpfungsrate von 64 % die angepeilte effektive Stichprobengröße von 1500 Personen. Die deutschen Stichprobe und die Stichprobe aus Großbritannien weisen allerdings eine weitaus geringere effektive Stichprobengröße auf. Die niedrigeren Ausschöpfungsraten stehen vielleicht in Zusammenhang mit dem Umstand, dass die Vorgaben der minimalen Kontaktversuche zur Verringerung der Nonresponse-Quote in den Ländern unterschiedlich eingehalten wurden (vgl. ESS Central Coordinating Team 2008:33)[48] So wurde zum Beispiel in Deutschland bei fast der Hälfte der Nonrespondenten kein zweiter Kontaktversuch vorgenommen und bei über 70% der Nonrespondenten nicht die geforderten vier Kontaktversuche. (vgl. ebd. 27) Dies könnte eine Folge davon sein, dass einerseits aus den regionalen Melderegistern gezogen wurde und so gewisse Eigenschaften der Nonrespondenten bekannt waren. Andererseits stand eine „Reservestichprobe“ zur

[46] http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation
[47] http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=547&itemid=80
[48] http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=547&itemid=80

Verfügung, falls die Ausschöpfungsrate zu niedrig sein sollte. Jedoch können diese Unterschiede bezüglich der Präzision der Substichproben durch die Designgewichtung korrigiert werden.

Wichtiger als die Ausschöpfungsraten erscheint die Betrachtung der Nonrespondenten hinsichtlich der Gründe ihrer Nichtteilnahme an der Befragung zu sein. Hierbei kann zwischen Gründen unterschieden werden, auf die die Studienverantwortlichen keinen Einfluss haben und solche die in einem gewissen Ausmaß beeinflussbar sind. (vgl Philippens/Billiet o.J.:10-17)[49] Die unterschiedlichen Anteile an Personen, die die Durchführung eines Interviews explizit verweigert haben, sind nicht beeinflussbar und als unterschiedliche Befragungsakzeptanz in den einzelnen Ländern zu interpretieren. Zudem sind länderspezifische Unterschiede bezüglich der Zeit die Individuen zu Hause verbringen ebenso zu erwarten. Falls zum Beispiel in zwei Ländern unterschiedliche Erwerbsquoten der weiblichen Bevölkerung existieren, wird sich dies in unterschiedlichen Kontaktraten dieser Subpopulation zeigen. Dies ist ebenfalls eine gegebene Tatsache und nicht beeinflussbar. Allerdings haben die Forscher auch gewisse Einflussmöglichkeiten auf diese Umstände wie zum Beispiel mehrfache Kontaktversuche, Konvertierung von Verweigerern, die Ausgabe von Incentives und gezielte Interviewerschulungen.

Es zeigt sich, dass die bereinigte deutsche Stichprobe verglichen mit den anderen zwei Ländern mit zirka 12% den höchsten Anteil an Personen aufweist, mit denen kein Interview auf Grund einer Verhinderung durchgeführt werden konnte. In Österreich beträgt der Anteil der verhinderten Personen hingegen lediglich 1 %. Deutschland weist auch mit 3% den mit Abstand höchsten Anteil an ungültigen Interviews auf. Großbritannien weist hingegen mit knapp 27% den höchsten Anteil an expliziten Interviewverweigerungen auf. In Österreich konnte hingegen mit 10% der für die Stichprobe vorgesehenen Personen kein Kontakt hergestellt werden. Worin diese Unterschiede begründet sind, kann auf Grund der Datenlage nur gemutmaßt werden. Dies ist allerding auch nicht wirklich für den Zweck dieser Arbeit relevant. Fakt ist, dass die Unterschiede zwischen den einzelnen Ländern bei den Gründen der Nichteilnahme an einem Interview sind zum Teil relativ groß sind. So kann letztlich die Möglichkeit der Existenz eines *Nonresponse errors* und somit die Verzerrung des

[49] http://www.ined.fr/fichier/t_rendezvous/126/rendezvous_fichier_philippens.pdf

Vergleichs der Ergebnisse der dritten Erhebungswelle des ESS zwischen diesen Ländern nicht mit Sicherheit ausgeschlossen werden.

5.2. Beschreibung der zu überprüfenden Skalen und Betrachtung der Übersetzung

In diesem Kapitel werden die Skalen dargestellt, die in weiterer Folge hinsichtlich der erreichten Ebene der Äquivalenz für einen Vergleich zwischen den Analyseeinheiten Großbritannien, Österreich, Ost- und Westdeutschland überprüft werden sollen. Im Anschluss daran wird die Übersetzung der Skalen aus dem Master-Fragebogen betrachtet, um die Äquivalenz dieser Übersetzungen abzuschätzen.

Die zwei zu testenden Skalen messen die Einstellung der Befragten zu Immigration und die Meinung bezüglich der Auswirkungen dieser anhand verschiedener Kriterien. Die Kriterien anhand derer die Einstellung der Befragten zur Einwanderung gemessen werden sollen stellen die „Rasse“ oder ethnische Gruppe dar, der die Migranten zugehörig sind und der Wohlstand des Herkunftslandes beziehungsweise, dass das Herkunftsland sich in nicht Europa befindet. Diese Ratingskala besteht aus drei Items**,** wobei zu den einzelnen Fragen jeweils dieselben vier Antwortkategorien ausgewählt werden können. Diese Antwortkategorien legen fest, in welcher Quantität der Befragte eine Immigration gemäß gewisser Merkmale der Immigranten befürwortet. Die zweite Ratingskala erfasst die Meinung der Befragten bezüglich der Auswirkungen von Einwanderung auf das Einwanderungsland. Dies beinhaltet die Erfassung der subjektiv empfundenen Auswirkungen auf die Wirtschaft, auf das kulturelle Leben und die generelle Qualität des Einwanderungslandes als Lebensort. Die Befragten haben die Möglichkeit auf elfstufigen Antwortskalen mit jeweils zwei Gegenpolen ihre Meinung hierzu mitzuteilen.

Die beiden Skalen unterscheiden sich deutlich in ihrer Struktur. Sie weisen eine unterschiedliche Anzahl an Skalenpunkten bei den Antwortskalen auf. Die zweite Skala enthält im Gegensatz zur ersten einen Skalenmittelpunkt, also eine neutrale Mittelkategorie, so dass Unentschlossene oder Meinungslose nicht zur Abgabe einer inhaltlichen Antwort „gezwungen“ werden. Die Befragten haben hierbei die

Möglichkeit die ihrer Meinung nach bestehenden Auswirkungen von Migration für das Einwanderungsland zu bewerten. Bei der ersten Skala ist hingegen jeder Punkt der Antwortskala verbalisiert und es ist die Einstellung von Interesse in welchen Mengen es Migranten mit verschiedenen Charakteristiken erlaubt sein sollte einzuwandern. Eine Verbalisierung ist bei der zweiten Skala nur für die beiden gegensätzlichen Extremkategorien vorhanden. Was beide Skalen gemeinsam haben, ist dass beide bipolare Rating-Skalen darstellen. Die in dieser Arbeit verwendeten Rating-Skalen weisen einige Unterschiede zu den typischen Charakteristiken von Likert-Skalen auf. So sollen die Items der beiden Skalierungsverfahren zwar auch beide monotone Itemcharakteristiken aufweisen und lediglich eine Dimension messen. Allerdings bestehen zum Beispiel die zwei in dieser Arbeit zu prüfenden Rating-Skalen aus relativ wenigen Items. Bei Likert-Skalen werden mittels Statements der Grad der Zustimmung beziehungsweise Ablehnung und somit die Einstellungsstärke erfasst. Bei diesen Rating-Skalen werden hingegen die Befragten gebeten auf Fragen ihre diesbezüglichen Einschätzungen preiszugeben. Diese Angaben bezüglich der Häufigkeiten und Bewertungen auf diese Fragen werden dann in weiterer Folge in gleicher Weise wie bei einer Likert-Skala verwendet. Das heißt, dass Skalensummenwerte gebildet werden können, die letztlich den Ausprägungen der Befragten auf der hinter den Items stehende Einstellungs- beziehungsweise Meinungsdimension entsprechen.

Betrachtung der Übersetzung der Skalen

In der folgenden Tabelle werden die länderspezifischen Skalenversionen dargestellt. Dies hat den Zweck die Skalen detailliert vorzustellen und zu prüfen, ob offensichtliche Verzerrungen existieren könnten, die eine linguistische und pragmatische Bedeutungsäquivalenz unwahrscheinlich erscheinen lassen. Der ESS basiert auf einem sequentiellem Design, genauer gesagt auf dem *Ask-the-Same-Question* Modell, was bedeutet, dass der Master-Fragebogen fertiggestellt wird, bevor die diversen Übersetzungen beginnen, die anschließend für jede Sprache parallel durchgeführt werden. (vgl. Harkness 2007) Der Master-Fragebogen ist in englischer Sprache verfasst und identisch mit der Version die in Großbritannien abgefragt wurde. Die Vorgangsweise bei der Fragebogenübersetzung entspricht dem TRAPD Ansatz (*Translation, Review, Adjudication, Pretesting and Documentation*). Ländern mit einer gemeinsamen Landessprache wurde zudem die Möglichkeit gegeben, die

Übersetzungsaufgaben aufzuteilen und sich somit gegenseitig Hilfestellungen und Beratung bei etwaigen Problemen geben zu können. (vgl. Harkness 2007)

In der englischen Version des ESS wird beim ersten und zweiten Items nach „race or ethnic group“ gefragt, während in den deutschen Übersetzungen nach der Volksgruppe oder der ethnischen Gruppe gefragt wird.(vgl. Tab. 4) Der Begriff der „Rasse“ ist im deutschsprachigen Raum nach den verheerenden Auswirkungen des Nationalsozialismus und seiner Ideologie nicht mehr gesellschaftlich akzeptiert. Diese negative Konnotation hätte im Vergleich zur englischen Fragebogenversion vermutlich massive Antwortverzerrungen zur Folge. Es musste also ein anderer möglichst bedeutungsgleicher Begriff gefunden werden. Das Problem hierbei ist, dass es sich bei dem Begriff der Volksgruppe in Österreich um einen rechtlich besetzten Begriff handelt, der unterschiedliche Assoziationen bei den Befragten in Österreich als in Deutschland evozieren könnte. So wird im Paragraph 1, Absatz 2 des Volksgruppengesetzes, einem Bundesgesetz, der Begriff der Volksgruppe folgendermaßen definiert: „Volksgruppen im Sinne dieses Bundesgesetzes sind die in Teilen des Bundesgebietes wohnhaften und beheimateten Gruppen österreichischer Staatsbürger mit nichtdeutscher Muttersprache und eigenem Volkstum.“ (vgl. Bundeskanzleramt Rechtsinformationssystem 2009)[50] Das entspricht einem Schutz für ethnische Minderheiten eines Landes. Demzufolge ist der Begriff nicht für die Bezeichnung der Mehrheit eines Landes vorgesehen. Es existiert zudem das Problem, dass die Begriffe „Rasse“ und Volksgruppe nicht komplett bedeutungsäquivalent erscheinen und damit ein Problem darstellen könnten. Dieses Problem war auch den Übersetzern klar, jedoch konnte vermutlich kein anderer äquivalenter Begriff gefunden werden. (vgl. ESS o.J.b:44)[51] Ob ihnen jedoch auch die Existenz der speziellen rechtlichen Bedeutung des Begriffs der Volksgruppe in Österreich bekannt war, ist allerdings nicht bekannt.

[50] http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10000602
[51] http://www.europeansocialsurvey.org/index.php?option=com_docman&task=cat_view&gid=95&Itemid=80

Tabelle 4: die Fragebogenversionen der einzelnen Länder

	Großbritannien	Deutschland	Österreich
Einleitungstext	Now some questions about people from other countries coming to live in [country].	Ich möchte ihnen nun ein paar Fragen zu Menschen stellen, die aus anderen Ländern nach Deutschland kommen, um hier zu leben.	Nun einige Fragen in Bezug auf Menschen aus anderen Ländern, die nach Österreich kommen, um hier zu leben.
Skala 1: Einstellungen zu Immigration Item 1 (imsmetn)	To what extent do you think [country] should allow people of the same race or ethnic group as most [country's] people to come and live here?	Zunächst geht es um die Zuwanderer, die derselben Volksgruppe oder ethnischen Gruppe angehören wie die Mehrheit der Deutschen. Wie vielen von ihnen sollte es Deutschland es erlauben, hier zu leben? Sollte Deutschland es	Zunächst geht es um Zuwanderer, die derselben Volksgruppe oder ethnischen Gruppe wie die meisten Österreicher angehören. Wie vielen von ihnen sollte es Österreich erlauben, sich hier niederzulassen?
Item 2 (imdfetn)	How about people of a different race or ethnic group from most [country] people?	Wie ist das mit Zuwanderern, die einer anderen Volksgruppe oder ethnischen Gruppe angehören als die Mehrheit der Deutschen? Sollte Deutschland es	Wie ist das mit Zuwanderern, die einer anderen Volksgruppe oder ethnischen Gruppe angehören als die Mehrheit der Österreicher? Sollte Österreich …?
Item 3 (impcntr)	How about people from the poorer countries outside Europe?	Und wie ist das mit Zuwanderern, die aus den ärmeren Ländern außerhalb Europas kommen? Sollte Deutschland es	Wie ist das mit Zuwanderern, die aus ärmeren Ländern außerhalb Europas stammen?
Antwortskala bei Item 1-3	1:Allow many to come and live here, 2:Allow some, 3: Allow a few, 4: Allow none	1:vielen erlauben, herzukommen und hier zu leben, 2: einigen erlauben, 3: ein paar wenigen erlauben, 4: niemandem erlauben	1: es vielen erlauben, sich hier niederzulassen, 2: es einigen erlauben, 3: es wenigen erlauben, 4: es keinem erlauben
Skala 2 Auswirkungen von Immigration Item 4 (imbgeco)	Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?	Was würden Sie sagen, ist es im Allgemeinen gut oder schlecht für die deutsche Wirtschaft, dass Zuwanderer hierher kommen?	Würden Sie sagen, dass es generell schlecht oder gut für die österreichische Wirtschaft ist, dass Zuwanderer nach Österreich kommen, um hier zu leben?
Antwortskala	0:Bad for the economy, 10:Good for the economy	0:Schlecht für die Wirtschaft, 10:gut für die Wirtschaft	0:Schlecht für die Wirtschaft, 10:gut für die Wirtschaft
Item 5 (imueclt)	Would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?	Würden Sie sagen, dass das kulturelle Leben in Deutschland im Allgemeinen von Zuwanderern, untergraben oder bereichert wird?	Würden Sie sagen, dass das kulturelle Leben in Österreich im Allgemeinen von Zuwanderern, die nach Österreich kommen, untergraben oder bereichert wird?
Antwortskala	0:Cultural life undermined, 10:Cultural life enriched	0:Kulturelles Leben wird untergraben, 10:kulturelles Leben wird bereichert	0:Kulturelles Leben wird untergraben, 10:kulturelles Leben wird bereichert
Item 6 (imwbcnt)	Is [country] made a worse or a better place to live by people coming to live here from other countries?	Wird Deutschland durch Zuwanderer zu einem schlechteren oder besseren Ort zum Leben?	Wird Österreich durch Zuwanderer zu einem schlechteren oder besseren Ort zum Leben?
Antwortskala	0:Worse place to live; 10:Better place to live	0:wird zu einem schlechteren Ort zum Leben, 10:wird zu einem besseren Ort zum Leben	0:wird zu einem schlechteren Ort zum Leben, 10:wird zu einem besseren Ort zum Leben

Quelle: Master-Fragebogen (GB)[52], deutsche Fragebogenversion[53] österreichische Fragebogenversion[54]
Anmerkung: relevant erscheinende Abweichungen der einzelnen Fragebogenversionen wurden farblich markiert.

[52] http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=questionaires
[53] http://www.europeansocialsurvey.de/dokumentation/dritte.htm
[54] Die österreichische Fragebogenversion wurde von WISDOM per e-mail zur Verfügung gestellt.

Das nicht alle Unterschiede zwischen den deutschen und österreichischen Itemformulierungen sinnvoll erscheinen, wird an den folgenden zwei Beispielen sichtbar. So wird beim ersten Item in der deutschen Fragebogenversion danach gefragt, wie vielen Zuwanderern es erlaubt sein sollte, in Deutschland zu leben. In der österreichischen Version wird hingegen danach gefragt, wie vielen es erlaubt sein sollte, sich in Österreich niederzulassen. Die deutsche Version entspricht hierbei eher dem englischen Fragebogen als die österreichische. „Wo zu leben“ und „sich wo niederzulassen“ enthält unterschiedliche Assoziationsmöglichkeiten. Während die erste Formulierung auch temporär sein könnte, enthält die zweite eine endgültigere Komponente in dem Sinn eine Familie zu gründen und eine Staatsbürgerschaft anzustreben.

Das zweite Beispiel stellt das erste Item der zweiten Skala, also das Item 4 in Tabelle 4 dar. In der österreichischen Fragebogenversion wird nach den subjektiv empfundenen Auswirkungen auf die Wirtschaft gefragt, wenn Zuwanderer nach Österreich kommen, um hier zu leben. In der deutschen Version wird hingegen nach Zuwanderern gefragt, die nach Deutschland kommen. In diesem Fall ist die österreichische Itemformulierung der englischen näher als der deutschen. Ob diese beiden Unterschiede wirklich ein unterschiedliches Antwortverhalten der Befragten provozieren ist unklar, da im Einleitungstext zu diesen Skalen bereits dargestellt wird, dass es sich bei den nachfolgenden Fragen um Zuwanderer handelt, die in das jeweilige Land kommen, um hier zu leben. Ob diese Unterschiede jedoch notwendig sind um eine linguistische Äquivalenz sei zu erreichen, sei dahingestellt.

Ein Unterschied zwischen den drei Fragebogenversionen, der jedoch von Bedeutung ist, da er ein unterschiedliches Antwortverhalten der Befragten verursachen könnte, stellt die länderspezifische Formulierung der Antwortskalen bei der ersten Skala dar. Wie bereits dargestellt, können sich die wahrgenommen Intensitäten von Begrifflichkeiten zwischen verschiedenen Sprachen unterscheiden. (vgl. Kap. 3.1.1.) Es sind allerdings wieder die deutsche und die österreichische Formulierung der Antwortskala, die problematisch erscheint. So lautet die Formulierung des dritten Antwortskalenpunkts im englischsprachigen Fragebogen „Allow a few“. In der deutschen Version wurde dies mit „ein paar wenigen erlauben“ übersetzt, während es in der österreichischen Version mit „wenigen erlauben“ übersetzt wurde. Es kann vermutet werden, dass „ein paar

wenigen“ von der Quantität her als weniger beziehungsweise als stärkere Restriktion der Zuwanderung wahrgenommen wird, als die Formulierung „wenigen“. Diese unterschiedliche Intensität hat eine Auswirkung auf die Symmetrie der Antwortskalen in diesen beiden Ländern. Die österreichische Version der Antwortskala dürfte eher „gleicherscheinende Intervalle“ repräsentieren als die deutsche Version, bei der „ein paar wenigen erlauben“ eher näher einer absoluten Ablehnung von Zuwanderung sein dürfte. So ist es letztlich nicht auszuschließen, dass Befragte die eigentlich realiter dieselbe Ausprägung bezüglich Immigration aufweisen, zu einem unterschiedlichen Antwortverhalten „provoziert“ werden. Auch wenn beide Übersetzungen aus dem Englischen durchaus vertretbar sind, wäre in diesem Fall eine Angleichung der Antwortkategorien sinnvoller gewesen.

Die Betrachtung der verschiedenen Übersetzungen hat gezeigt, dass es nicht unproblematisch ist, einen Fragebogen in andere Sprachen zu übersetzen, um einen internationalen Vergleich zu ermöglichen. Es ist nämlich sehr schwer, die ursprünglich geplante Bedeutung einer Frage zu hundert Prozent in eine andere Sprache zu übersetzen, oder anders ausgedrückt äquivalent zu gestalten. Jedoch hat sich gezeigt, dass die größten Unterschiede zwischen den beiden Fragebögen in der eigentlich selben Sprache bestehen. Wie bereits mehrfach dargestellt, stellen linguistische Unterschiede hinsichtlich der Erreichung einer äquivalenten Formulierung der Items und der Antwortskalen per se kein Problem dar. Unterschiede können mitunter sogar notwendig sein, um Äquivalenz herzustellen. Es besteht nämlich die Möglichkeit, dass pragmatische und semantische Bedeutungsunterschiede zwischen verschiedenen Gruppen existieren, obwohl diese dieselbe Sprache sprechen. (vgl. Kap. 3.1.1) Die Gründe für die gefundenen Unterschiede zwischen den Fragebögen können im Rahmen der Diplomarbeit nicht geklärt werden, da keine Informationen über den Übersetzungsprozess verfügbar sind. Bis zu welchem Grad die hier gefundenen Unterschiede zwischen den deutschsprachigen Fragebogenversionen notwendig sind, um eine Äquivalenz herzustellen, oder ob es sich hierbei um Eitelkeiten der nationalen Umfragekoordinatoren oder gar um Kommunikationsprobleme handelt, ist dementsprechend fraglich.

5.3. Deskriptive Statistiken

Die Berechnungen die notwendig sind, um die Ebene der Äquivalenz der Skalen zu überprüfen werden fast ausschließlich mit dem Statistikprogramm SPSS 17 durchgeführt. Lediglich bei der Berechnung der konfirmatorischen Faktorenanalyse wird das Programm AMOS 17 verwendet. Es wird immer mit einer fünfprozentigen Irrtumswahrscheinlichkeit davon ausgegangen, dass die in den Stichproben entdeckten Effekte in der Grundgesamtheit tatsächlich existieren. Anders formuliert wird die Nullhypothese ab einem Signifikanzniveau von 0,05 verworfen und die getestete Alternativhypothese angenommen. Die in den Tabellen dargestellten Zahlenwerte werden nach mathematischer Konvention jeweils auf zwei Nachkommastellen gerundet. So können zum Beispiel Summen von Anteilswerten vom Sollwert geringfügig abweichen.

Bevor die deskriptiven Statistiken und die Lage-, Streuungs- und Formmaße der einzelnen Items betrachtet und Missing-Value Analysen durchgeführt werden, wurde die Kodierung der Daten betrachtet. So war es notwendig die Items der zweiten Skala, die die Meinung über die Auswirkungen von Immigration misst, umzupolen. Ein Unterschied bei der Kodierung zwischen den Analyseeinheiten konnte ebenfalls entdeckt werden. So wird im deutschen und österreichischen Datensatz zwischen expliziten Antwortverweigerungen und einer „weiß-nicht“-Angaben unterschieden, während im Datensatz Großbritanniens lediglich „weiß-nicht“-Kategorien existieren. Diese Unterscheidungsmöglichkeit hat keinen Effekt auf den Befragten, da dieser Umstand lediglich den Interviewern bekannt war, da auf dem Fragebogen diese Kategorien nicht explizit dem Befragten zur Beantwortung angeboten wurden.

5.3.1. Häufigkeitsauszählungen

Zuerst werden die Häufigkeitsauszählungen der einzelnen Items betrachtet, um festzustellen ob hierbei offensichtliche länderpezifische Unterschiede bezüglich der anteilsmäßigen Zellenbesetzungen entdeckt werden können. (vgl. Tab. A.1-A.6; Anhang:144-146) Zuerst werden die Ergebnisse der Skala betrachtet, die die Einstellung bezüglich Migration misst. Beim Item „Allow immigrants of same race/ethnic group as majority" ist in Großbritannien und in Ostdeutschland eine stärkere Ablehnung gegenüber diesem Migrantentypus festzustellen als in den anderen zwei Analyseeinheiten. In Westdeutschland ist die Zustimmung zu diesem Item hingegen mit Abstand am größten. Rund ein Viertel der Befragten ist der Ansicht, dass es vielen von diesen Migranten erlaubt sein sollte einzuwandern. Bezüglich dieser Migrantengruppe herrscht in allen Analyseeinheiten relativ zu den anderen abgefragten Charakteristiken der Migranten die positivste Einstellung vor. Bei dem Item „Allow immigrants of different race/ethnic group as majority" weist die Einstellung der Westdeutschen ebenso die positivste Ausprägung auf, während die Österreicher bei diesem Item die stärkste negative Einstellung hierzu kundgeben. Der Modalwert in der österreichischen Stichprobe befindet sich bei der Kategorie „es wenigen zu erlauben", während dieser sich bei den anderen Analyseeinheiten bei der Kategorie „es einigen erlauben" liegt. Die Ostdeutschen weisen jedoch den bei weitem größten Anteilswert bezüglich der Befragten auf, die die Einstellung aufweisen, dass es keinem Migranten mit diesen Merkmal erlaubt werden sollte, in ihr Land zu kommen um hier zu leben. Bei dem Item „Allow immigrants from poorer countries outside Europe" sind die anteilsmäßigen Unterschiede zwischen der österreichischen, der westdeutschen und der großbritannischen Stichprobe relativ ähnlich. Jedoch weisen die Befragten aus Westdeutschland erneut die positivste Einstellung zu diesem Item auf, während die Ostdeutschen die mit Abstand negativste Haltung hierzu einnehmen. Zirka ein Viertel der Befragten aus Ostdeutschland ist der Ansicht, dass es keinem Migranten, der aus einem ärmeren Land außerhalb Europas kommt, erlaubt sein sollte nach Deutschland zu imigrieren. Generell kann zu den anteilsmäßigen Zellenbesetzungen der ersten Skala festgestellt werden, dass bei allen Items in allen Ländern auf die zwei mittleren Antwortkategorien jeweils mehr als zwei Drittel der Antworten entfallen. Die extremen

Antwortkategorien wurden bei lediglich zwei Fällen von zirka einem Viertel der Befragten in einem Land ausgewählt.

Nun werden die wichtigsten Erkenntnisse bezüglich der gefundenen Unterschiede und Gemeinsamkeiten zwischen den Analyseeinheiten der anteilsmäßigen Zellenbesetzungen der zweiten Skala betrachtet, die die Meinung bezüglich der Auswirkungen von Migration misst. Bei der Betrachtung der drei Skalenitems fällt auf, dass die neutrale Mittelkategorie bei allen Items in jeder Analyseeinheit die Modalkategorie darstellt. Die Antweilswerte sind hierbei zwischen Ost- und Westdeutschland und Österreich ähnlich und in Großbritannien deutlich niedriger. Bei den ersten beiden Items „Immigration bad or good for country's economy" und „Country's cultural life undermined or enriched by immigrants" haben in den erstgenannten Analyseeinheiten sich zirka ein Viertel der Befragten für die neutrale Mittelkategorie entschieden. Bei dem dritten Item „Country is made a worse or a better place by immigrants" sind dies sogar in etwa ein Drittel aller Befragten. In Großbritannien hingegen hat ein Fünftel beim ersten Item, ein Viertel beim zweiten Item und ein Viertel der Befragten beim dritten Item die Mittelkategorie gewählt. Vielleicht sind die hohen Anteilswerte in der Mittelkategorie bei dem dritten Item darauf zurückzuführen, dass dieses das allgemeinste der drei Items darstellt. Diese allgemeine Formulierung könnte die Befragten davor abgeschreckt haben, klar Stellung zu diesem Thema zu beziehen. Es ist allerdings auch denkbar, dass bei vielen Befragten keine klar ausgeprägte Meinung hierzu vorliegt und die Abgabe eines dementsprechenden Urteils nicht möglich war.

Ob die gefundenen Unterschiede zwischen den Analyseeinheiten einstellungsbedingt oder messbedingt sind, kann zu diesem Zeitpunkt der Analyse noch nicht bestimmt werden. Es wurde allerdings festgestellt, dass Unterschiede bezüglich der Häufigkeitsverteilungen in den Analyseeinheiten vorliegen. So wurden zum Beispiel in Westdeutschland bei der ersten Skala überdurchschnittlich viele positive Einstellungen zu den Migranten angegeben, während sich dies in Ostdeutschland genau umgekehrt darstellt. Hierbei kann es sich um tatsächliche Einstellungsunterschiede handeln. Es kann allerdings auch ein *uniform bias* zwischen Ost- und Westdeutschland vorliegen, der messbedingte Unterschiede zwischen diesen Analyseeinheiten verursacht. (vgl. Kap. 3.1.2) Bei der Betrachtung der Ergebnisse der zweiten Skala fällt hingegen auf, dass die

Meinungen zu den Auswirkungen der Migration zwischen den einzelnen Ländern wesentlich ähnlicher sind als die Einstellungen zur Migration. Besonders auffällig ist hierbei der hohe Anteil an Befragten, die sich für die neutrale Mittelkategorie entschieden haben. Ob dies eine inhaltsunabhängige Antworttendenz darstellt, wird im Kapitel 5.3.5 erörtert.

5.3.2. Verteilungen der Items

Die Überprüfung, ob eine Normalverteilung vorhanden ist, ist durch den Umstand notwendig, dass die zumeist bei der Durchführung einer konfirmatorischen Faktorenanalyse für die Parameterschätzung favorisierte Maximum-Likelihood-Methode eine Normalverteilung der manifesten Variablen eines Modells vorraussetzt. *(vgl. Reinecke 2005:109-110)* Obwohl auf Grund des ordinalen Skalenniveaus eine Normalverteilung grundsätzlich nicht möglich ist, wird trotzdem geprüft, ob die Verteilung der Items in den Analyseeinheiten zumindest annähernd normalverteilt ist. Denn sollte dies der Fall sein, wäre die Durchführung der Parameterschätzung mit der Maximum-Likelihood-Methode vertretbar. (vgl. ebd.) Der hierfür durchgeführte Kolmogorov-Smirnov-Test ergab für jedes Item in jeder Analyseeinheit einen signifikanten Prüfwert. Demzufolge kann also mit einer Irrtumswahrscheinlichkeit von 5 % davon ausgegangen werden, dass diese Items in der Grundgesamtheit nicht normalverteilt sind. Da dieser Test jedoch die Alternativhypothese der Nicht-Normalverteilung in der Grundgesamtheit bereits bei geringen Abweichungen der Verteilung der Items von der Normalverteilung annimmt, werden zudem die Histogramme und die Schiefe- und Wölbungskoeffizienten betrachtet. Diese Kennwerte setzen ebenfalls zumindest intervallskalierte Daten voraus. Da es sich um Daten mit intervallskalierten Charakter handelt und die durch diese Kennwerte gewonnenen Informationen bezüglich der Beschaffenheit der einzelnen Items nicht unwichtig sind, werden diese dennoch betrachtet werden.

Die Betrachtung der Histogramme mit der Normalverteilungskurve zeigt, dass die Verteilungen der Items zumindest unimodal sind und einige einer Normalverteilung ähneln. Dies wird aber besser durch die Darstellung der Schiefe- und Wölbungskoeffizienten sichtbar. (vgl. Tab. A.7+A.8; Anhang:146-147) Die Verteilung

des ersten Items, das Migranten der selben ethnischen Gruppe wie die Mehrheit der Bevölkerung thematisiert, weist im Vergleich zu den anderen Items in allen Analyseeinheiten die größte Abweichung von einer Normalverteilung auf. Das Item weist allerorts eine relativ starke rechtsschiefe Verteilung auf, wobei dies in Westdeutschland besonders stark der Fall ist. Lediglich bei dem fünften Item, das die Auswirkungen von Immigration auf das kulturelle Leben thematisiert, ist in Ost- und Westdeutschland eine hierzu vergleichbare Abweichung von einer Normalverteilung zu beobachten. Dieses Item ist in beiden Fällen stark rechtsschief verteilt. Bezüglich der Verteilungen der Items in den einzelnen Analyseeinheiten zeigt sich, dass in Großbritannien alle Items außer dem ersten Item nur leicht schief verteilt sind. Bis auf das letzte Item, welches leicht linksschief verteilt ist, weisen dort alle Items rechtsschiefe Verteilung auf. Ähnliches ist für Westdeutschland feststellbar, wo allerdings die Items etwas schiefer verteilt sind und lediglich das dritte Item eine leicht linksschiefe Verteilung aufweist. In Österreich und Ostdeutschland sind die Items hingegen überwiegend linksschief verteilt. In Ostdeutschland weisen lediglich das erste und fünfte Item stark rechtsschiefe Verteilungen auf, während dies in Österreich beim ersten und vierten Item der Fall ist.

Die Betrachtung der Wölbungskoeffizienten zeigt, dass die Verteilungskurve bei allen Items zum Teil deutlich flachgipfliger ist, als im Fall einer Normalverteilung. Lediglich das sechste Item weist in Deutschland eine hochgipflige Verteilung auf. Basierend auf diesen Informationen kann letztlich zwar festgestellt werden, dass einige Items in einigen Analyseeinheiten eine Verteilung aufweisen, die annähernd normalverteilt sein dürfte. Jedoch existieren ebenso viele Items, bei denen dies nicht der Fall sein dürfte. Der Umstand, dass keine normalverteilten Daten vorliegen, muss bei der Durchführung der konfirmatorischen Faktorenanalyse berücksichtigt werden. (vgl. Kap. 6.4)

5.3.3. Vergleich der Mittelwerte und Standardabweichungen

Es werden nun die Mittelwerte und Standardabweichungen der einzelnen Items dargestellt, um eventuelle Auffälligkeiten zwischen den Analyseeinheiten zu entdecken. (vgl. Tab. 5) In dieser Tabelle werden zudem die Mittelwerte der auf einen Wertebereich von 0 bis 1 transformierten Items dargestellt. Die Mittelwerte entsprechen der Schwierigkeit eines Items, welche als der Anteil der Personen aufgefasst wird, die einem Item zustimmen beziehungsweise eine positive Haltung zu diesem einnehmen. (vgl. Bortz/Döring 1995:199) Da die beiden Skalen so rekodiert wurden, dass ein niedrigerer Wert einer positiveren Einstellung beziehungsweise Meinung entspricht, bedeutet ein niedriger normierter Mittelwert, dass das Item eine geringe Schwierigkeit aufweist. Da die Mittelwerte und Itemschwierigkeiten dieselbe Information über die Beschaffenheit der Items enthalten, werden im folgenden Teil lediglich erstere miteinander verglichen. Es sind allerdings weniger die absoluten Werte von Bedeutung als die Relation dieser zwischen den einzelnen Items in den Analyseeinheiten. Liegt bei einem einzelnen Item ein höherer Mittelwert als bei den restlichen Items einer Skala vor und ist dies in den anderen Ländern für dieses Item so nicht beobachtbar, kann dies erste Hinweise auf die Existenz einer Verzerrung liefern.

Tabelle 5: Mittelwerte und Standardabweichung der Skalenitems

Land/Item		1	2	3	4	5	6
ESS	Mittelw	2,17 (0,39)	2,52 (0,51)	2,58 (0,53)	5,05 (0,50)	4,52 (0,45)	5,21 (0,52)
	Std.abw	0,88 (0,29)	0,90 (0,30)	0,92 (0,31)	2,48 (0,25)	2,56 (0,26)	2,31 (0,23)
GB	Mittelw	2,39 (0,46)	2,58 (0,53)	2,64 (0,55)	5,46 (0,55)	5,21 (0,52)	5,52 (0,55)
	Std.abw	0,81 (0,27)	0,84 (0,28)	0,86 (0,29)	2,5 (0,25)	2,6 (0,26)	2,4 (0,24)
Ost-D	Mittelw	2,31 (0,44)	2,67 (0,56)	2,79 (0,60)	5,55 (0,56)	4,47 (0,45)	5,80 (0,58)
	Std.abw	0,88 (0,29)	0,87 (0,29)	0,89 (0,3)	2,6 (0,26)	2,5 (0,25)	2,3 (0,23)
West-D	Mittelw	2,09 (0,36)	2,49 (0,50)	2,56 (0,52)	5,13 (0,51)	4,33 (0,43)	5,33 (0,53)
	Std.abw	0,82 (0,28)	0,84 (0,28)	0,88 (0,29)	2,32 (0,23)	2,37 (0,24)	2,13 (0,21)
Ö	Mittelw	2,19 (0,4)	2,61 (0,54)	2,58 (0,53)	4,78 (0,48)	5,20 (0,52)	5,79 (0,58)
	Std.abw	0,79 (0,26)	0,81 (0,27)	0,82 (0,27)	2,46 (0,25)	2,58 (0,26)	2,23 (0,22)

Anmerkung: Gewichtung mit Designgewicht. In den Klammern neben den Mittelwerten und befinden sich dieselben Werte für die auf einen Wertebereich von 0 bis 1 normierten Items. Der Wert 0 entspricht hierbei einer „perfekten“ Zustimmung, also dem geringstmöglichen erreichbaren Schwierigkeitsgrad eines Items. Der Wert 1 entspricht einer „perfekten“ Ablehnung, also dem höchstmöglichen erreichbaren Schwierigkeitsgrad eines Items.
Lesebeispiel: Im Gesamtdatensatz mit allen Ländern des ESS hat das Item 1 einen Mittelwert von 2,17 und eine Standardabweichung von 0,88. Wird dieses Item normiert, weist dieses Item einen Mittelwert beziehungsweise eine Itemschwierigkeit von 0,39 auf.

Die Relationen der Standardabweichungen der einzelnen Items der beiden Skalen sind über alle Analyseeinheiten hinweg ähnlich. (vgl. Tab. 5) Die Betrachtung der Standardabweichungen zeigt, dass diese auch auf eine zwischen den Items der beiden Skalen vergleichbare Streuung der Antworten in den Analyseeinheiten hinweisen. Somit besteht nicht die Gefahr, dass einzelne Items, die eine höhere Streuung aufweisen, die Faktorenlösung der explorativen und konfirmatorischen Faktorenanalyse stärker beeinflussen und somit die Genauigkeit der Parameterschätzungen beeinflussen.

Bei der Betrachtung der Mittelwerte fällt auf, dass die einzelnen Skalenitems in Großbritannien vermutlich als ähnlicher wahrgenommen werden als dies in den anderen Analyseeinheiten der Fall ist. (vgl. Tab. 5) Zudem ist ersichtlich, dass die Relationen der Mittelwerte der Items in Großbritannien, Ost- und Westdeutschland bis auf geringfüge Abweichungen relativ ähnlich sind. In Österreich sind jedoch hiervon unterschiedliche Relationen der Mittelwerte feststellbar. So weist das zweite Item „Allow immigrants of different race/ethnic group as majority" in Österreich einen höheren Mittelwert als das dritte Item „Allow immigrants from poorer countries outside Europe" auf. Dies ist in den restlichen Analyseeinheiten nicht der Fall. Inhaltlich bedeutet dies, dass im Fall der Existenz einer *Measurement unit equivalence*, in Österreich im Gegensatz zu den anderen Ländern eine negativere Einstellung zu Migranten einer anderen ethnischen Gruppe vorherrscht, als zu solchen, die aus ärmeren Ländern außerhalb Europas kommen. Es existiert jedoch noch ein weiterer Unterschied. In Großbritannien, Ost- und Westdeutschland weist das fünfte Item „Country's cultural life undermined or enriched by immigrants" den niedrigsten Mittelwert aller Items der zweiten Skala auf, während dies in Österreich für das vierte Item „Immigration bad or good for country's economy" zutreffend ist. Dieser Umstand wird als ein Anzeichen gewertet, dass in Österreich das Item „Country's cultural life undermined or enriched by immigrants" eine zu den restlichen Analyseeinheiten unterschiedliche Bedeutung für die Ausprägung der gemessenen Einstellung beziehungsweise Meinung zur Migration aufweisen. Dieses Item erscheint nämlich generell problematisch zu sein. Ein Vergleich zwischen Großbritannien, West- und Ostdeutschland zeigt nämlich, dass auch hier die Relationen der Mittelwerte dieses Items zu denen der restlichen Items unterschiedlich sind.

Dies wirft die Frage auf, ob diese entdeckten Abweichungen bezüglich der Verteilungen der genannten Items zwischen den einzelnen Ländern nun Anzeichen für eine mangelnde Äquivalenz der betreffenden Items oder tatsächlich existierende Länderunterschiede darstellen. Diese Frage kann letztlich nur durch die Verfahren, die auf Korrelationsmuster der Items basieren, mit Sicherheit beantwortet werden. Es ist auf Grund der gefundenen Erkenntnisse aber zumindest zu vermuten, dass bei dem Item „Country's cultural life undermined or enriched by immigrants" ein *non uniform Item bias* und dadurch keine *Measurement unit equivalence* zwischen allen Analyseeinheiten vorliegt. Denn es scheint so zu sein, dass lediglich dieses Item in verschiedenen Gruppen eine unterschiedliche Bedeutung besitzt. Dies beeinträchtigt nicht die funktionale Äquivalenz des Konstrukts Meinung bezüglich der Auswirkungen von Immigration und somit die Verhinderung einer gemeinsamen Grundlage, die für einen Vergleich notwendig ist. Es dürfen bei Vorliegen eines *Item bias* aber keine direkten Vergleiche von Häufigkeiten getätigt werden, da keine identischen Skaleneinheiten für dieses Item in den betreffenden Analyseeinheiten existent sind. Unter der Vorraussetzung, dass zumindest eine funktionale Äquivalenz der Skalen vorliegt, dürfen lediglich Korrelationskoeffizienten und Faktorenstrukturen verglichen werden.

5.3.4. Analyse der fehlenden Werte

In der folgenden Tabelle sind die Anteile der fehlenden Werte bei den Items in den Analyseeinheiten sowie der Anteil der Befragten dargestellt, die bei mindestens einem Item der ersten und der zweiten Skala die Antwort explizit verweigert oder angegeben haben, keine Antwort zu wissen. Eine Unterscheidung zwischen diesen beiden Gründen nicht auf ein Item zu antworten, ist wie bereits dargestellt, in Großbritannien nicht möglich. Ein Vergleich der fehlenden Werte ist sowohl über die Analyseeinheiten hinweg als auch innerhalb der einzelnen Analyseeinheiten sinnvoll. So kann zum Beispiel ein höherer Anteil an fehlenden Werten bei einem Item innerhalb eines Landes als Anzeichen dafür gewertet werden, dass dieses Item eine höhere Sensitivität als die restlichen Items aufweist.

In Österreich sind bei allen Items im Vergleich mit den anderen Analyseeinheiten die mit Abstand höchsten Anteile an fehlenden Werten feststellbar. (vgl. Tab. 6) Der Anteil

ist bei manchen Items sogar doppelt so hoch, wie in Westdeutschland, wo ebenfalls verglichen mit Großbritannien und Ostdeutschland hohe Anteilswerte vorliegen. Bei den zuletzt genannten Analyseeinheiten sind die Anteile der fehlenden Werte insgesamt relativ niedrig und zudem auf einem vergleichbaren Niveau. Die erste Skala weist über alle betrachteten Einheiten hinweg höhere Anteilswerte bei den expliziten Antwortverweigerungen auf als die zweite Skala. Besonders stark ausgeprägt ist dies in Westdeutschland der Fall. Lediglich beim letzten Item der zweiten Skala, welches die Meinung erfasst, ob durch Migranten das jeweilige Land ein besserer oder schlechterer Ort zum Leben wird, sind ähnliche Anteilswerte bei den expliziten Antwortverweigerungen vorhanden, wie bei den Items der ersten Skala.

Dieser Unterschied zwischen den beiden Skalen kann mehrere Ursachen haben. So ist es möglich, dass die erste Skala eine höhere Sensitivität aufweist als die zweite Skala. Dies kann durch die Art der Fragestellung beziehungsweise der Thematik verursacht sein. Die Einstellung zur Migration wurde mittels einer indirekten Fragestellung erfasst, während die Meinung zu den Auswirkungen direkt erfragt wurde. Bei der ersten Skala wurden die Respondenten gefragt, in welcher Quantität es Personen mit verschiedenen Merkmalen erlaubt sein sollte einzuwandern.[55] Die angegebene Menge entspricht der Einstellung zu den Migranten mit diesem Merkmal. Es erscheint letztlich einfacher und weniger „belastend“ für die Befragten zu sein, Meinungen bezüglich der Auswirkungen von Migration preiszugeben, als zu „entscheiden“ wie vielen Zuwanderern es erlaubt sein sollte in ein Land einzuwandern. Die höheren Anteilswerte der ersten Skala bei den expliziten Antwortverweigerungen können aber auch durch die unterschiedliche Struktur der Antwortskalen verursacht sein. (vgl. Kap. 5.2) Denn die zweite Skala enthält im Gegensatz zur ersten einen Skalenmittelpunkt. Es existiert also eine neutrale Mittelkategorie, so dass Unentschlossene oder Meinungslose nicht zur Abgabe einer inhaltlichen Antwort „gezwungen“ werden. Der Umstand, dass die „weiß-nicht“-Kategorie den Befragten nicht explizit zur Beantwortung angeboten wird, verschärft diesen Unterschied noch zusätzlich. Welche dieser vermuteten Ursachen schließlich zutreffend ist, kann nicht mit Sicherheit festgestellt werden. Für die Beantwortung dieser Frage empfiehlt sich die Durchführung von Methodenexperimenten, bei denen unter anderem die Antwortskalen variiert werden müssten.

[55] So lautete zum Beispiel die Fragestellung in der österreichischen Fragebogenversion: „Wie vielen von ihnen [den Zuwanderern] sollte es Österreich erlauben, sich hier niederzulassen?“

Tabelle 6: Anteile der fehlenden Werte nach Item und Skala

in %	ESS	GB	Ost-D	West-D	Ö
1	3,2 (0,2)	1,3	1 (0,4)	2,5 (1,2)	4,7 (0,5)
2	3,5 (0,2)	1,2	1,1 (0,4)	3,1 (1,1)	4,5 (0,5)
3	4,3 (0,2)	1,3	1,6 (0,3)	2,6 (0,9)	5,9 (0,7)
4	5,7 (0,1)	1,9	1,8 (0,1)	5,8 (0,4)	5,5 (0,2)
5	5,5 (0,1)	1,8	1,8 (0,1)	2,7 (0,3)	4,8 (0,2)
6	5,7 (0,1)	1,3	1,8 (0,3)	3,6 (0,7)	4,6 (0,6)
mind. 1 Missing bei Items 1-3	5,6	2,4	1,8	3,7	7,2
mind. 1 Missing bei Items 4-6	10,3	3,9	4,7	8,1	10
mind. 1 Missing bei Items 1-6	13,2	5,6	6,2	11,4	14,5

Anmerkung: Gewichtung mit Designgewicht, Angabe des Anteils der fehlenden Werte an der Anzahl der gültigen Interviews, die Werte in Klammern entsprechen dem Anteil der expliziten Verweigerungen, Die Differenz zwischen den beiden Werten in einer Zelle ist der Anteil der „weiß-nicht"-Angaben an der Anzahl der gültigen Interviews.
Lesebeispiel: In der österreichischen Befragung haben zirka 15 % der Befragten auf zumindest ein Item der beiden Skalen die Antwortkategorie „weiß-nicht" gewählt oder eine Antwortabgabe verweigert.

Das vierte Item, das die Meinung der Befragten erfasst, ob Immigration gute oder schlechte Auswirkungen auf die Ökonomie hat, weist in allen Analyseeinheiten die höchsten Anteile der „weiß-nicht"-Antworten auf. Dies ist ein Anzeichen dafür, dass eine diesbezügliche Abschätzung der Auswirkungen für die Befragten sich als schwierig erwiesen hat. Dieses Item dürfte also das schwierigste Item der beiden Skalen darstellen. Dies ist jedoch nicht mit der Bedeutung der zuvor dargestellten Itemschwierigkeit gleichzusetzen, sondern mit der Schwierigkeit der Urteilsbildung zu einem Sachverhalt. Es ist zudem interessant, dass das dritte Item, welches die Einstellung gegenüber Migranten aus ärmeren Ländern außerhalb Europas erfasst, in Österreich den relativ zu den anderen Items höchsten Anteil an fehlenden Werten aufweist. In den anderen Analyseeinheiten sind bei diesem Item vergleichsweise niedrige Werte zu beobachten. Im vorhergehenden Kapitel wurde bereits festgestellt, dass die Mittelwerte des dritten und vierten Items in Österreich nicht mit denen der anderen Analyseeinheiten vergleichbar sind. (vgl. Tab. 5) Deswegen wurde geprüft, ob in Österreich Zusammenhänge zwischen der Ausprägung der Einstellung zur Migration beziehungsweise der Ausprägung der Meinung zu den Auswirkungen der Migration und dem Umstand vorliegt, dass auf diese Items eine Anwortabgabe verweigert wurde. Es können jedoch solche Zusammenhänge und somit die Existenz von systematischen

Verzerrungen dieser Items durch einen *Item-Nonresponse error* ausgeschlossen werden.[56]

Die Feststellung des Anteils der Befragten die auf die erste oder die zweite Skala oder über beide Skalen hinweg betrachtet keine gültige Antwort gegeben haben, hat hingegen den Zweck festzustellen, wie stark sich die Fallzahl bei einem listenweisen Ausschluss der Fälle verringert. Dieser listenweise Ausschluss wird im weiteren Verlauf dieser Arbeit zum Beispiel bei der Bestimmung der internen und externen Konsistenz der Skalen, sowie der explorativen und konfirmatorischen Faktorenanalyse angewandt. Es werden also die Befragten, die mindestens einen fehlenden Wert bei einem Item aufweisen in diesen Analysen nicht beachtet, um konstante Fallzahlen und bessere Parameterschätzungen bei diesen Verfahren zu erhalten. Bei der Betrachtung dieser Anteilswerte bestätigen sich die bereits zuvor getätigten Feststellungen. (vgl. Tab. 6) So zeigt sich, dass bei einem listenweisen Ausschluss der Fälle mit einem fehlenden Wert sich in Österreich die Fallzahl bei der weiteren Analyse der beiden Skalen um zirka 15 % verringert, in Westdeutschland um ungefähr 11 % und in Ostdeutschland und Großbritannien um jeweils etwa 6 %. Zudem wird sichtbar, dass der Anteil der Personen die zumindest einen fehlenden Wert aufweisen bei der zweiten Skala bedeutend höher ausfällt als bei der ersten Skala. So weisen zirka 7 % der Befragten einen fehlenden Wert bei der ersten Skala in Österreich auf, während dies in Westdeutschland bei etwa 4 % und in Ostdeutschland und Großbritannien bei lediglich 2 % der befragten Personen zutreffend ist. Bei der zweiten Skala trifft selbiges auf 10 % der Befragten in Österreich, auf 8 % in Westdeutschland, auf 5 % in Ostdeutschland und auf 4 % der Befragten in Großbritannien zu. Die Anteile der fehlenden Werte bei den zwei Skalen weisen also große Unterschiede zwischen den Untersuchungseinheiten auf. Warum diese Anteile zwischen den Analyseeinheiten variieren, kann nicht festgestellt werden. Die Unterschiede zwischen den Skalen können jedoch ihre Ursache in der Art der Itemformulierungen bei der Skala zur Meinungsmessung zu den Auswirkungen von Migration haben.[57] Diese ähneln nämlich in gewisser Weise Wissensfragen und es wäre möglich, dass dies einige Befragte von einer Urteilsbildung

[56] Die Berechnungen ergaben jeweils ein Lambda von Null, das sich als nicht signifikant erwiesen hat.

[57] So lautete die Formulierung des vierten Items: „Was würden Sie sagen, ist es im Allgemeinen gut oder schlecht für die deutsche Wirtschaft, dass Zuwanderer hierher kommen?“ Die restlichen Items der zweiten Skala weisen ähnliche Formulierungen hierzu auf.

beziehungsweise von der Bekanntgabe dieses Urteils abgeschreckt hat.[58] Diese Unterschiede bei den Anteilen der fehlenden Werte zwischen den Analyseeinheiten und zwischen den beiden Skalen sind jedoch nicht problematisch, solange diese Unterschiede zufällig zustande kommen.

Um sicherzustellen, dass diese unterschiedliche Verringerung der Fallzahl in den einzelnen Analyseeinheiten nicht zu einer Verzerrung der mit diversen statistischen Verfahren gewonnenen Erkenntnissen führt, wurde geprüft, ob ein Zusammenhang zwischen der Anzahl der fehlenden Werte der Befragten bei beiden Skalen und diversen sozialstatistischen Merkmalen besteht.[59] (vgl. Tab. A.9 Anhang:147) Ein listenweiser Ausschluss von Fällen mit einem fehlenden Wert setzt nämlich voraus, dass das Vorliegen eines fehlenden Wertes zufällig erfolgt und nicht mit einer anderen Eigenschaft der Befragten assoziiert ist. Sonst wäre ein alternativer Umgang mit diesen fehlenden Werten, wie zum Beispiel die Imputation des Mittelwertes im Fall eines fehlenden Wertes, empfehlenswert. Die Betrachtung von Spearmans Rangkorrelationskoeffizienten zeigt, dass in Großbritannien, Ostdeutschland und Österreich ein ganz schwacher Zusammenhang zwischen dem Alter und der Anzahl der fehlenden Werte besteht. Diese Korrelationen sind in Großbritannien und Ostdeutschland auf einem 0,05 Niveau signifikant, während die Korrelation in Österreich auf einem 0,01 Niveau signifikant ist.

Es besteht weiters eine schwache Korrelation zwischen der Anzahl der fehlenden Werte und der höchsten erreichten Schulbildung einer Person in Großbritannien, Westdeutschland und Österreich. Diese Korrelationen sind alle auf dem 0,01 Niveau signifikant. Zwischen dem verfügbaren Einkommen pro Kopf beziehungsweise dem Geschlecht der Befragten und der Anzahl der fehlenden Werte dieser ist hingegen von einer statistischen Unabhängigkeit in der Grundgesamtheit der vier Analyseeinheiten auszugehen. Da die berichteten Zusammenhänge allesamt ganz schwach ausfallen, können statistische Verfahren angewandt werden, bei denen ein listenweiser Ausschluss der Fälle sinnvoll erscheint, ohne dass eine starke Verzerrung der Ergebnisse zu erwarten ist.

---

[58] Wissensfragen werden in der Literatur auch mitunter als Überzeugungsfragen bezeichnet. (vgl. Schnell/Hill/Esser 2008:326)

[59] Es werden deswegen sozialstatistischen Merkmale wie Alter, Bildung, Einkommen und Geschlecht gewählt, weil davon auszugehen ist, dass diese in einem geringeren Ausmaß fehlerbehaftet sind, als dies bei Einstellungsfragen der Fall ist.

5.3.5. Betrachtung möglicher Antworttendenzen

In diesem Kapitel wird dargestellt, welcher Anteil an den befragten Personen über alle Items der ersten und der zweiten Skala beziehungsweise bei beiden Skalen ausschließlich dieselben Antwortkategorien ausgewählt hat**.** Da beide Skalen jeweils aus lediglich drei Items bestehen, waren die Items der jeweiligen Skala alle in dieselbe Richtung gepolt. Gemäß der ursprünglichen Kodierung entspricht einem niedrigen Wert bei der ersten Skala eine positive Einstellung, während bei der zweiten Skala ein hoher Wert mit einer positiven Meinung zu den Auswirkungen von Migration für das Einwanderungsland gleichzusetzen ist. Ob bei den befragten Personen, die bei einem jeden Item dieselbe Antwortkategorie gewählt haben, nun eine inhaltsunabhängige Antworttendenz vorliegt, bleibt aber letztlich auf Grund der Skalenkonstruktion unklar. Um eine Antworttendenz bei Befragten festzustellen, wäre es notwendig, dass die Items der Skalen in eine unterschiedliche Richtung gepolt sind. (vgl. Diekmann 2004: 386-389) Dies stellt allerdings kein spezifisches Problem eines Vergleichs dar, sondern zeigt, dass die Fragebogenkonstruktion durchaus verbesserungswürdig wäre. Trotzdem werden die Anteile der Personen, die jeweils konsistente Antworten abgegeben haben, zwischen den Analyseeinheiten verglichen, da länderspezifische Unterschiede auf eine mögliche Verzerrung des Vergleichs hinweisen können. (vgl. Tab. 7)

Tabelle 7: Betrachtung möglicher Antworttendenzen

in %	ESS	GB	Ost-D	West-D	Ö
Skala 1					
Bei allen allow many	10	6,2	5,8	9	7,1
Bei allen allow none	6,8	7,6	8,7	4,3	4
Skala 2					
Bei allen Items positivste Kategorie	1,2	1,2	1,4	0,9	1,1
Bei allen Items mittlere Kategorie	5,3	4,5	5,8	6	7
Bei allen Items negativste Kategorie	2,5	4,1	3,3	1,7	2,5
Beide Skalen					
Bei Items allen positivste Kategorie	0,7	0,7	1,2	0,7	1
Bei allen Items negative Kategorie	1,1	2,2	2,0	0,9	1

Anmerkung: gewichtet mit Gewichtung mit Designgewicht; Angabe des Anteils der Antwortmuster an der Anzahl der gültigen Interviews
Lesebeispiel: In Österreich haben 7,1 % der Befragten bei allen 3 Items der ersten Skala, die positivste Antwortkategorie „es vielen erlauben“ gewählt.

Bei Betrachtung der Tabelle fällt auf, dass die Anteile der Befragten, die bei der ersten Skala, die die Einstellung zur Migration erfasst, immer eine der beiden extremen Antwortkategorien gewählt haben, zwischen den einzelnen Analyseeinheiten in etwa vergleichbar sind. Ähnliches zeigt sich auch bei der zweiten Skala, welche die Meinung zu den Auswirkungen von Migration erfasst, für die positivste Kategorie. Die Prozentsätze der Befragten, die sich bei bei allen Items der zweiten Skala immer für die neutrale Mittelkategorie oder für die negativste Kategorie entschieden haben, unterscheiden sich jedoch zwischen den Analyseeinheiten. Besonders auffällig sind die hohen Anteilswerte der Personen, die sich bei einer elfstufigen Skala für die Mittelkategorie entschieden haben. Sollte dies eine Antworttendenz darstellen, wären in Zusammenhang mit dem ebenfalls hohen Anteil an fehlenden Werten Verzerrungen bei dieser Skala nicht auszuschließen.

6. Korrelationsbasierte Verfahren zur Überprüfung der erreichten Ebene der Äquivalenz

Für eine erste Überprüfung der Ähnlichkeit der Items zueinander werden die Interkorrelationen dieser betrachtet. Dies stellt die explorative Vorstufe der Prüfung der internen Konsistenz der Skalen dar. Jedes einzelne Item wird mit jedem anderen Item korreliert. Sollte sich zeigen, dass die relationalen Muster dieser Zusammenhänge zwischen den Analyseeinheiten vergleichbar sind, stellt dies ein Anzeichen für die Existenz einer funktionalen Äquivalenz dar. Es werden zudem auch die Items der zwei unterschiedlichen Skalen miteinander korreliert, da anzunehmen ist, dass die Einstellung und Meinung zu den Auswirkungen der Migration nicht unabhängig voneinander sind.[60] Für die Festellung der Interitemkorrelationen wurde Spearmans Rangkorrelationskoeffizient verwendet, wobei sich alle Zusammenhänge als hochsignifikant erwiesen haben.

Die Korrelationskoeffizienten der Interitemkorrelationen bei der Skala zur Messung der Einstellung zur Migration variieren in den Analyseeinheiten von 0,58 bis 0,78. (vgl. Tab. A.10-14, Anhang:148-149) Diese starken Zusammenhänge lassen auf eine gute interne Konsistenz der Skala schließen. Wichtiger für die Festellung der erreichten Ebene der Äquivalenz dieser Skala ist jedoch der Umstand, dass die Relation der Stärke der Zusammenhänge zwischen den Items in allen Ländern vergleichbar ist. Dies ist ein erstes Anzeichen für die Existenz einer vergleichbaren Skalenstruktur in den Analyseeinheiten. Anders formuliert ist es wahrscheinlich, dass das Konstrukt Einstellung zur Migration in den betrachteten Ländern eine vergleichbare Beschaffenheit aufweist, also funktional äquivalent ist. Das zweite Item der ersten Skala scheint in allen Analyseeinheiten das zentrale Item dieser Skala zu sein, da es die höchsten Korrelationskoeffizienten zu den anderen beiden Items der Skala aufweist.[61] Inhaltlich ist es durchaus nachvollziehbar, dass das zweite Item mit dem dritten Item überall den stärksten Zusammenhang offenbart. Denn wenn ein Befragter eine negative

[60] Die Festellung, ob zwischen den zwei Konstrukten ein Zusammenhang besteht, ist auch für die Durchführung der explorativen und konfirmatorischen Faktorenanalyse von Bedeutung. (vgl. Kap. 6.3+6.4)

[61] Unter Zentralität ist die Wichtigkeit zu verstehen, die ein Item für die Beschaffenheit eines Konstrukts aufweist. Das zentrale Item einer Skala ist jenes Item, bei dem die größte Übereinstimmung zwischen der Ausprägung der Befragten und der Ausprägung bezüglich des betreffenden Konstrukts vorherrscht.

Einstellung gegenüber Migranten hat, die einer anderen ethnischen Gruppe als der Mehrheit der Bevölkerung angehören, wird dieser mit einer hohen Wahrscheinlichkeit ebenfalls eine negative Einstellung gegenüber Migranten aus ärmeren Ländern außerhalb Europas aufweisen. Dies ist dadurch begründbar, dass Migranten aus diesem Ländertypus zumeist auch einer anderen ethnischen Gruppe angehören.

Bei der Skala zur Messung der Meinung der Befragten zu den Auswirkungen von Migration weisen die Interitemkorrelationen in den Analyseeinheiten Werte zwischen 0,52 und 0,77 auf, was eine gute interne Konsistenz dieser Skala indiziert. (vgl. Tab. A.10-14) Bei der zweiten Skala scheint das sechste Item, welches die Meinung erfasst, ob durch Migranten das jeweilige Land ein besserer oder schlechterer Ort zum Leben wird, das zentrale Item zu sein. Dieses Item weist nämlich die höchsten Korrelationskoeffizienten mit den anderen Items dieser Skala auf. Das ist nachvollziehbar, da dieses Item das Allgemeinste der drei Items darstellt. Wenn Befragte der Meinung sind, dass Migration negative Auswirkungen auf das Einwanderungsland hat, wird mit einer hohen Wahrscheinlichkeit auch die Meinung vorherrschen, dass dies die Qualität des Lebensortes beeinträchtigt. Bis auf eine Ausnahme ist die Struktur der Zusammenhänge zwischen den Items dieser Skala in den Analyseeinheiten ähnlich. In drei der vier Analyseeinheiten weist die Meinung der Befragten bezüglich der Auswirkungen von Migration auf die Qualität des Lebensortes mit der Meinung bezüglich der Auswirkungen von Migration auf das kulturelle Leben überall den stärksten Zusammenhang auf. Lediglich in Ostdeutschland besteht zwischen der Meinung der Befragten zu den ökonomischen Auswirkungen und den Auswirkungen auf das kulturelle Leben von Migration ein jeweils gleich starker Zusammenhang mit der Evaluation bezüglich der Auswirkungen auf die Qualität des Lebensortes. Da dieser Unterschied zwischen Ostdeutschland und den restlichen Untersuchungseinheiten jedoch nicht groß ist, ist auch bei dieser Skala vorerst von einer funktionalen Äquivalenz auszugehen. Zudem ist zu diesem Zeitpunkt der Analyse noch nicht auszuschließen, dass dieser Unterschied zwischen den Analyseeinheiten realiter existent ist und keine methodischen Ursachen hat.

Nun wird noch auf die Fragestellung eingegangen, bis zu welchem Grad die Items der zwei Skalen untereinander korrelieren. Die Skalenitems der ersten und der zweiten Skala weisen starke Zusammenhänge auf. (vgl. Tab. A.10-14, Anhang) Die

Koeffizienten reichen in den einzelnen Analyseeinheiten von 0,4 bis 0,59. Die Korrelationsstruktur ist hierbei wieder in den einzelnen Gruppen sehr ähnlich. So zeigt sich, dass die zwei Items, die zuvor als die zentralen Items der jeweiligen Skalen identifiziert wurden, also das zweite und das sechste Item, in allen Analyseeinheiten die größten Zusammenhänge mit den Items der anderen Skalen aufweisen. Der Zusammenhang zwischen diesen beiden Items ist stärker als so manche Korrelation zwischen den eigentlich zur selben Skala zugehörigen Items. Es zeigt sich also, dass diese Skalen zwar zwei unterschiedliche Konstrukte messen, die beiden Einflussfaktoren allerdings realiter eine starke und zudem eine zwischen den Analyseeinheiten ähnliche Korrelation aufweisen. Der Umstand, dass auch die relationalen Muster der Zusammenhänge zwischen den Einstellungs- und Meinungsitems in den Analyseeinheiten ähnlich sind, verstärkt die Vermutung, dass die Skalen funktional äquivalent sind.

Die Betrachtung der Interkorrelationen der Items hat gezeigt, dass eine vergleichbare Struktur der Skalen in allen Ländern existiert. Dies ist ein Anzeichen für die Existenz einer funktionalen Äquivalenz der mit den beiden Skalen gemessenen Konstrukte und somit für die Existenz einer gemeinsamen Basis, die für einen Vergleich zwischen den Analyseeinheiten notwendig ist. In den folgenden Kapiteln wird diese Feststellung insbesonders mittels der Durchführung der explorativen und konfirmatorischen Faktorenanalyse überprüft. Zudem wurde festgestellt, dass die Einstellung zur Migration und die Meinung der Befragten zu den Auswirkungen der Migration für das Einwanderungsland nicht unabhängig voneinander sind. Es kann sogar von einem starken Zusammenhang zwischen diesen beiden Dimensionen ausgegangen werden. Dies wird bei der explorativen Faktorenanalyse berücksichtigt, in dem eine oblique Rotation der hinter den Skalen stehenden Einflussfaktoren vorgenommen wird. (vgl. Kap 6.3)

6.1. Interne Konsistenz der Skalen

Nun wird die interne Konsistenz der beiden Skalen berechnet. So wird geprüft, ob die Reliabilität der Skalen zwischen den einzelnen Analyseeinheiten vergleichbar ist, beziehungsweise ob Items existieren, deren Ausschluss sich positiv auf die Reliabilität auswirkt.

Tabelle 8: Interne Konsistenz der Skala Einstellung zu Migranten

	ESS		GB		Ost-D		West-D		Ö	
Cronbachs α	0,87		0,895		0,873		0,868		0,877	
Items	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted
1	0,68	0,88	0,77	0,87	0,72	0,85	0,67	0,88	0,70	0,88
2	0,83	0,75	0,85	0,80	0,82	0,76	0,83	0,74	0,82	0,77
3	0,75	0,82	0,77	0,88	0,73	0,85	0,75	0,81	0,77	0,82

Anmerkung: Gewichtung mit Designgewicht.
Lesebeispiel: In Großbritannien weist diese Skala ein Cronbachs Alpha von 0,9 auf. Das erste Item weist eine Trennschärfe von zirka 0,8 auf und Cronbachs Alpha würde sich bei einem Ausschluss dieses Items aus der Skala nicht weiter verbessern.

Der Cronbach-Alpha-Koeffizent weist in allen Analyseeinheiten Werte zwischen 0,87 und 0,9 auf, was auf eine gute interne Konsistenz der ersten Skala schließen lässt. (vgl. Tab. 8) Den höchsten Grad der internen Konsistenz weist hierbei die Skala in Großbritannien auf, was angesichts der hohen Korrelationskoeffizienten der Items zu vermuten war. Zusätzlich zeigt die Betrachtung der Trennschärfekoeffizienten, dass sich die anfängliche Vermutung, dass das zweite Item in allen Analyseeinheiten das zentrale Item der zweiten Skala darstellt, als zutreffend erwiesen hat. Das erste Item weist in überall den niedrigsten Trennschärfekoeffizienten aller Items auf. Dies war auf Grund der in diesen Analyseeinheiten existierenden niedrigen Itemschwierigkeit und der niedrigen Interitemkorrelationen dieses Items zu erwarten. In Großbritannien und Ostdeutschland weist das erste Item „Allow immigrants of same race/ethnic group as majority" einen vergleichbaren Trennschärfekoeffizienten, wie das dritte Item der Skala auf. Es zeigt sich zudem, dass eine geringfügig höhere interne Konsistenz der Skalen in Westdeutschland und Österreich erreichbar ist, wenn das erste Item aus der Skala ausgeschlossen wird. Da dies eben nur eine geringfügige Verbesserung darstellt und die erreichte Reliabilität der Skalen für diese geringe Anzahl der Skalenitems an sich exzellent ist, wird in der weiteren Analyse die Skala in ihrer bestehenden Form

weiterverwendet. Durch den Umstand, dass die interne Konsistenz der Skalen überall gut ist und auch die zuvor betrachteten Interkorrelationen der Items keine nennenswerten Unterschiede zwischen den Analyseeinheiten aufgewiesen haben, ist das Vorliegen eines *Construct bias* unwahrscheinlich. Dies stellt eher ein Anzeichen dafür dar, dass bei dem Item „Allow immigrants of same race/ethnic group as majority" in Österreich und Westdeutschland im Vergleich zu Großbritannien und Ostdeutschland unterschiedliche Skaleneinheiten vorliegen. Dies hat einen Einfluss auf die betreffende Itemschwierigkeit und kann einen *Item bias* verursachen. Diesem Item muss also in den noch durchzuführenden Analysen eine besondere Aufmerksamkeit gewidmet werden.

Tabelle 9: Interne Konsistenz der Skala Meinung zu den Auswirkungen von Immigration

	ESS		GB		Ost-D		West-D		Ö	
Cronbachs α	0,845		0,888		0,847		0,821		0,852	
Items	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted	Trenn-schärfe	Alpha if deleted
4	0,68	0,81	0,74	0,87	0,69	0,81	0,63	0,80	0,69	0,83
5	0,71	0,79	0,77	0,85	0,71	0,80	0,68	0,75	0,74	0,78
6	0,74	0,76	0,83	0,80	0,76	0,75	0,72	0,71	0,75	0,77

Anmerkung: Gewichtung mit Designgewicht
Lesebeispiel: In Großbritannien weist diese Skala ein Cronbachs Alpha von 0,89 auf. Das erste Item weist eine Trennschärfe von zirka 0,7 auf und Cronbachs Alpha würde sich bei einem Ausschluss dieses Items aus der Skala nicht weiter verbessern.

Die zweite Skala weist ebenfalls eine gute interne Konsistenz auf. Denn der Cronbach-Alpha-Koeffizent variiert in den Untersuchungseinheiten in einem Wertebereich von 0,82 bis 0,89. (vgl. Tab. 9) Bis auf Großbritannien, wo die interne Konsistenz der beiden Skalen ähnlich ist, ist die interne Konsistenz der zweiten Skala durchwegs deutlich niedriger als die der ersten Skala. Dieser Umstand ist besonders stark in Westdeutschland ausgeprägt und auf den niedrigen Trennschärfekoeffizienten des vierten Items zurückzuführen. Dieses Item weist in allen Analyseeinheiten die geringste Korrelation mit der Gesamtskala auf. Es zeigt sich außerdem, dass in allen Fällen keine höhere interne Konsistenz der Skalen bei Ausschluss eines Items aus der Skala erreicht würde. Die anfängliche Vermutung, dass das sechste Item in allen Analyseeinheiten das zentrale Item der zweiten Skala darstellt, wird bestätigt. Lediglich in Österreich weist das fünfte Item „Country's cultural life undermined or enriched by immigrants" eine ähnliche Korrelation mit der Gesamtskala auf, wie das sechste Item. Dies ist ein Indiz dafür, dass das fünfte Item in Österreich von zentralerer Bedeutung für das zu messende

Konstrukt ist, als dies in den anderen Gruppen der Fall ist. Dies weist auf die Existenz eines *Item bias* in Österreich hin, zumal schon bei der Betrachtung der Mittelwerte eine unterschiedliche Itemschwierigkeit dieses Items festgestellt wurde. (vgl. 5.3.3)

Es hat sich gezeigt, dass die interne Konsistenz der zwei Skalen in allen Analyseeinheiten gut ist. Jedoch wurden auch Hinweise für die Existenz unterschiedlicher Skaleneinheiten bei zwei Items entdeckt. Beim ersten Item, das die Einstellung der Befragten gegenüber Migranten derselben ethnischen Gruppe misst, könnte in Österreich und Westdeutschland im Vergleich zu Großbritannien und Ostdeutschland ein *Item bias* vorliegen. Selbiges wurde beim fünften Item, welches die Meinung bezüglich der Auswirkungen von Migration auf das kulturelle Leben misst, für Österreich im Vergleich zu den restlichen Analyseeinheiten festgestellt. Sollte sich im weiteren Verlauf der Analyse zeigen, dass bei diesen Items wirklich ein *Item bias* und dadurch keine *Measurement unit equivalence* vorliegt, dürfen zum Beispiel keine Häufigkeiten zwischen den Analyseeinheiten direkt verglichen werden. Jedoch ist die Feststellung des Vorliegens eines *Item bias* und somit der Befund, dass es sich bei den gefundenen Unterschiede mit Sicherheit um keine tatsächlichen Länderunterschiede handelt, erst nach der Durchführung der konfirmatorischen Faktorenanalyse möglich.

6.2. Externe Konsistenz der Skalen

Es wird nun die externe Konsistenz der Skalenitems in den Analyseeinheiten anhand der Korrelationen mit der Kriteriumsvariable Alter überprüft. Dies ist jedoch nicht mit einer Prüfung der Kriteriumsvalidität der Skala gleichzusetzen. Es ist kein theoretisch fundiertes Kriterium zu den Einstellungen gegenüber Migration beziehungsweise Meinung bezüglich der Auswirkungen der Migration bekannt, das im ESS enthalten wäre. Zudem erscheint es schwierig zu sein, die Kriteriumsvalidität für Skalen in einer international vergleichenden Perspektive festzustellen. Es kann nämlich nicht davon ausgegangen werden, dass ein als valide geltendes Kriterium in allen Ländern eine ähnliche Beziehung zu den mit den Skalen gemessenen Konstrukten aufweist. Demzufolge ist eine Variation der Zusammenhänge zwischen den einzelnen Indikatoren beziehungsweise der Skalensummenwerte und dem Alter der befragten Personen zwischen den Analyseeinheiten zu erwarten und nicht aussagekräftig für die Äquivalenz

der Skalen. (vgl. Braun 2000) Bei der Prüfung der externen Konsistenz der Skalen geht es, ähnlich wie bei der bereits durchgeführten bivariaten Korrelationsanalyse darum herauszufinden, ob die einzelnen Items der Skalen zu den restlichen Items in den Analyseeinheiten ähnliche Korrelationsmuster mit dem Alter der Befragten aufweisen. Die funktionale Äquivalenz dieser Drittvariablen stellt die Vorraussetzung dar, dass diese Überprüfung einen Aufschluss über die Äquivalenz der Indikatoren geben kann. (vgl. ebd.) Deswegen wurde als Kriteriumsvariable auch das Alter gewählt.

Tabelle 10: Überprüfung der externen Konsistenz: Korrelation der Skalenitems mit Alter

	ESS	GB	Ost-D	West-D	Ö
Skala 1 gesamt	0,16**	0,21**	0,10**	0,14**	0,11**
Skala 2 gesamt	0,10**	0,17**	0,07*	0,14**	0,05*
Beide Skalen zusammen	0,15**	0,21**	0,10**	0,15**	0,09**
allow immigrants of same race/ethnic group as majority	0,10**	0,15**	0,04	0,05*	0,05**
allow immigrants of different race/ethnic group as majority	0,15**	0,19**	0,11**	0,14**	0,09**
allow immigrants from poorer countries outside Europe	0,16**	0,23**	0,12**	0,17**	0,13**
immigration bad or good for country's economy	0,07**	0,11**	-0,02	0,07**	-0,01
country's cultural life undermined or enriched by immigrants	0,09**	0,18**	0,11**	0,15**	0,06**
country is made a worse or a better place by immigrants	0,09**	0,15**	0,1**	0,14**	0,08**

Anmerkung: Gewichtung mit Designgewicht; listenweiser Fallausschluss; **= Korrelation ist auf dem 0,01 Niveau signifikant (einseitig), *= Korrelation ist auf dem 0,05 Niveau signifikant (einseitig); Korrelationen mit Spearmans Rho.
Lesebeispiel: Die Korrelation zwischen den Skalensummenwerten der ersten Skala, zur Messung der Einstellung von Befragten zu Migration weist in Westdeutschland einen Korrelationskoeffizienten von 0,1 auf, der auf dem 0,01 Niveau signifikant ist.

Zwischen den Skalensummenwerten der beiden Skalen und dem Alter der Befragten sind nur schwache Korrelationen feststellbar. (vgl. Tab. 10) Die stärksten Zusammenhänge liegen hierbei in Großbritannien und die schwächsten in Ostdeutschland und Österreich vor. Es zeigt sich, dass die Relation der Stärke der Zusammenhänge zwischen dem Alter und den Items der ersten Skala bis auf zwei Ausnahmen zwischen den Analyseeinheiten ähnlich sind. In Westdeutschland ist die Korrelation zwischen dem Item „allow immigrants of same race/ethnic group as majority" und dem Alter etwas schwächer, als es auf Grund der anderen Zusammenhänge zu vermuten wäre. Da dieser Unterschied sehr gering ausfällt, erscheint die Existenz einer Verzerrung unwahrscheinlich. In Ostdeutschland kann

hingegen nicht davon ausgegangen werden, dass überhaupt ein Zusammenhang zwischen dem Alter und diesem Item in der Grundgesamtheit vorliegt.

Die Korrelationsmuster der Items der zweiten Skala mit dem Alter der Befragten sind zwischen Großbritannien und Westdeutschland vergleichbar**.** In Ostdeutschland und Österreich existiert im Gegensatz zu den anderen zwei Analyseeinheiten kein Zusammenhang zwischen dem Item „immigration bad or good for country's economy" und dem Alter der befragten Personen. In Österreich ist zudem die Korrelation zwischen dem fünften Item und dem Alter schwächer beziehungsweise die Korrelation zwischen dem sechsten Item und dem Alter stärker als auf Grundlage der übrigen Korrelationsmuster zu erwarten wäre. Es wurde schon bei der Betrachtung der internen Konsistenz der Skalen und der Mittelwerte der einzelnen Items vermutet, dass das Item „country's cultural life undermined or enriched by immigrants" in Österreich eine unterschiedliche Bedeutung als in den anderen Analyseeinheiten besitzt. Die Existenz eines *non uniform Item bias* bei diesem Item und dadurch die Absenz einer *Measurement unit equivalence* zwischen Österreich und den restlichen Analyseeinheiten ist somit mit ziemlicher Sicherheit bestätigt.

6.3. Überprüfung der funktionalen Äquivalenz mittels Faktorenanalysen

Die explorative Faktorenanalyse hat den Zweck, die Eindimensionalität der Skalen, sowie die Existenz einer funktionalen Äquivalenz dieser für die vier Analyseeinheiten festzustellen. Die funktionale Äquivalenz der beiden Skalen kann dann angenommen werden, wenn die Faktorenstruktur zwischen den einzelnen Analyseeinheiten ähnlich ist. Für diese Überprüfung werden separate Faktorenanalysen für die einzelnen Untersuchungseinheiten, sowie eine gemeinsame Faktorenanalyse für alle Untersuchungseinheiten berechnet. Die erhaltenen Faktorenstrukturen werden zuerst paarweise zwischen den Analyseeinheiten und schließlich ebenfalls paarweise mit der Faktorenstruktur des Gesamtdatensatzes verglichen. Die Betrachtung der Relation der Faktorladungen zueinander in den einzelnen Ländern entscheidet über das Ausmaß der Ähnlichkeit der Faktorenstrukturen und somit über das Ausmaß der Äquivalenz. Es ist also weder der erklärte Varianzanteil eines Faktors noch die Höhe der Faktorladungen

für die Bestimmung der funktionalen Äquivalenz relevant, sondern lediglich deren Relation.

Der Bartlett-Test auf Sphärizität ergibt für alle Untersuchungseinheiten ein hochsignifikantes Ergebnis der Chi-Quadrat Werte. Es kann also mit einer Wahrscheinlichkeit von über 99% davon ausgegangen werden, dass zumindest einige der in der Stichprobe beobachteten Zusammenhänge zwischen den Items auch in der interessierenden Grundgesamtheit existieren. Die Maße der Stichprobeneignung nach Kaiser-Meyer-Olkin nehmen in den einzelnen Stichproben Werte zwischen 0,85 bis 0,87 an. Dies ist ein gutes Ergebnis und es kann demzufolge davon ausgegangen werden, dass die Items in einem hohen Ausmaß korreliert sind, beziehungsweise dass nur niedrige partielle Korrelationen zwischen den Items existieren. Die Maße der Stichprobeneignung der einzelnen Items variieren hierbei innerhalb eines Wertebereichs von 0,79 bis 0,91. (vgl. Tab. A.15-A.19, Anhang:150-151) Das zweite Item, das in allen Untersuchungseinheiten die niedrigste Stichprobeneignung aufweist, unterschreitet im westdeutschen Datensatz den erforderlichen Schwellenwert von 0,8 leicht. Da dieses Item das zentrale Item der ersten Skala darstellen dürfte, es aber nur minimal unter diesem Schwellenwert liegt, wird von einem Ausschluss dieses Items aus der Faktorenanalyse abgesehen. Es sind also alle Vorraussetzungen für die Durchführung einer explorativen Faktorenanalyse erfüllt und die Ergebnisse können sinnvoll interpretiert werden.

Da auf Grund der hohen Interitemkorrelationen der zwei Skalen anzunehmen ist, dass beide Einflussfaktoren realiter korreliert sind, wird eine Hauptachsenanalyse mit der Oblimin-Methode durchgeführt, die dieser vermuteten Korrelation der Faktoren gerecht wird.[62] Die Betrachtung der Tabelle 11 zeigt, dass vergleichbare Korrelationen zwischen den Faktoren in den Analyseeinheiten existieren.[63] Die

[62] Durch die Anwendung einer obliquen Rotation wird die Annahme der Orthogonalität der Faktoren bewusst aufgegeben. Die Aufgabe der Orthogonalität der Faktoren hat den Nachteil, dass die Faktorladungen nicht mehr den Korrelationskoeffizienten zwischen dem Faktor und den Variablen entsprechen. Deswegen besteht die Ausgabe des Ergebnisses der Faktorenanalyse nicht mehr nur aus der Matrix der direkten Faktorladungen eines Faktors auf seine Komponenten, sondern auch aus einer Matrix mit den Gesamtkorrelationen der Komponenten mit den Faktoren. Die erste Matrix wird in diesem Fall als Mustermatrix bezeichnet und die zweite Matrix als Strukturmatrix. (vgl. Brosius 1998:660)

[63] Die Kommunalitäten der Komponenten nach Faktorextraktion werden nicht extra dargestellt, da etwaige länderspezifische Unterschiede für die funktionale Äquivalenz der Skalen unbedeutend sind. Dies ist vergleichbar mit dem Umstand, dass Variationen der Eigenwerte der Faktoren zwischen den

Korrelationskoeffizienten weisen Werte zwischen 0,77 und 0,84 auf. Die Betrachtung der Faktordiagramme im gedrehten Faktorbereich zeigt, dass die Annahme der in dieser Stärke korrelierten Faktoren gerechtfertigt ist, da diese die Komponenten gut abbildet. (vgl. Abb. A.1-5, Anhang:153) In diesen Faktordiagrammen sind die Faktorladungswerte der Mustermatrix in einem zweidimensionalen Raum abgebildet. Die Betrachtung von sowohl den Werten der Mustermatrix als auch der Faktordiagramme ist trotz der in ihnen enthaltenen redundanten Information hilfreich, da gerade die grafische Darstellung der Faktorenstruktur die Prüfung der Ähnlichkeit und Unterschiede zwischen diesen für die Analyseeinheiten erleichtert.

Werden die Werte der Faktorladungen in den einzelnen Analyseeinheiten betrachtet, wird sichtbar, dass nach der obliquen Rotation die Items der beiden Skalen hohe Ladungen auf einen der beiden extrahierten Faktoren aufweisen. (vgl. Tab. 11) Die Ladungen auf dem jeweils anderen Faktor sind sehr nahe bei Null. Das zweite Item weist überall die höchsten Faktorladungen auf den ersten Faktor auf, während das sechste Item überall die höchsten Faktorladungen auf den zweiten Faktor aufweist. Wie bereits mehrfach dargestellt, sind dies die zentralen Items für diese beiden Dimensionen.

Bei der Betrachtung der Faktorladungen des Faktors, der die Einstellung gegenüber Migration abbildet, wird sichtbar dass die Faktorenstruktur in Ost- und Westdeutschland und in Österreich relativ ähnlich ist. In diesen Analyseeinheiten weist das erste Item die schwächsten Faktorladungen auf, das dritte Item die zweitstärksten und das zweite Item die stärksten Faktorladungen auf diesem Faktor. Jedoch existieren auch zwischen diesen drei Analyseeinheiten kleine Unterschiede. So weist das erste Item in Westdeutschland eine niedrigere Faktorladung in Relation zu den Faktorladungen der anderen beiden Items auf als in Ostdeutschland und Österreich. Zudem zeigt sich, dass die Höhe der Faktorenladung des zweiten Items in Österreich ähnlicher den anderen beiden Items ist, als dass in Ost- und Westdeutschland der Fall ist. Die Faktorenladungen der Komponenten dieses Faktors weichen hingegen in Großbritannien im Vergleich zu den anderen Analyseeinheiten ab. So sind in dieser Gruppe die Faktorladungen des ersten Items weitaus höher als beim dritten Item. Die Höhe der Faktorladungen dieser beiden

Analyseeinheiten für die funktionale Äquivalenz keine Rolle spielen. (vgl. Kap. 4.2.2) Eine Tabelle mit den Kommunalitäten befindet sich jedoch im Anhang. (vgl. Abb. A.20, Anhang:xy)

Komponenten in Großbritannien ist hierbei genau reziprok wie dies in Österreich der Fall ist.

Tabelle 11: Ergebnis der Faktorenanalyse nach Oblimin-Rotation

	D,GB,Ö		GB		Ost-D		West-D		Ö	
KMO	0,86		0,85		0,86		0,85		0,87	
Anzahl der Faktoren	2		2		2		2		2	
Korr. zw. Faktoren	0,809		-0,774		0,819		0,795		0,835	
Winkel zw. Faktorenachsen	36º		141º (= 39º)		35º		37º		33º	
Faktor	1	2	1	2	1	2	1	2	1	2
Mustermatrix										
allow immigrants of same race/ethnic group as majority	0,79	-0,04	-0,11	-0,90	0,75	0,04	0,70	0,04	0,78	-0,04
allow immigrants of different race/ethnic group as majority	1,07	-0,15	-0,12	-1,04	1,08	-0,17	1,09	-0,18	1,05	-0,14
allow immigrants from poorer countries outside Europe	0,84	-0,01	0,06	-0,78	0,82	-0,05	0,82	0,00	0,89	-0,06
immigration bad or good for country's economy	0,03	0,72	0,79	-0,01	0,01	0,76	0,09	0,65	-0,04	0,80
cultural life undermined/enriched by immigrants	-0,09	0,88	0,90	0,08	-0,07	0,85	-0,13	0,88	-0,08	0,90
country is made a worse or a better place by immigrants	-0,12	0,98	1,02	0,12	-0,12	0,97	-0,06	0,90	-0,11	0,94
Strukturmatrix										
Item 1	0,76	0,61	0,59	-0,82	0,78	0,65	0,73	0,59	0,75	0,61
Item 2	0,94	0,71	0,69	-0,95	0,94	0,71	0,95	0,69	0,94	0,74
Item 3	0,83	0,67	0,66	-0,83	0,79	0,63	0,82	0,65	0,85	0,69
Item 4	0,62	0,75	0,80	-0,62	0,63	0,77	0,60	0,72	0,62	0,76
Item 5	0,62	0,81	0,84	-0,62	0,62	0,79	0,57	0,78	0,67	0,84
Item 6	0,67	0,88	0,93	-0,67	0,67	0,87	0,66	0,86	0,68	0,85

Anmerkung: Gewichtung mit Designgewicht; Extraktionsmethode: Hauptachsenanalyse mit Oblimin-Rotation; listenweiser Fallausschluss; Bartlett-Test auf Sphärizität in allen Fällen hochsignifikant; es wurden nur Faktoren extrahiert, die einen Eigenwert $\geq 0,5$ aufweisen. Die in der Tabelle enthaltenen Werte entsprechen der mit der Oblimin-Methode (Delta=0,2) rotierten Faktorenlösung. Ein Delta-Wert, der größer als Null ist entspricht einem Abhängigkeitskriterium einer hohen Korrelation zwischen den Faktoren. (Langer 1999a:27-28) [64] Der Winkel zwischen den Faktorenachsen ergibt sich aus der Umrechnung der Korrelation zwischen den Faktoren mit der Arkus-Kosinus Funktion. Die Mustermatrix entspricht den direkten Faktorladungen des Faktors auf die Komponenten. Die Strukturmatrix enthält Gesamtkorrelation zwischen Komponenten und Faktor. Die Vorzeichen der Werte der Muster- und Strukturmatrizen dürfen nicht inhaltlich interpretiert werden, da diese lediglich die Richtung der Winkelverschiebung der Faktorenachsen wiedergeben. (vgl. Langer 1999b: 21-22)[65]
Lesebeispiel: Das Item „allow immigrants of same race/ethnic group as majority" lädt in Österreich nach der obliquen Rotation mit 0,78 auf den ersten Faktor und mit 0,04 auf den zweiten Faktor.

[64] http://www.soziologie.uni-halle.de/langer/lisrel/index.html
[65] http://www.soziologie.uni-halle.de/langer/lisrel/skripten/faktxeno.pdf

Werden die Faktorenstrukturen der einzelnen Analyseeinheiten mit der gemeinsamen Faktorenstruktur aller Analyseeinheiten verglichen, bestätigen sich die eben getätigten Feststellungen. Durch den Umstand, dass die Faktorenstruktur Großbritanniens relativ zu den anderen Analyseeinheiten so verschieden ist, stimmt diese allerdings mit keiner der Faktorenstruktur der Analyseeinheiten wirklich gut überein.

Beim zweiten Faktor, der die Einstellung gegenüber Migration abbildet, ist zumindest die Abfolge der Komponenten gemäß deren Höhe der Faktorladungen zwischen den Analyseeinheiten gleich. In allen Gruppen zeigen sich beim vierten Item die niedrigsten Faktorladungen, beim fünften Item die zweithöchsten und beim sechsten Item die höchsten Faktorladungen. Dass die Faktorenstruktur der Analyseeinheiten bei dem zweiten Faktor vergleichbarer als bei dem ersten Faktor sind bestätigt auch der paarweise Vergleich dieser mit der gemeinsamen Faktorenstruktur aller Analyseeinheiten. Die Faktorenstrukturen sind zwischen Großbritannien und Ostdeutschland sehr ähnlich, während die Faktorenstrukturen in Westdeutschland und Österreich hiervon verschieden sind. Bei letzteren zeigt sich, dass die Stärke der Faktorladungen des fünften und des sechsten Items sich ähnlicher sind, als dies in den anderen beiden Analyseeinheiten der Fall ist. Zudem weist das vierte Item in Westdeutschland im Vergleich zu den anderen Untersuchungseinheiten eine weitaus niedrigere Faktorenladung auf.

Der Vergleich der Ergebnisse der Faktorenanalyse hat gezeigt, dass die Skalen überall ein jeweils eindimensionales Konstrukt messen und diese beiden Dimensionen überall einen ähnlichen Zusammenhang aufweisen dürften. Allerdings zeigen die Ergebnisse auch, dass unterschiedliche Faktorenstrukturen in den Analyseeinheiten existieren. So unterscheidet sich die Faktorenstruktur des Faktors, der die Einstellung gegenüber Migration abbildet in Großbritannien deutlich von der der restlichen Untersuchungseinheiten. Die Faktorenstruktur des zweiten Faktors, der Einstellung gegenüber Migration abbildet, unterscheidet sich, wenn auch in einem geringeren Umfang zwischen Westdeutschland und Österreich und den restlichen Analyseeinheiten. Letztlich fällt die Entscheidung, ob und wie gut das Faktorenmodell mit den Daten übereinstimmt und somit die erreichte Ebene der Äquivalenz anhand der noch durchzuführenden konfirmatorischen Faktorenanalyse.

6.4. Überprüfung des Ausmaßes der funktionalen Äquivalenz mittels einer konfirmatorischen Faktorenanalyse

Es wird nun mittels der Durchführung einer konfirmatorischen Faktorenanalyse und eines multiplen Gruppenvergleichs, das Ausmaß der Äquivalenz der Skalen für den Vergleich in den betrachteten Ländern bestimmt. Es existieren verschiedene Verfahren zur Schätzung der Modellparameter, zahlreiche *Goodness-of-Fit*-Indizes und Teststatistiken zur Bestimmung der Güte des Modells.[66] Diese basieren auf verschiedenen Annahmen und stellen unterschiedliche Anforderungen an das Skalenniveau und Verteilung der Daten. So erfordert zum Beispiel die oftmals für die Parameterschätzung favorisierte Maximum-Likelihood-Methode eine Normalverteilung der manifesten Variablen eines Modells. (vgl. Reinecke 2005:109-110) Da die Daten nicht normalverteilt sind muss bei der Durchführung der konfirmatorischen Faktorenanalyse auf ein Parameterschätzverfahren zurückgegriffen werden, welches keine Normalverteilung voraussetzt. (vgl. Kap. 3.5.2) Zudem wird die Verbesserung oder Verschlechterung der Güte der Modelle beim multiplen Gruppenvergleich mittels der Teststatistik der Chi-Quadrat-Differenzentest festgestellt.[67] Demzufolge ist ein Parameterschätzverfahren auszuwählen, das keine Normalverteilung voraussetzt und für das Chi-Quadrat-basierte Inferenzstatistiken verfügbar sind. Deswegen werden die Parameterschätzungen bei der konfirmatorischen Faktorenanalyse mittels der ADF-Methode vorgenommen. Diese asymptotisch verteilungsfreie Schätzmethode ermöglicht auch effiziente Parameterschätzungen bei nicht normalverteilten Daten und ermöglicht zudem die Berechnung von Chi-Quadrat-basierten Inferenzstatistiken. (vgl. Backhaus/Erichson/Plinke/Weiber 2006 370-371) Die Anforderung die dieses Schätzverfahren an die Stichprobengröße stellt, wird erfüllt. Es ist außerdem erforderlich, dass ein listenweiser Ausschluss der Fälle, bei Vorliegen eines fehlenden Wertes bei einem der Items vorgenommen wird. Dies ist insofern rechtzufertigen, dass bei der Analyse der fehlenden Werte keine nennenswerten systematischen Verzerrungen entdeckt werden konnten. (vgl. Kap.5.3.4)

[66] Für eine detaillierte Darstellung der verschiedenen Verfahren zur Parameterschätzung und deren Vorraussetzungen an die Daten sei auf Backhaus/Erichson/Plinke/Weiber 2006:368-371 und Reinecke 2005:107-115 verwiesen.

[67] Der Chi-Quadrat-Differenzentest wird in der gängigen Literatur auch als Likelihood-Ratio-Test bezeichnet. (vgl. Reinecke 2005:153-154)

Gütemaße, Teststatistiken und Testkriterien

Es werden lediglich jene Gütemaße, Teststatistiken und Testkriterien beziehungsweise deren Schwellenwerte beschrieben, die auch in dieser Arbeit zur Überprüfung der Anpassungsgüte der Daten der einzelnen Untersuchungseinheiten an das theoretische zweifaktorielle Modell verwendet werden.[68] Da die Parameterschätzungen Schätzungen eines konkreten Parameterwertes auf Basis einer Stichprobenverteilung darstellen, werden statistische Testkriterien betrachtet, die Rückschlüsse auf die Zuverlässigkeit dieser Schätzungen erlauben. So werden die Standardfehler der Parameterschätzungen betrachtet, um die Streuung dieser festzustellen. Sind diese groß, ist dies ein Anzeichen für die Unzuverlässigkeit der betreffenden Parameterschätzungen. Ein weiterer Weg, um die Zuverlässigkeit der Parameterschätzungen zu überprüfen, stellt die Betrachtung der quadrierten multiplen Korrelationskoeffizienten dar. Diese Koeffizienten entsprechen bei der konfirmatorischen Faktorenanalyse den quadrierten Faktorenladungen und sind ein Maß der Reliabilität der Messung der Indikatoren. Die Reliabilität der jeweiligen Konstrukte entspricht dem Mittelwert der quadrierten Faktorenladungen der zugehörigen Indikatoren. Die Indikatorreliabilitäten sollten hierbei größer als 0,5 sein. Der Einflussfaktor sollte also mindestens 50 Prozent der Varianz eines jeden einzelnen Indikators erklären können. Zusätzlich hierzu werden sogenannte *Critical Ratio*-Werte betrachtet, die dem Quotienten aus einer unstandardisierten Werte der Parameterschätzungen und deren Standardfehler entsprechen. Liegt dieser Wert über 1,96 kann mit einer fünfprozentigen Irrtumswahrscheinlichkeit davon ausgegangen werden, dass sich die Parameterschätzungen signifikant von Null unterscheiden.

Es werden zur Beurteilung der Güte der Gesamtstruktur des Modells in dieser Arbeit in erster Linie *Goodness-of-Fit*-Indizes eingesetzt, die auf einen Wertebereich zwischen Null und Eins normiert sind. Dies erleichtert die Interpretation, da hierbei auch in der gängigen Literatur Konsens darüber vorherrscht, ab welchen Schwellenwerten von einem guten beziehungsweise akzeptablen Modellfit ausgegangen werden kann. Die in dieser Arbeit verwendeten Gütemaße sind:

[68] Für eine detaillierte Darstellung der zahlreiche *Goodness-of-Fit* Maße und Teststatistiken sei auf Backhaus/Erichson/Plinke/Weiber 2006:376-387, Reinecke 2005:115-128 und Baltes-Götz 2008 (www.uni-trier.de/fileadmin/urt/doku/amos/v16/amos16.pdf) verwiesen. Alle wiedergegebenen Charakteristiken dieser Maße entstammen diesen Werken.

- Die Chi-Quadrat-Statistik ist die am häufigsten verwendete Teststatistik zur Evaluation eines Modells. Der Chi-Quadrat-Wert wird umso größer je weniger das Modell mit den Daten übereinstimmt. Der Chi-Quadrat Wert wird in ein Verhältnis zur Anzahl der Freiheitsgrade gesetzt und erst dann interpretiert. So wird ein Verhältnis von drei zu eins zwischen dem Chi-Quadrat Wert und der Anzahl der Freiheitsgrade als ein Indiz dafür gewertet, dass der Modellfit gut ist. Die Chi-Quadrat-Statistik wird in der vorliegenden Arbeit allerdings in erster Linie als deskriptive Statistik und nicht als Teststatistik für die Evaluation der Modellgüte eingesetzt.[69]
- Der *Goodness-of-Fit-Index* (GFI), beschreibt, das Ausmaß der durch das Modell erklärten Varianz und Kovarianz, die durch das Modell erklärt wird. Liegt der GFI über 0,9 kann von einem akzeptablen Modellfit ausgegangen werden. Liegt er hingegen über 0,95 kann von einem guten Modellfit gesprochen werden.
- Der *Adjusted-Goodness-of-Fit-Index* (AGFI) stellt eine Korrektur des GFI um die Anzahl der Freiheitsgrade eines Modells dar. Die Schwellenwerte für die Bestimmung der Anpassungsgüte des Modells sind identisch mit jenen des GFI.
- Der *Comparative-Fit-Index* basiert auf einem Vergleich der Modellgüte eines Unabhängigkeitsmodells, bei dem angenommen wird das alle Indikatoren unkorreliert sind, mit dem bestehenden Modell. Es wird hierbei ein Verhältnis zwischen den Quotienten aus Chi-Quadrat-Werten und den Freiheitsgraden der beiden Modelle gebildet. So wird überprüft, ob das bestehende Modell eine ausschlaggebende Verbesserung zu dem Unabhängigkeitsmodell darstellt. Die Schwellenwerte als Beurteilungsgrundlage sind hierbei identisch mit den beiden obigen Maßen.
- Der *Root Mean Square Error of Approximation* (RMSEA) ist eigentlich als *Badness-of-Fit*-Index zu bezeichnen. Der RMSEA basiert auf der Abweichung der Kovarianzmatrix des Modells von der Kovarianzmatrix der empirischen Daten. Liegt der RMSEA unter 0,08 kann von einem akzeptablen Modellfit

[69] So überprüft die Chi-Quadrat-Teststatistik die exakte Übereinstimmung des Modells mit den empirischen Daten, was eine nicht realistische und dadurch zu strikte Annahme ist. Jedoch weitaus bedeutender für die Ablehnung als Teststatistik ist die hochgradige Abhängigkeit der die Chi-Quadrat-Teststatistik von der Größe der Stichprobe. Je höher die Fallzahl ist, desto höher ist auch der Chi-Quadrat Wert. Zudem erhöht sich mit einer steigenden Fallzahl die Wahrscheinlichkeit, dass fälschlicherweise eine Übereinstimmung des Modells mit den empirischen Daten abgelehnt wird. Dieser Umstand erscheint gerade im Hinblick auf die sehr großen Substichproben der Analyseeinheiten als nicht akzeptabel. Die anderen dargestellten *Goodness-of-Fit*-Indizes sind hingegen unabhängig vom Stichprobenumfang und setzen keine exakte Übereinstimmung des Modells mit den empirischen Daten voraus, sondern lediglich eine näherungsweise Übereinstimmung.

ausgegangen werden. Liegt er hingegen unter 0,05 kann von einem guten Modellfit gesprochen werden. Zudem existiert ein Test der Nullhypothese, dass der Wert tatsächlich kleiner als 0,05 ist. Sollte dieser Test insignifikant sein, kann mit gegebener Irrtumswahrscheinlichkeit davon ausgegangen werden, dass der Wert des RMSEA tatsächlich kleiner oder gleich 0,05 ist.

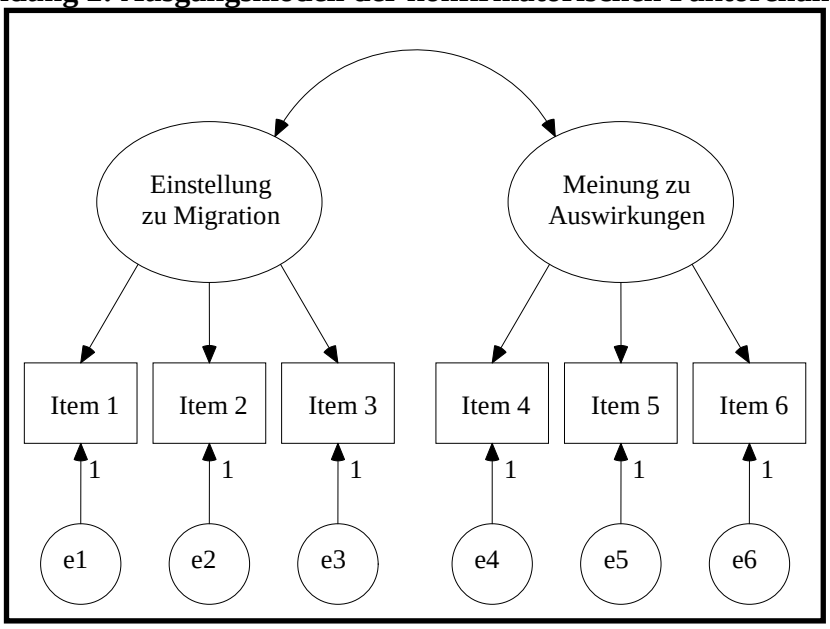
Sollten diese Gütemaße für die Beurteilung der Gesamtstruktur indizieren, dass das theoretische Faktorenmodell der zwei Skalen in einem ungenügenden Ausmaß mit den empirischen Daten übereinstimmt, werden Gütekriterien für die Beurteilung der Teilstrukturen des Modells betrachtet. So soll festgestellt werden, ob ein bestimmter Teil des Modells für die schlechte Gesamtgüte des Modells verantwortlich ist.

Ergebnisse der konfirmatorischen Faktorenanalysen

Für die Berechnung der separaten Lösungen der konfirmatorischen Faktorenanalysen wird die Varianz der latenten Variablen auf den Wert Eins fixiert, da dies den Vergleich der Regressionskoeffizienten der Items zwischen den Analyseeinheiten erleichtert. Bei der Durchführung der multiplen Gruppenvergleiche wird allerdings die Referenzindikatorstrategie angewandt. Da die größte Messäquivalenz zwischen den Analyseeinheiten in den bisherigen Analysen die Items zwei und sechs aufgewiesen haben, werden diese als Referenzindikatoren für die Analyse ausgewählt. Es werden die Referenzindikatoren jedoch auch variiert, um die Stabilität der erhaltenen Lösung abschätzen zu können. Diese Ergebnisse dieser Variationen werden jedoch nicht ausdrücklich dargestellt.

Die Abbildung 2 enthält das Ausgangsmodell der konfirmatorischen Faktorenanalyse, welches auf Grund der theoretischen Vorüberlegungen und der Ergebnisse der explorativen Faktorenanalyse in allen Analyseeinheiten Gültigkeit aufweisen sollte. Mit diesem Modell wird ein gemeinsames konfirmatorisches Faktorenmodell für alle Analyseeinheiten, vier separate konfirmatorische Faktorenmodelle für die Analyseeinheiten und abschließend ein multipler Gruppenvergleich berechnet. Zuerst werden die Parameterschätzungen, Gütemaße und Testkriterien für die separaten konfirmatorischen Faktorenanalysen in den einzelnen Analyseeinheiten dargestellt.

Abbildung 2: Ausgangsmodell der konfirmatorischen Faktorenanalyse



In der Tabelle 12 sind die Gütemaße für die Gesamtstruktur des Modells in den einzelnen Analyseeinheiten enthalten. Bei der Betrachtung der Gütemaße wird sichtbar, dass die Anpassungsgüte des Modells mit den Daten zwischen den einzelnen Gruppen unterschiedlich ist. Der Quotient aus dem Chi-Quadrat-Wert und den Freiheitsgraden des Modells übersteigt in Großbritannien, Westdeutschland und beim Modell mit allen Ländern den Wert 3. Es wurde bereits darauf hingewiesen, dass die Chi-Quadrat-Statistik hochgradig abhängig vom Stichprobenumfang ist und demzufolge als deskriptive Statistik eingesetzt wird. Trotzdem können Unterschiede hierbei als Indiz dafür gewertet werden, dass Unterschiede bezüglich der Anpassungsgüte der Gesamtstruktur existieren und das Modell in Österreich den besten Modellfit aufweist. Bei der Betrachtung des Quotienten des Chi-Quadrat-Wertes und der Freiheitsgrade sollten allerdings die unterschiedlichen Stichprobengrößen der Gruppen berücksichtigt werden. Besonders für Ostdeutschland allerdings auch für Westdeutschland wäre eine dementsprechende Erhöhung des Quotienten bei einem zu den anderen zwei Analyseeinheiten vergleichbaren Stichprobenumfangs zu erwarten. Die Betrachtung der *Goodness-of-Fit*-Indizes zeigt hingegen, dass durchaus von einem guten Modellfit für alle Analyseeinheiten, sowie für das Modell mit allen Analyseeinheiten ausgegangen werden kann. Dies ist ein erstes, wenn auch letztlich ein schwaches Anzeichen dafür, dass die Struktur des Modells in allen Analyseeinheiten gleich sein dürfte, also eine funktionale Äquivalenz vorhanden ist. Der GFI, AGFI und CFI ist in allen Gruppen größer als 0,95 und der RMSEA in fast allen Gruppen kleiner als 0,05. Lediglich in

Westdeutschland wird dieser Schwellenwert, der einen guten Modellfit indiziert, knapp überschritten. Das Modell weist die größte Übereinstimmung mit den Daten der österreichischen Stichprobe auf und die geringste Übereinstimmung mit den Daten der westdeutschen Stichprobe. Von einer Eindimensionalität der beiden Skalen kann also in allen Analyseeinheiten ausgegangen werden.

Tabelle 12: Gütemaße der jeweiligen separat berechneten Modelle

	GB,D,Ö	GB	Ost-D	West-D	Ö
N	6934	2253	952	1684	2045
χ^2	99,128	47,672	19,722	43,950	14,666
χ^2/df	12,391	5,959	2,465	5,494	1,833
GFI	,993	,990	,992	,986	,997
AGFI	,982	,974	,978	,963	,991
CFI	,982	,980	,982	,966	,996
RMSEA	,041	,047	,039	,052	,020

Anmerkung: listenweiser Ausschluss der Fälle mit einem fehlenden Wert. Das Modell enthält 8 Freiheitsgrade, Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der RMSEA weist in Österreich einen Wert von 0,2 auf, was ein Indiz für einen guten Modellfit darstellt.

Da eine leicht unterschiedliche Anpassungsgüte der Gesamtstruktur des Modells zwischen den Analyseeinheiten konstatiert wurde, wird in einem weiteren Schritt die Anpassungsgüte der Teilstrukturen des Modells betrachtet. Damit sollen etwaige Unterschiede zwischen der Anpassungsgüte der Teilstrukturen des Modells zwischen den Gruppen identifiziert werden, die einen Einfluss auf die Differenzen der Anpassungsgüte des Gesamtmodells haben. Die Indikatorreliabilitäten sind in allen Analyseeinheiten immer höher als 0,5. Bei der Betrachtung der Indikatorreliabilitäten bestätigen sich erneut einige Feststellungen der bisherigen Analyse. So weisen das erste und vierte Item in allen Analyseeinheiten die geringste Reliabilität auf, während das zweite und sechste Item die höchste Reliabilität aller Items aufweisen. Zudem zeigt sich, dass in Österreich ausgehend von den anderen Reliabilitätswerten der Items, das Item „country's cultural life undermined or enriched by immigrants" einen bedeutend höheren Reliabilitätswert aufweist als dies in den anderen Analyseeinheiten der Fall ist. Es wird also mehr Varianz dieses Items durch den zweiten Faktor erklärt. Dies war zu erwarten, da die Existenz eines *non uniform Item bias* bei diesem Item und dadurch die Absenz einer *Measurement unit equivalence* zwischen Österreich und den restlichen Untersuchungseinheiten bereits festgestellt wurde.

Nun werden noch die einzelnen Parameterschätzungen zwischen den Analyseeinheiten verglichen, um zu überprüfen, ob offensichtliche Unterschiede zwischen den Analyseeinheiten evident sind.[70] (vgl. Abb. A.6-15; Anhang:154-158) Die Korrelationskoeffizienten zwischen den beiden Faktoren reichen von 0,71 in Großbritannien bis 0,79 in Österreich. Die Einflussfaktoren der beiden Skalen weisen also überall einen sehr hohen Zusammenhang auf. Dies bestätigt die Ergebnisse der explorativen Faktorenanalyse, bei der allerdings die Korrelation der beiden Faktoren leicht überschätzt wurde. (vgl. Kap. 6.3) Die dort festgestellte Reihenfolge der Analyseeinheiten bezüglich der Stärke des Zusammenhangs ist allerdings korrekt gewesen. Wenig überraschend zeigt sich auch, dass in allen Analyseeinheiten das zweite und das sechste Item, die höchsten Faktorladungen auf die jeweiligen Faktoren aufweisen, während die niedrigsten Faktorladungen überall für das erste und vierte Item festzustellen sind. Dies entspricht der Erkenntnis der bisher durchgeführten Analysen. Zudem wurden bereits die Indikatorreliabilitäten betrachtet, die ja bei der konfirmatorischen Faktorenanalyse den multiplen Faktorladungen entsprechen. Werden allerdings die Regressionskoeffizienten verglichen, offenbaren sich Unterschiede zwischen den Untersuchungseinheiten. So entspricht zwar die Reihenfolge der Items des ersten Faktors bezüglich der Höhe der Regressionskoeffizienten in allen Gruppen der Abfolge bezüglich der Höhe der Faktorladungen, allerdings ist diese beim zweiten Faktor zwischen den Analyseeinheiten unterschiedlich. Während die Rangfolge der betreffenden Items in Großbritannien und Westdeutschland übereinstimmt, ist diese in Ostdeutschland, Österreich, sowie bei der Lösung über alle Länder hinweg unterschiedlich hierzu. So weist in der Lösung für alle Gruppen und in Österreich das vierte Item die niedrigsten Regressionskoeffizienten auf und das fünfte Item die höchsten. In Ostdeutschland hingegen, wo die Regressionskoeffizienten der drei Items zueinander ähnlichere Werte aufweisen als in den restlichen Gruppen, weist das sechste Item den niedrigsten Regressionskoeffizienten auf und das fünfte Item den höchsten. Ob und inwieweit dies und die entdeckten Unterschiede der Teilstrukturen des Modells zwischen den Untersuchungseinheiten das Ausmaß der Äquivalenz zwischen den

[70] Es ist in allen Analyseeinheiten eine hohe Zuverlässigkeit der Parameterschätzungen gegeben.Die Standardfehler der einzelnen Parameterschätzungen sind niedrig und zwischen den einzelnen Analyseeinheiten auf einem vergleichbaren Niveau. Zudem sind alle *Critical Ratio*-Werte der einzelnen Parameterschätzungen in allen Untersuchungseinheiten signifikant. Es kann also mit einer fünfprozentigen Irrtumswahrscheinlichkeit davon ausgegangen werden, dass die Parameterschätzungen sich signifikant von Null unterscheiden.

beiden Skalen beeinträchtigen, soll nun durch die Durchführung eines multiplen Gruppenvergleichs festgestellt werden.

Durchführung des multiplen Gruppenvergleichs

In der folgenden Tabelle sind die Gütemaße und Teststatistik für den multiplen Gruppenvergleich zwischen allen Analyseeinheiten dargestellt. Mit dem simultanen Gruppenvergleich wird festgestellt, welches Ausmaß der Äquivalenz der Skalen in den betrachteten Analyseeinheiten vorliegt. Hierfür wird geprüft, inwieweit die einzelnen Parameter zwischen den Gruppen invariant sind. Sollte sich die Güte des Models zwischen dem restriktiveren und dem weniger restriktiven Modell verschlechtern, wird davon ausgegangen, dass die betreffenden Parameter des Modells zwischen den Gruppen nicht gleich sind.[71] Der Quotient aus dem Chi-Quadrat-Wert und den Freiheitsgraden des Modells beträgt bei dem Basismodell, bei dem keine Restriktionen vorherrschen zirka 3,9. (vgl. Tab. 13) Dieser Umstand war angesichts der fast 7000 Fälle, die dieses Modell beinhaltet, zu erwarten. Die Betrachtung der *Goodness-of-Fit*-Indizes zeigt aber, dass das Ausgangsmodell einen guten Modellfit aufweist. Die Werte des GFI, AGFI und CFI sind jeweils größer als 0,95 und der RMSEA ist kleiner als 0,05. Es kann also von einer Invarianz der Modellstruktur zwischen den einzelnen Analyseeinheiten und somit von der Existenz einer funktionalen Äquivalenz der zwei Skalen in allen Analyseeinheiten ausgegangen werden. Deswegen wird dieses Modell mit dem Modell verglichen, bei dem die Faktorladungen zwischen den Gruppen restringiert wurden. Während der GFI und der CFI beim restriktiveren Modell im Vergleich zum Basismodell etwas niedrigere Werte offenbaren, ist der AGFI stabil und der RMSEA weist einen etwas niedrigeren Wert auf. Allerdings ergibt der Chi-Quadrat-Differenzentest eine Chi-Quadrat-Differenz von 44,8; die bei einer Differenz von 12 Freiheitsgraden hochsignifikant ist. Das restriktivere Modell weist im Vergleich zum weniger restriktiven Modell eine signifikante Verschlechterung des Modellfits auf. Es kann also nicht von einer Invarianz der Faktorladungen und somit nicht von der Existenz von identischen Skaleneinheiten zwischen allen Analyseeinheiten ausgegangen werden. Dies bedeutet, dass bei den beiden Skalen keine *Measurement unit equivalence* zwischen Großbritannien, Österreich, Ost- und Westdeutschland vorliegt.

[71] Wenn die Differenz der Chi-Quadrat-Werte zwischen dem weniger restriktiven und dem restriktiveren Modell bei gegebener Differenz der Freiheitsgrade signifikant ist, heisst das, dass das restriktivere Modell im Vergleich zum weniger restriktiven Modell eine Verschlechterung des Modellfits aufweist.

Tabelle 13: Gütemaße für den multiplen Gruppenvergleich zwischen allen Analyseeinheiten

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,938	,991	,978	,982	,021	-	-	-
Faktorladungen invariant	3,882	,988	,978	,976	,020	44,803	12	,000

Anmerkung: Das erste Modell enthält 32 Freiheitsgrade und das mit den invarianten Faktorladungen 44. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 44,8; die bei einer Differenz von 12 Freiheitsgraden hochsignifikant ist.

Es werden nun die Modelle der einzelnen Gruppen paarweise miteinander verglichen, um festzustellen, ob zwischen zwei oder sogar drei Analyseeinheiten ein höheres Ausmaß an Parameterinvarianz und somit eine höhere Ebene der Äquivalenz vorhanden ist als über alle Einheiten. Die paarweisen Vergleiche zeigen, dass die Grundstruktur des Modells zwischen Ostdeutschland und Österreichs am ähnlichsten ist und zwischen Großbritannien und Westdeutschland am unterschiedlichsten. (vgl. Tab. A.21-26, Anhang:159-160) Jedoch weisen die *Goodness-of-Fit*-Indizes darauf hin, dass das unrestringierte Basismodell in allen Analyseeinheiten einen guten Modellfit aufweist. Der GFI, AGFI und CFI ist jeweils größer als 0,95 und der RMSEA ist in allen Fällen kleiner als 0,05. Allerdings ergeben alle bis auf einen Chi-Quadrat-Differenzentest signifikante Chi-Quadrat-Differenzen. In diesen Fällen ist von einer signifikanten Verschlechterung des Modellfits der unrestringierten Basismodelle im Vergleich zu den restriktiveren Modellen auszugehen. Besonders interessant ist hierbei der Umstand, dass die Faktorladungen zwischen Großbritannien und Ostdeutschland ähnlicher sind, als zwischen Ost- und Westdeutschland. (vgl. Tab. A.25-26) Der Gruppenvergleich zwischen Großbritannien und Westdeutschland zeigt, dass zwar für das Modell mit invarianten Faktorladungen der GFI einen etwas niedrigeren Wert annimmt, als beim Basismodell, allerdings verbessern sich die restlichen Gütemaße. (vgl. Tab. 14) Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 4,6; die bei einer Differenz von 4 Freiheitsgraden insignifikant ist. Dementsprechend ist keine Verschlechterung des Modellfits zu beobachten, wenn die Faktorladungen zwischen den beiden Gruppen fixiert werden. Es herrscht also für Großbritannien und Westdeutschland eine Messinvarianz im Sinne der Existenz identischer Skaleneinheiten und somit eine *Measurement unit equivalence* der beiden Skalen vor. Allerdings ergab der Vergleich des Modells mit invarianten Faktorladungen im Vergleich zu dem nächstrestriktiveren Modell, bei dem die Varianzen der beiden Konstrukte und deren

Kovarianz zwischen den zwei Gruppen restringiert wurden, eine signifikante Chi-Quadrat-Differenz. (vgl. Tab. A.26) So wurden die betreffenden drei Parameterschätzungen zwischen den Gruppen nur partiell restringiert. Es hat sich letztlich gezeigt, dass zusätzlich zu der Invarianz der Faktorladungen ebenfalls die Varianz des Konstruktes der Einstellung zur Migration zwischen den beiden Gruppen gleich ist. (vgl. Tab. 14)

Tabelle 14: Gütemaße für den multiplen Gruppenvergleich zwischen Großbritannien und Westdeutschland

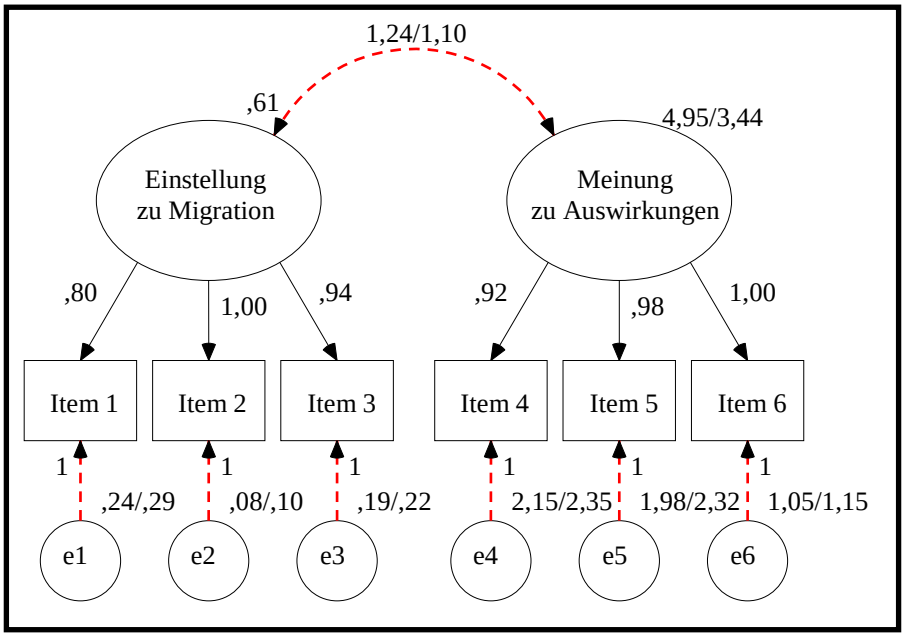
	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	5,726	,989	,970	,975	,035	-	-	-
Faktorladungen invariant	4,811	,988	,975	,975	,031	4,592	4	,332
zus. 1. latente Var. invariant	4,669	,988	,975	,975	,031	1,835	1	,176

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade, das Modell mit den invarianten Faktorladungen 20 und das Modell bei dem zusätzlich die Varianz der latenten Variablen der ersten Skala restringiert wurde 21. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant. Der Chi-Quadrat-Differenzentest bezieht sich jeweils auf das vorhergehende Modell.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt für den Modellvergleich der Modelle, bei dem die Faktorladung zwischen den beiden Gruppen fixiert wurden und dem wo zusätzlich der Einflussfaktor der ersten Skala zwischen den Gruppen fixiert wurde eine Chi-Quadrat-Differenz von 1,8; die bei einer Differenz von einem Freiheitsgrad nicht signifikant ist.

Die folgende Abbildung verdeutlicht das Ausmaß der Messinvarianz zwischen den beiden Analyseeinheiten. Zudem wird wiedergegeben, auf welche Werte die die Regressionskoeffizienten und die Varianz des ersten Konstruktes fixiert werden konnten und warum die Varianz des zweiten Konstruktes und die Kovarianz der Konstrukte nicht invariant sind. Es liegt eine *Measurement unit equivalence* bei beiden Skalen zwischen Großbritannien und Westdeutschland vor. Dies bedeutet, dass identische Skaleneinheiten zwischen diesen beiden Analyseeinheiten existieren.[72]

[72] Was dies für die Art der möglichen Vergleiche zwischen diesen Analyseeinheiten bedeutet, wird im Kapitel 7.1 dargestellt.

Abbildung 3: multipler Gruppenvergleich zwischen Großbritannien und Westdeutschland



Anmerkung: unstandardisierte Lösung. Die rot markierten gestrichelten Pfade sind freigesetzt und die schwarzen Pfade wurden invariant gesetzt. Der linke Parameterwert bei den freigesetzten Pfaden entspricht dem Wert für Großbritannien, der rechte dem Wert für Westdeutschland.
Lesebeispiel: Das Item 2 weist in beiden Gruppen einen Regressionskoeffizienten von 0,8 auf den Faktor Einstellung zu Migration auf.

Da außer für den Vergleich zwischen Großbritannien und Westdeutschland die Faktorladungen variieren, werden in weiterer Folge lediglich Teile des Modells restringiert. Es wird also zugelassen, dass zwischen den Gruppen jeweils eine Faktorladung variiert. Dabei werden abwechselnd die Faktorladungen von verschiedenen Items freigesetzt. Somit soll überprüft werden, ob nur die Faktorladung eines Items zwischen den Gruppen nicht vergleichbar ist, oder ob mehrere Items zu den Konstrukten unterschiedliche Zusammenhänge zwischen den Gruppen aufweisen. Sollte sich zeigen, dass sich der Modellfit trotz der Freisetzung einer Faktorladung weiterhin zwischen den restriktiveren und dem Basismodell weiter verschlechtert, wird zusätzlich die Variation einer zweiten Faktorladung zwischen den Gruppen zugelassen. Es werden hierbei wieder alle möglichen Freisetzungskombinationen überprüft. Diese Vorgangsweise wird solange fortgesetzt bis letztlich keine signifikante Verschlechterung des Modellfits anzunehmen ist. Für den Fall, dass über alle Analyseeinheiten betrachtet die Mehrzahl der Faktorladungen variiert, werden lediglich die Faktorladungen Ostdeutschlands oder Österreichs mit jenen von Großbritannien und Westdeutschland verglichen. Denn diese zwei Analyseeinheiten weisen ja identische Skaleneinheiten auf. Auch wenn sich dies auf den ersten Blick wie ein „*trial and error*-

Prinzip“ anhört, sind letztlich nur so die betreffenden Items zu identifizieren, die einen unterschiedlichen Zusammenhang zu den jeweiligen Konstrukten in den Analyseeinheiten aufweisen. Es existieren natürlich auch Erkenntnisse aus den bisherigen Analyseschritten, die letztlich in Annahmen münden bei welchen Items Verzerrungen vermutet werden. Zudem sollte erwähnt sein, dass zusätzlich die Referenzindikatoren variiert werden, da deren Fixierung auf den Wert 1 unterstellt, dass diese zwischen den Gruppen invariant sind. Denn sollten diese Annahmen nicht zutreffend sein, hätte dies nämlich eine Verzerrungen der Beurteilung der Veränderung der Modellgüte zwischen den geschachtelten Modellen zur Folge und es würden falsche Schlüsse über die Gleichheit der Faktorladungen zwischen den Gruppen gezogen. (vgl. Temme/Hildebrandt 2008:20)[73] Es ist also erkennbar, dass eine aufwändige Vorgehensweise notwendig ist, um zu validen Ergebnissen zu gelangen. Es werden die einzelnen Schritte allerdings nicht gesondert dargestellt, sondern lediglich die Endergebnisse präsentiert.

So haben die einzelnen partiellen Fixierungen der Faktorladungen zwischen den Analyseeinheiten letztlich gezeigt, dass über alle Analyseeinheiten hinweg betrachtet keine nennenswerte Invarianz der Faktorladungen vorherrscht. So wurde auf Grund der Erkenntnisse der bisherigen Analyse eine Variation der Faktorladungen des ersten und des fünften Items zugelassen, da für diese ein niedrigeres Ausmaß an Äquivalenz als für die restlichen Items vermutet wurde. Es zeigte sich jedoch auch hierbei, dass dieses restriktivere Modell zu einem signifikant schlechteren Modellfit führt, wenn auch in einem geringeren Ausmaß als bei den bisher dargestellten multiplen Gruppenvergleichen. (vgl. Tab. A.27, Anhang:160) Es wurde ebenfalls geprüft, ob zumindest die Faktorladungen der Items der jeweiligen Skalen eine gleiche Variation zwischen den Analyseeinheiten aufweisen. Dieser Umstand muss allerdings ebenfalls negiert werden. (vgl. Tab. A.28-29, Anhang:161) Die Ähnlichkeit der Faktorladungen ist hierbei bei den Items der ersten Skala eher gegeben als bei den Items der zweiten Skala. Dass heißt aber trotzdem, dass letztlich keine identischen Skaleneinheiten für diese Skalen in den einzelnen Analyseeinheiten existent sind. Der Grund dafür ist, dass die Struktur des Modells zwar zwischen Ostdeutschland und Österreich insgesamt die größte Ähnlichkeit aufweist, allerdings die Skaleneinheiten zwischen diesen beiden Untersuchungseinheiten die größten Differenzen offenbaren. Dies wird auch

[73] http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2008-042.pdf

verdeutlicht durch die Ergebnisse der Vergleiche zwischen Großbritannien und Westdeutschland mit jeweils einer der beiden Gruppen.

Der Vergleich zwischen Großbritannien und Ost- und Westdeutschland offenbart, dass alle Faktorladungen bis auf jene des ersten Items zwischen den Gruppen invariant sind. (vgl. Tab. 15) Der Chi-Quadrat-Differenzentest ergibt nämlich eine Chi-Quadrat-Differenz von 11,6, die bei einer Differenz von 6 Freiheitsgraden insignifikant ist. Die Betrachtung der Veränderungen der *Goodness-of-Fit*-Indizes untermauert diese Tatsache. Während der GFI und der CFI jeweils geringfügig niedrigere Werte für das restriktivere Modell ausgeben, weisen der etwas höhere AGFI und der niedrigere RMSEA auf einen besseren Modellfit des restriktiveren Modells in Relation zum Ausgangsmodell hin. Das heißt, dass identischen Skaleneinheiten für diese fünf Items in den betreffenden Analyseeinheiten existent sind.

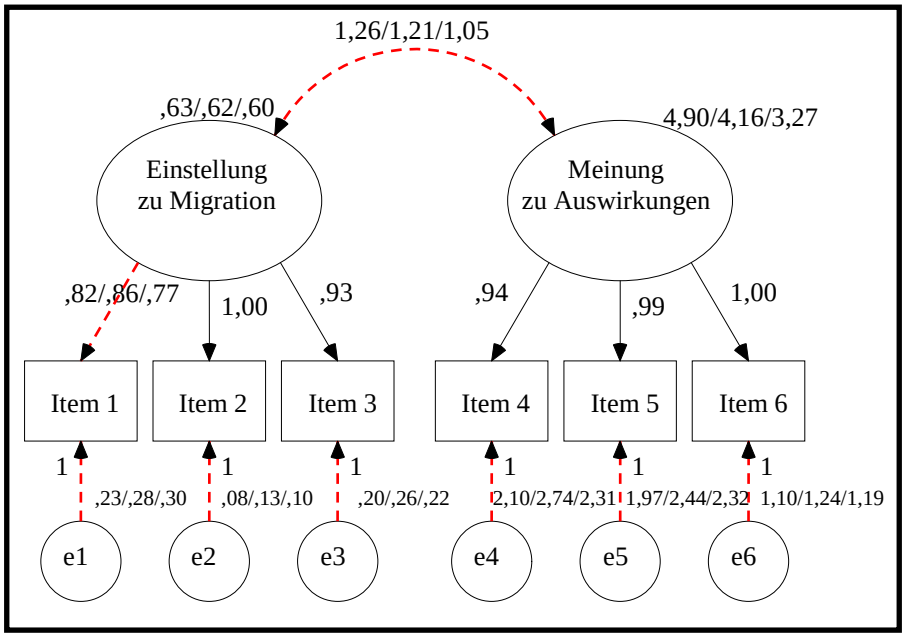
Tabelle 15: Gütemaße multipler Gruppenvergleich Großbritannien, Ostdeutschland und Westdeutschland (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	4,639	,989	,972	,977	,027	-	-	-
Faktorladungen part. invariant	4,097	,988	,975	,975	,025	11,573	6	,072

Anmerkung: Es wurden zwischen den drei Gruppen alle Regressionskoeffizienten bis auf den des ersten Items fixiert. Das erste Modell enthält 24 Freiheitsgrade und das mit den invarianten Faktorladungen 30. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 11,6; die bei einer Differenz von 6 Freiheitsgraden nicht signifikant ist.

Die Betrachtung der Abbildung 4 veranschaulicht, dass zwischen Großbritannien und Ost- und Westdeutschland bei fünf der sechs Items eine *Measurement unit equivalence* vorliegt. Beim Item „allow immigrants of same race/ethnic group as majority“ besteht jedoch ein *Item bias* in Ostdeutschland, dass heißt es bestehen dort unterschiedliche Skaleneinheiten im Vergleich zu den anderen zwei Untersuchungseinheiten. Dies bedeutet, dass bezüglich der Meinungen zu den Auswirkungen der Migration zwischen den betreffenden Analyseeinheiten eine Messinvarianz existiert. Bei den Einstellungen zu Migration trifft dies nur auf Teilaspekte dieses Konstruktes zu. So liegen lediglich bei den Items „allow immigrants of different race/ethnic group as majority“ und „allow immigrants from poorer countries outside Europe“ identische Skaleneinheiten vor.

Abbildung 4: multipler Gruppenvergleich Großbritannien, Ostdeutschland und Westdeutschland



Anmerkung: unstandardisierte Lösung, Die rot markierten gestrichelten Pfade sind freigesetzt und die schwarzen Pfade wurden invariant gesetzt. Der erste Parameterwert bei den freigesetzten Pfaden entspricht dem Wert für Großbritannien, der zweite Wert dem Wert für Ostdeutschland und der dritte Wert dem für Westdeutschland.
Lesebeispiel: Das Item 1 weist in Großbritannien einen Regressionskoeffizienten von 0,82, in Ostdeutschland von 0,86 und in Westdeutschland von 0,77 auf den Faktor Einstellung zu Migration auf.

Der Vergleich zwischen Großbritannien, Westdeutschland und Österreich zeigt hingegen, dass die alle Faktorladungen bis auf jene des fünften Items zwischen den Gruppen invariant sind. (vgl. Tab. 16) So ergibt der Chi-Quadrat-Differenzentest nämlich eine Chi-Quadrat-Differenz von 7,7, die bei einer Differenz von 6 Freiheitsgraden insignifikant ist. Dass heißt, dass letztlich identischen Skaleneinheiten für diese Items in den betreffenden Analyseeinheiten existent sind. Der Vergleich der Veränderung der restlichen Gütemaße zwischen den beiden geschachtelten Modellen lässt zudem darauf schließen, dass zwischen Großbritannien, Westdeutschland und Österreich ein höheres Ausmaß an partieller faktorieller Invarianz vorherrscht als dies für Großbritannien und Ost- und Westdeutschland der Fall ist. Während der GFI und der CFI jeweils die gleichen Werte für das restriktivere Modell ausgeben, weisen der höhere AGFI und der niedrigere RMSEA auf einen besseren Modellfit des restriktiveren Modells in Relation zum Ausgangsmodell hin. Zudem ist die beobachtete Chi-Quadrat-Differenz bei der gleichen Anzahl an Freiheitsgraden für diesen Modellvergleich niedriger als beim zuvor durchgeführten Gruppenvergleich.

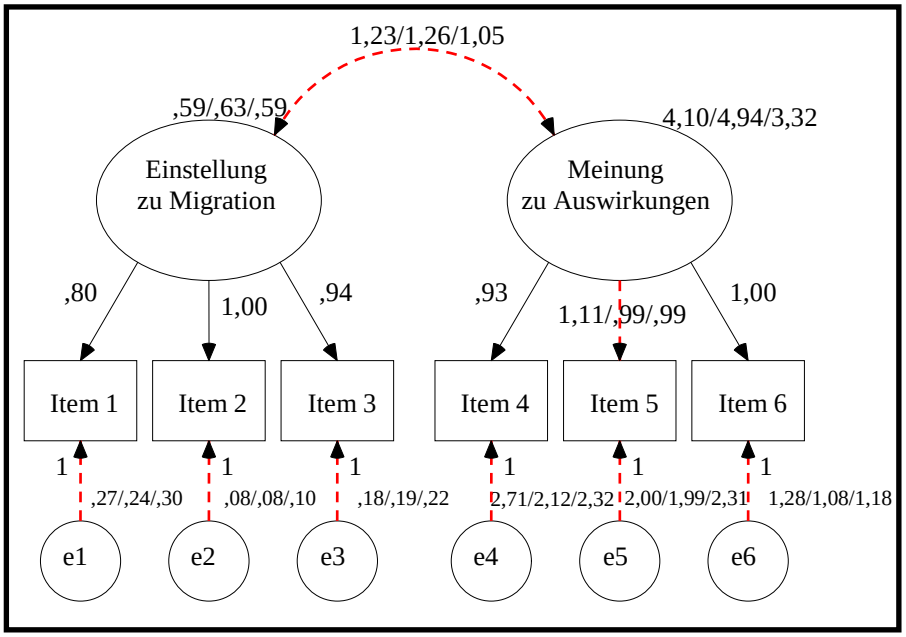
Tabelle 16: Gütemaße multipler Gruppenvergleich Großbritannien, Westdeutschland und Österreich (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	4,429	,991	,978	,982	,024	-	-	-
Faktorladungen part. invariant	3,799	,991	,981	,982	,022	7,683	6	,262

Anmerkung: Es wurden zwischen den drei Gruppen alle Regressionskoeffizienten bis auf den des fünften Items fixiert. Das erste Modell enthält 24 Freiheitsgrade und das mit den invarianten Faktorladungen 30. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 7,7; die bei einer Differenz von 6 Freiheitsgraden nicht signifikant ist.

Die folgende Abbildung verdeutlicht das Ausmaß der Invarianz der Parameter zwischen den Untersuchungseinheiten. Es wird sichtbar, warum der Regressionskoeffizient des Items „country's cultural life undermined or enriched by immigrants" freigesetzt werden muss, damit sich der Modellfit dieses Modells nicht signifikant im Vergleich zum Ausgangsmodell verschlechtert. Es existiert also bei fünf Items eine *Measurement unit equivalence* zwischen Großbritannien, Österreich und Westdeutschland. Die Erreichung einer vollständigen Messinvarianz ist auf Grund einer Verzerrung des fünften Items in Österreich durch einen *Item bias* nicht gegeben.

Abbildung 5: multipler Gruppenvergleich Österreich, Großbritannien und Westdeutschland



Anmerkung: unstandardisierte Lösung, Die rot markierten gestrichelten Pfade sind freigesetzt und die schwarzen Pfade wurden invariant gesetzt. Der erste Parameterwert bei den freigesetzten Pfaden entspricht dem Wert für Österreich, der zweite Wert dem Wert für Großbritannien und der dritte Wert dem für Westdeutschland.
Lesebeispiel: Das Item 1 weist in allen drei Gruppen einen Regressionskoeffizienten von 0,8 auf den Faktor Einstellung zu Migration auf.

Da die Mehrheit durchgeführten Vergleiche im Rahmen von Sekundäranalysen mit diesen Skalen vermutlich ohne eine Aufteilung Deutschlands in Ost- und Westdeutschland durchgeführt werden, wird in einem letzten Schritt überprüft, inwieweit eine partielle Invarianz der Faktorladungen auf Ebene der drei Nationalstaaten gegeben ist. Diese Vermutung ist damit begründet, dass eine Trennung Deutschlands in Ost- und Westdeutschland im Datensatz des ESS nicht vorgesehen ist und erst durch eine Integration einer zusätzlichen Variable des deutschen nationalen Moduls des ESS in den Hauptdatensatz ermöglicht wird. Die Betrachtung der partiellen Faktorinvarianz auf Länderebene soll keineswegs bedeuten, dass die Erkenntnisse der bisherigen Analyse bedeutungslos wären. Es wird vielmehr untersucht, ob bei Analysen auf einer höheren Ebene Unterschiede, die bei einer niederen Ebene der Analyse gefunden wurden verschwinden.

Die Ergebnisse des multiplen Gruppenvergleichs zeigen, dass dies tatsächlich der Fall ist. (vgl. Tab. 17) Während das Ausgangsmodell erwartungsgemäß eine gute Anpassungsgüte für die drei Länder aufweist, weist die partielle Invarianz der Faktorladungen die gleiche Struktur auf wie beim Vergleich zwischen Großbritannien, Westdeutschland und Österreich. So ergibt der Chi-Quadrat-Differenzentest eine Chi-Quadrat-Differenz von 4,8; die bei einer Differenz von 6 Freiheitsgraden insignifikant ist. Die Betrachtung der Veränderungen der *Goodness-of-Fit*-Indizes untermauert diese Tatsache. Während der GFI einen etwas niedrigeren Wert für das restriktivere Modell aufweist als für das Ausgangsmodell, sind die Werte des AGFI und des CFI jeweils höher und der Wert des RMSEA niedriger. Die in Relation zu den anderen durchgeführten multiplen Gruppenvergleichen äußerst niedrige Chi-Quadrat-Differenz würde bei Betrachtung auf Ebene der Länder auf ein höheres Ausmaß an faktorieller Invarianz schließen lassen, als dies tatsächlich zutreffend ist. Eine Analyse auf Ebene der Nationalstaaten würde also die Existenz einer *Measurement equivalence* in allen Ländern für alle Items außer dem fünften Item ergeben. Die Analyse bei der einer Aufteilung Deutschlands in Ost- und Westdeutschland vorgenommen wurde hat allerdings gezeigt, dass das erste Item keine identischen Skaleneinheiten zwischen Ost- und Westdeutschland sowie zwischen Ostdeutschland und Großbritannien aufweist. Diese Feststellung ist ebenfalls zutreffend für den Vergleich zwischen Ostdeutschland und Österreich, wobei allerdings mehr als zwei Items keine identischen Skaleneinheiten aufweisen.

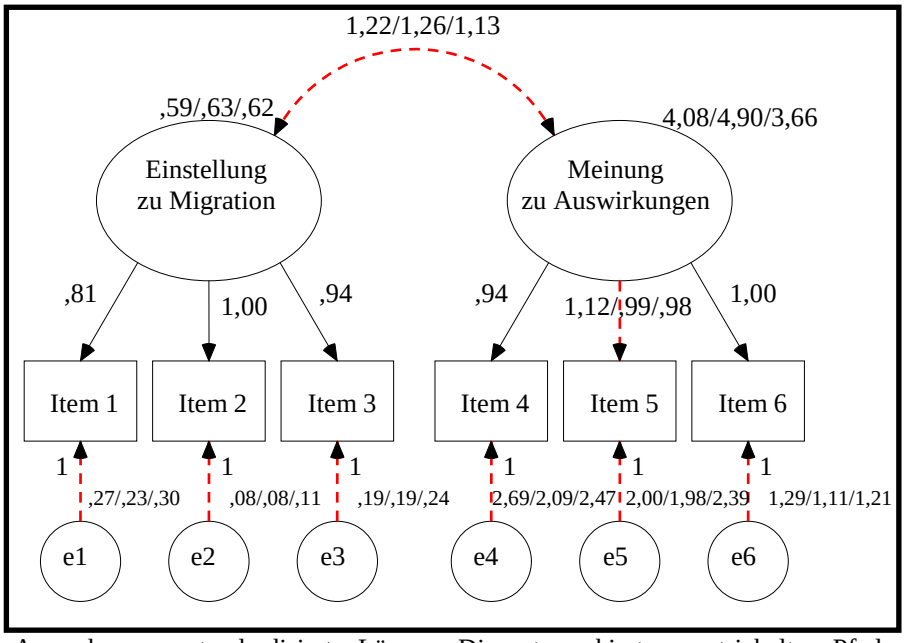
Tabelle 17: Gütemaße multipler Gruppenvergleich Großbritannien, Gesamtdeutschland und Österreich (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	5,076	,992	,978	,981	,024	-	-	-
Faktorladungen part. invariant	4,222	,991	,982	,982	,022	4,838	6	,565

Anmerkung: Es wurden zwischen den drei Gruppen alle Regressionskoeffizienten bis auf den des fünften Items fixiert. Das erste Modell enthält 24 Freiheitsgrade und das mit den invarianten Faktorladungen 30. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 4,8; die bei einer Differenz von 6 Freiheitsgraden nicht signifikant ist.

Der Vergleich zwischen Großbritannien, Österreich und dem „wiedervereinten" Deutschland zeigt, dass fünf Items dieselben Maßeinheiten in diesen Ländern aufweisen. (vgl. Abb. 6) Weil beim Item „country's cultural life undermined or enriched by immigrants" in Österreich ein *Item bias* vorliegt, ist nicht bei beiden Skalen eine *Measurement equivalence* existent.

Abbildung 6: multipler Gruppenvergleich Österreich, Großbritannien und Gesamtdeutschland



Anmerkung: unstandardisierte Lösung, Die rot markierten gestrichelten Pfade sind freigesetzt und die schwarzen Pfade wurden invariant gesetzt. Der erste Parameterwert bei den freigesetzten Pfaden entspricht dem Wert für Österreich, der zweite Wert dem Wert für Großbritannien und der dritte Wert dem für Gesamtdeutschland.
Lesebeispiel: Das Item 1 weist in allen drei Gruppen einen Regressionskoeffizienten von 0,81 auf den Faktor Einstellung zu Migration auf.

7. Zusammenfassung und Diskussion der Ergebnisse

Es werden in dieser Arbeit verschiedene Arten von international vergleichender Forschung dargestellt, um letztlich die für die Sozialwissenschaften bedeutendste Erscheinungsform, die der international vergleichenden Umfrageforschung auszuwählen. Das Ziel dieser Arbeit war die Überprüfung, ob und in welchem Ausmaß eine Vergleichbarkeit von Einstellungen und Meinungen zu Immigration gemäß des Konzepts der Äquivalenz, zwischen Deutschland, Großbritannien und Österreich gegeben ist. Deutschland wurde hierbei in zwei separate Analyseeinheiten aufgeteilt, um möglichst homogene Untersuchungseinheiten für die Analyse im Sinne eines *Most Different Systems Design* zu verwenden.

Die Daten, die es zu überprüfen galt, entstammen der dritten Erhebungswelle des European Social Surveys. Der ESS hat auf Grund ressourcenintensiver Qualitätskontrollen, den Ruf einer qualitativ hochwertigen international vergleichenden Umfragestudie. Gewissermaßen kann der ESS als Vorzeigeprojekt in der international vergleichenden Umfrageforschung betrachtet werden. So kann anhand dieser Studie gut demonstriert werden, dass ein Reihe von Anstrengungen unternommen wurden, um Verzerrungen zu minimieren, deren Existenz allerdings nicht ganz eliminiert werden kann. Der zweite Grund für die Auswahl dieser Studie ist hingegen praktischer Natur. Die lückenlose Dokumentation des Forschungsablaufes ermöglicht eine sehr detaillierte Analyse der erreichten Äquivalenz. Gemessen wurde diese anhand von zwei exemplarisch ausgewählten Skalen des ESS, welche die Einstellung zu Immigration sowie die Meinung zu den Auswirkungen von Migration der Befragten erfassen.

Es gibt zahlreiche Fehlerquellen, die die Äquivalenz der Daten in einem unterschiedlichen Ausmaß gefährden und dazu führen können, dass bei Vergleichen zwischen verschiedenen Ländern entdeckte Unterschiede und Gemeinsamkeiten messbedingte und keine realen Unterschiede darstellen. Die Äquivalenz ist hierbei in verschiedenen Phasen des Forschungsprozesses gefährdet. Dies fängt bei der Formulierung der Forschungsfrage an und reicht bis zur Kodierung der gegebenen Antworten. So existiert zum Beispiel die Möglichkeit, dass ein Konstrukt, welches verglichen werden soll eine unterschiedliche Beschaffenheit in verschiedenen Ländern

aufweist. Vergleiche zwischen verschiedenen Ländern können allerdings auch durch methodologische Aspekte einer Studie verzerrt sein. Das beinhaltet zum Beispiel die Nichtvergleichbarkeit von Substichproben einzelner Länder oder systematisch unterschiedliche Reaktionen von Befragten auf verschiedene Eigenschaften eines Erhebungsinstruments, wie etwa Antwortverzerrungen auf Grund unterschiedlicher semantischer Bedeutungen von Frageinhalten. Diese Fehlerquellen stellen keine inhärenten Charakteristiken eines Erhebungsinstruments beziehungsweise Studiendesigns dar, sondern entstehen erst in der Anwendung. Solche Fehlerquellen können nur beschränkt durch Pretests ausgeschaltet werden. Um die Wahrscheinlichkeit des Auftretens von Verzerrungen zu minimieren, müssen zusätzlich Methodenexperimente, wie zum Beispiel Multitrait-Multimethod- oder Split Ballot Verfahren eingesetzt werden. Diese führen im Idealfall zur Entwicklung eines adäquaten Fragebogendesigns. Es ist außerdem notwendig, dass nach der Erhebung nachträglich diverse statistische Verfahren eingesetzt werden, um die Äquivalenz der mittels einer Umfrage erhobenen Daten zu untersuchen. So wird deutlich, dass eine Vergleichbarkeit von gemessenen Werten nicht einfach angenommen werden darf, sondern vielmehr nachgewiesen werden muss.

Für die Beantwortung der Frage, welches Ausmaß der Äquivalenz hinsichtlich der untersuchten Skalen in den vier Analyseeinheiten vorliegt, wurden mehrere statistische Verfahren angewandt. Es wurden die Stichprobendesigns, die semantischen und pragmatischen Bedeutungen der Skalenitems, die Verteilungen in den Antworten, die interne und externe Konsistenz der Skalen, die Dimensionalität, die Faktorenstruktur und die Identität der Skalenwerte der zwei Skalen verglichen. Die Erkenntnisse, die mittels dieser einzelnen Analyseschritte gewonnen wurden, werden im folgenden Kapitel zusammenfassend dargestellt.

7.1. Ergebnisse

Beim ESS wurde eine effektive Stichprobengröße vorgegeben, um eine vergleichbare Präzision zwischen den Stichproben der einzelnen Länder herzustellen. Während in Österreich die angepeilte Stichprobengröße realisiert werden konnte, weisen die Stichproben Großbritanniens und Deutschlands weitaus geringere Stichprobengrößen auf. Dies ist auf die unterschiedliche Einhaltung der Vorgaben bezüglich der Kontaktversuche zur Verringerung der Nonresponse-Quote in den Ländern zurückzuführen. Diese Unterschiede wurden durch die Implementierung von Designgewichten neutralisiert. Weitere Verzerrungen der Stichproben wurden nicht entdeckt. Allerdings konnte die Existenz eines *Nonresponse errors* auf Grundlage der zur Verfügung gestellten Daten auch nicht endgültig ausgeschlossen werden. Relativierend muss hierzu erwähnt sein, dass dies bei den meisten nationalen Studien ebenfalls nicht möglich ist.

Die Betrachtung der Übersetzungen des Erhebungsinstruments hat gezeigt, dass die größten Unterschiede zwischen den beiden Fragebögen in der eigentlich selben Sprache bestehen. Die Gründe für die gefundenen Unterschiede zwischen den Fragebögen konnten im Rahmen der Diplomarbeit nicht geklärt werden, da keine Informationen über den Übersetzungsprozess verfügbar sind. Diese Unterschiede haben jedoch keine Auswirkungen auf die erreichte Ebene der Äquivalenz. Denn es hat sich gezeigt, dass in einem ähnlichen Ausmaß Unterschiede bezüglich der erreichten Ebene der Äquivalenz zwischen Ost- und Westdeutschland bestehen, wie zwischen den restlichen Analyseeinheiten. Also kann eine Verzerrung durch unterschiedliche sprachliche Formulierungen der Items und Antwortskalen ausgeschlossen werden. Um Verzerrungen auf Grund mangelnder linguistischer Äquivalenz auszuschließen, wäre es letztlich sinnvoll, zusätzliche Methoden anzuwenden. Eine der im Rahmen eines *post hoc* Ansatzes möglichen Vorgangsweisen, ist die Durchführung von kognitiven Interviews in den betreffenden Ländern. Bei kognitiven Interviews wird den Befragten der Fragebogens vorgelegt, mit der Intention, durch zusätzliche mündliche Informationen Probleme der Fragebogenkonstruktion aufzudecken, indem man einen Einblick „in die kognitiven Prozesse bekommt, die beim Beantworten von Fragen

ablaufen (Prüfer/Rexroth 2005: 3)“.[74] Werden bei der Durchführung von kognitiven Interviews keine systematischen länderspezifischen Unterschiede sichtbar, können semantische und pragmatische Bedeutungsunterschiede der Inhalte der Fragen und Antwortkategorien ausgeschlossen werden.[75] In dieser Diplomarbeit wurde auf die Durchführung solcher Interviews allerdings verzichtet.

Die Untersuchung, ob länderspezifischen Variationen bezüglich inhaltsunabhängigen Antworttendenzen existieren, hat eine Schwäche des Fragebogendesigns offenbart. Da die Skalen aus lediglich drei Items bestehen, waren diese alle in dieselbe Richtung gepolt. Um eine Antworttendenz bei Befragten festzustellen, wäre es notwendig, dass die Items der Skalen in eine unterschiedliche Richtung gepolt sind. Dies stellt allerdings kein spezifisches Problem eines Vergleichs dar, sondern zeigt, dass für die Entdeckung und Bestimmung der Ursachen von Verzerrungen auch ein angemessenes Fragebogendesign notwendig ist. Es wurden bei den Meinungen zu den Auswirkungen der Migration länderspezifische Variationen bei den Anteilswerten der Befragten, die bei allen Items konsistente Antworten gegeben haben, festgestellt. Ob bei den befragten Personen, die bei einem jeden Item dieselbe Antwortkategorie gewählt haben, nun eine inhaltsunabhängige Antworttendenz vorliegt und dies den Vergleich der Meinung zu den Auswirkungen der Migration verzerrt, bleibt letztlich auf Grund der Skalenkonstruktion unklar.

Der entscheidende Schritt bei der Überprüfung des erreichten Ausmaßes der Äquivalenz der zwei Skalen bestand in der Anwendung von Verfahren, die auf Korrelationen und Kovarianzen basieren. So zeigten die Ergebnisse dieser Verfahren, dass beide Skalen eine hohe interne Konsistenz aufweisen, die sich auf einem vergleichbaren Niveau in allen Untersuchungseinheiten befindet. Die Skalen messen in allen Analyseeinheiten jeweils eindimensionale Konstrukte, zwischen denen überall ein sehr hoher Zusammenhang besteht. Es wurde letztlich festgestellt, dass die Skalen überall eine vergleichbare Struktur aufweisen. Dies bedeutet, dass die beiden Skalen in den Analyseeinheiten funktional äquivalent sind, also die von ihnen gemessenen Konstrukte

[74] Für eine Darstellung diverser kognitiver Interviewtechniken vgl. Prüfer/Rexroth (2005) und Ahola (2004). Harkness/Schoua-Glusber stellen hingegen andere Methoden dar, um die linguistische Äquivalenz zu überprüfen. (vgl. 1998) Diese betreffen aber fast alle die Phase der Fragebogenkonstruktion.

[75] Es bestehen aber auch einige Problemfelder bezüglich der Validität der Erkenntnisse kognitiver Interviews. (vgl. Miller u.a. 2005, Beatty 2004)

eine vergleichbare Beschaffenheit aufweisen. Eine gemeinsame Grundlage und somit eine Vergleichbarkeit der mit den beiden Skalen erhaltenen Ergebnissen zwischen den Analyseeinheiten ist somit gegeben. Da für die beiden Skalen in den Analyseeinheiten die grundlegende Ebene der Äquivalenz, die funktionale Äquivalenz, attestiert wurde, können statistische Verfahren die auf Korrelationen und Kovarianzen basieren angewandt und deren Ergebnisse zwischen den Ländern interpretiert werden. Die so gefundenen Unterschiede und Gemeinsamkeiten von Korrelationskoeffizienten und Faktorenstrukturen zwischen Großbritannien, Österreich und West- und Ostdeutschland bezüglich der Einstellung zu Immigration und der Meinung zu den Auswirkungen von Migration sind einstellungsbedingt und nicht messbedingt.

Eine höhere Ebene der Äquivalenz der Skalen, also eine *Measurement unit equivalence*, ist zwischen Großbritannien, Österreich, Ost- und Westdeutschland allerdings nicht gegeben. Es bestehen nämlich keine identischen Skaleneinheiten zwischen allen Analyseeinheiten, da eine Verzerrung in zumindest einer der Analyseeinheiten existiert. Da die Messeinheiten der Items nicht äquivalent sind, können keine niveauorientierte Aussagen über Unterschiede bei den Einstellungen und Meinungen zu Immigration zwischen diesen Ländern getroffen werden. Über alle Untersuchungseinheiten betrachtet, weisen die zwei Skalen also lediglich eine funktionale Äquivalenz auf.

Um festzustellen, zwischen welchen Analyseeinheiten eine *Measurement unit equivalence* besteht, beziehungsweise welche Items Antwortverzerrungen verursachen, wurden die Analyseeinheiten in einem weiteren Analyseschritt paarweise verglichen. Dies hat gezeigt, dass lediglich für Großbritannien und Westdeutschland eine *Measurement unit equivalence* der beiden Skalen vorliegt. Da identischer Skaleneinheiten bestehen, sind niveauorientierte Aussagen über länderspezifische Unterschiede und Gemeinsamkeiten bezüglich der Einstellungen und Meinungen zu Immigration möglich. Dies beinhaltet zum Beispiel die Durchführung von Mittelwertvergleichen mittels t-Tests und Varianzanalysen und von Vergleichen absoluter Häufigkeiten zwischen Großbritannien und Westdeutschland.

Die Untersuchung, ob zumindest einige der Items identische Skaleneinheiten in mehr als zwei Analyseeinheiten aufweisen, hatte den Zweck zu herauszufinden, für welche Konstrukte und Items eine *Measurement unit equivalence* besteht, beziehungsweise

welche Items die Erreichung eines höheren Ausmaßes der Vergleichbarkeit zwischen den Ländern verhindern. Dies hat gezeigt, dass zwischen Großbritannien und Ost- und Westdeutschland bei fünf der sechs Items eine *Measurement unit equivalence* vorliegt. Beim Item „allow immigrants of same race/ethnic group as majority" besteht jedoch ein *non uniform Item bias* in Ostdeutschland. Das Item weist also in Ostdeutschland eine im Vergleich zu den anderen Analyseeinheiten unterschiedliche Bedeutung für die Ausprägung der gemessenen Einstellung zur Migration auf. Da bezüglich der Meinungen zu den Auswirkungen der Migration zwischen den betreffenden Analyseeinheiten eine *Measurement unit equivalence* existiert, können deren Ausprägungen zwischen Großbritannien und Ost- und Westdeutschland direkt verglichen werden. Bei den Einstellungen zu Migration trifft dies nur auf Teilaspekte dieses Konstruktes zu. So liegen lediglich bei den Items „allow immigrants of different race/ethnic group as majority" und „allow immigrants from poorer countries outside Europe" identische Skaleneinheiten vor. Da aber lediglich eine funktionale Äquivalenz bezüglich dieses Konstrukts in den Analyseeinheiten besteht, dürfen nur Korrelationskoeffizienten und Faktorenstrukturen zwischen diesen verglichen werden.

In Großbritannien, Westdeutschland und Österreich weisen ebenfalls fünf der sechs Items identische Skaleneinheiten auf. Es existiert also bei fünf Items zumindest eine *Measurement unit equivalence* zwischen Großbritannien, Österreich und Westdeutschland. Die Erreichung einer vollständigen *Measurement unit equivalence* ist auf Grund einer Verzerrung des Items „country's cultural life undermined or enriched by immigrants" in Österreich durch einen *non uniform Item bias* nicht gegeben. Die Ausprägungen der Einstellung zu Migration können zwischen Großbritannien, Westdeutschland und Österreich direkt verglichen werden. Bei den Meinungen zu den Auswirkungen von Migration dürfen lediglich Korrelationskoeffizienten und Faktorenstrukturen verglichen werden.

Warum in Ostdeutschland und Österreich jeweils ein Item eine im Vergleich zu den anderen Analyseeinheiten unterschiedliche Bedeutung für die Ausprägung eines Konstrukts aufweist, kann im Rahmen dieser Diplomarbeit aus verschiedenen Gründen nicht mit Sicherheit beantwortet werden. Die Vorraussetzungen für eine solche Ursachenforschung werden im Rahmen des nächsten Kapitels diskutiert.

Da die Mehrheit der durchgeführten Vergleiche im Rahmen von Sekundäranalysen des ESS vermutlich ohne eine Aufteilung Deutschlands in Ost- und Westdeutschland durchgeführt werden, wurde in einem letzten Schritt überprüft, welches Ausmaß der Äquivalenz der zwei Skalen zwischen den drei Ländern festgestellt würde. Auf diese Weise wird untersucht, ob bei Analysen auf einer höheren Ebene Verzerrungen, die bei einer niederen Ebene der Analyse gefunden wurden verschwinden. Der Vergleich zwischen Großbritannien, Österreich und dem „wiedervereinten" Deutschland zeigt, dass fünf Items dieselben Maßeinheiten in diesen Ländern aufweisen. Weil beim Item „country's cultural life undermined or enriched by immigrants" in Österreich ein *non uniform Item bias* vorliegt, ist allerdings nicht bei beiden Skalen eine *Measurement equivalence* existent. Die Analyse bei der einer Aufteilung Deutschlands in Ost- und Westdeutschland vorgenommen wurde, hat allerdings gezeigt, dass beim Item „allow immigrants of same race/ethnic group as majority" keine identischen Skaleneinheiten zwischen Ostdeutschland und den anderen Analyseeinheiten existieren. Würde eine Untersuchung nur auf dieser Analyseebene stattfinden, würde dies zu falschen Rückschlüssen bezüglich der erreichten Ebene der Äquivalenz der Skalen führen.

7.2. Schlussfolgerungen

Im abschließenden Kapitel dieser Arbeit werden einige der gewonnenen Erfahrungswerte bezüglich der international vergleichenden Umfrageforschung dargestellt und gezeigt, welche Schlussfolgerungen daraus zu ziehen sind. Dies betrifft die Tauglichkeit der Analyseebene der Nationalstaaten für eine kausale Argumentationsstruktur, die Eignung der angewandten statistischen Methoden zur Überprüfung des Ausmaßes der Äquivalenz und die Gründe warum eine Identifizierung von Verzerrungen sich wesentlich einfacher gestaltet, als die Ursachen für deren Existenz zu bestimmen.

Tauglichkeit der Analyseebene der Nationalstaaten für eine kausale Argumentationsstruktur

Es wurden die Ergebnisse der Analyse des Ausmaßes an Äquivalenz zwischen Großbritannien, Österreich und Deutschland mit der Analyse verglichen, bei der Deutschland auf Grund theoretischer Überlegungen in zwei Analyseeinheiten aufgeteilt wurde. Dies hat gezeigt, dass die Untersuchung der erreichten Ebene der Äquivalenz auf der Analyseebene der Nationalstaaten betrachtet, zu falschen Rückschlüssen bezüglich der Vergleichbarkeit der Skalen führt. So würde bei Betrachtung auf Ebene der Länder ein höheres Ausmaß an Äquivalenz der beiden Skalen konstatiert werden, als dies tatsächlich zutreffend ist. Bei Betrachtung der Länder ist nämlich die Verzerrung eines Items zwischen Ostdeutschland und den restlichen Untersuchungseinheiten nicht sichtbar. Aber nur weil die auf einer niedrigeren Ebene der Analyse festgestellten Unterschiede bezüglich der Identität der Skaleneinheiten eines Items bei der Wahl einer höheren Ebene der Analyse verschwinden, heißt das nicht, dass diese nicht existieren. Dies verdeutlicht, dass Nationalstaaten nicht per se eine ideale Analyseebene für die international vergleichende Forschung darstellen.

So können zum Beispiel regionale Unterschiede innerhalb gewisser Länder als länderspezifischer Unterschiede aufgefasst werden. Bestimmte Länder können aber auch transnationale Subsysteme eines größeren internationalen Systems sein und dies kann Gemeinsamkeiten durch Diffusionsprozesse verursachen. (vgl. Scheuch 1993b) Nationalstaaten können also unter Umständen entweder zu große oder zu kleine Analyseeinheiten darstellen, um kausale Erklärungen für nationale Unterschiede und Gemeinsamkeiten zu finden. Die Schwierigkeit besteht in erster Linie darin herauszufinden, welche Einflussfaktoren letztlich für die Unterschiede der abhängigen Variable(n) verantwortlich sind. (vgl. Scheuch 1993a) Solange keine Klarheit darüber existiert, wie die einzelnen Länder beschaffen sind und welche möglichen Drittvariablen für Erklärungen von Unterschieden und Gemeinsamkeiten zwischen den Ländern vorliegen, werden keine plausiblen Erklärungen für diese möglich sein. (vgl. Scheuch 1993b) Außerdem ist die Varianz zumeist zwischen den Ländern geringer als die Variation innerhalb dieser, so dass Erklärungen auf Länderebene zum Teil nicht als plausible Erklärung für gefundene Unterschiede beziehungsweise Ähnlichkeiten herangezogen werden dürfen. (vgl. ebd.)

Daraus erwächst die Notwendigkeit, dass auch das erreichte Ausmaß der Äquivalenz von Messinstrumenten zwischen verschiedenen Gruppen innerhalb einzelner Nationalstaaten untersucht werden sollte, bevor Ergebnisse letztlich in ihrer Gesamtheit mit anderen Ländern verglichen werden. In dieser Arbeit wurde dies am Beispiel Deutschlands aufgezeigt, da bestehende politische Unterschiede durch die ungleiche historische Entwicklung und die zeitweise Teilung des Landes vermutet wurden. Regionale Unterschiede dieser Art sind auch für Österreich und Großbritannien nicht auszuschließen. Eine diesbezügliche Untersuchung wäre in weiteren Arbeiten zu dieser Thematik äußerst empfehlenswert. Eine solche Untersuchung setzt allerdings eine Vergleichbarkeit der Stichproben des ESS auf regionaler Ebene voraus, die für den ESS vermutlich nicht gegeben ist. So war eine Unterteilung Deutschlands in zwei separate Analyseeinheiten lediglich möglich, da für Ost- und Westdeutschland getrennte Stichproben gezogen wurden und diese mittels eines dementsprechenden Gewichtungsverfahrens äquivalent gestaltet wurden. Solche Gewichte auf regionaler Basis sind allerdings für Großbritannien und Österreich nicht im Datensatz des ESS enthalten. Dementsprechend sind valide Erkenntnisse bezüglich regionaler Unterschiede innerhalb dieser Länder nur bedingt möglich. Die Implementierung einer Gewichtung auf regionaler Ebene wäre definitiv eine Anregung für zukünftige Erhebungswellen des ESS. Denn so wären nicht nur internationale Vergleiche auf Ebene der Länder, sondern auch valide internationale Vergleiche zwischen verschiedenen Regionen, sowie intranationale Vergleiche möglich.

Eignung der Methoden zur Untersuchung des Ausmaßes an Äquivalenz

Die Vorgangsweise bei der Untersuchung der Äquivalenz in dieser Diplomarbeit bestand aus drei Phasen. Die explorative Phase, bei der die Dokumentation der einzelnen Forschungsvorgänge betrachtet wurde erwies sich als nützlich, da einige mögliche Fehlerquellen, wie zum Beispiel die Existenz von *Sampling error* und *Mode bias* ausgeschlossen werden konnten. Zudem wurde durch die Durchsicht der Dokumentierung ein besseres Verständnis für die diversen Bereiche des Forschungsablaufs gewonnen, was insgesamt zu einer effektivere Vorgangsweise bei der Untersuchung der erreichten Ebene der Äquivalenz beigetragen hat. Die Untersuchung der verschiedenen Übersetzungen der Skalen, hat einige semantische Bedeutungsunterschiede zwischen den Ländern offenbart. In diesem Zusammenhang

wäre allerdings eine Durchführung von kognitiven Interviews, um etwaige Probleme der Fragebogenkonstruktion in den betreffenden Ländern aufzudecken, die rationalere, wenn auch im Rahmen dieser Arbeit nicht realisierbare Alternative gewesen.

Die nützlichsten Verfahren der explorative Datenauswertungsphase stellten die Analyse der fehlenden Werte und die Betrachtung möglicher Antworttendenzen dar. Während mithilfe ersterer ein *Item-Nonresponse error* ausgeschlossen werden konnte, hat die Untersuchung der Antworttendenzen länderspezifische Variationen offenbart, die allerdings auf Grund eines mangelhaften Fragebogens nicht näher überprüft werden konnten. Die Betrachtung der Häufigkeitsauszählungen und Verteilungen der Antworten sowie der Vergleich der Mittelwerte haben sich für die Untersuchung der erreichten Äquivalenz allerdings als nicht zweckmäßig erwiesen. Es ging bei den einzelnen Verfahren der explorative Datenauswertungsphase aber auch darum, die Eignung der Daten für die Durchführung der eigentlich relevanten Verfahren zur Bestimmung der Äquivalenz in Erfahrung zu bringen.

Die korrelationsbasierten Verfahren, also die einfachen Korrelationsanalysen, die Verfahren zur Überprüfung der internen und externen Konsistenz der Skalen und die explorative Faktorenanalyse, welche zur Überprüfung der funktionalen Äquivalenz angewandt wurden, erwiesen sich im Nachhinein betrachtet als wenig hilfreich für die Untersuchung der erreichten Ebene der Äquivalenz. So waren zwar die mittels dieser Verfahren gewonnen Erkenntnisse bezüglich problematischer Items zumeist korrekt, es war allerdings auch mitunter unklar, von welcher der Analyseeinheiten die Verzerrung ausgeht. Es fehlen bei den korrelationsbasierten Verfahren Maßzahlen, um die erreichte funktionale Äquivalenz im Sinne einer vergleichbaren Faktorenstruktur beziehungsweise vergleichbaren Korrelationskoeffizienten zwischen den einzelnen Analyseeinheiten definitiv feststellen zu können. Zudem besteht bei diesen Methoden die Gefahr einen Fehler zweiter Art zu machen, in dem Items als verzerrt klassifiziert werden, die dies in Wahrheit nicht sind. (vgl. Van de Vijver 2003b) Die konfirmatorische Faktorenanalyse hat sich hingegen als die geeignetste der in dieser Diplomarbeit angewandten Methoden erwiesen, um die erreichten Ebene der Äquivalenz zu untersuchen. Bei diesem Verfahren können nämlich Gütemaße für das Ausmaß der Äquivalenz berechnet werden. Da dieses Verfahren im Gegensatz zu den anderen korrelationsbasierten Methoden auf den Kovarianzen der Items basiert, ist

zudem nicht nur die Überprüfung der Existenz einer funktionalen Äquivalenz, sondern auch die der höheren Ebenen der Äquivalenz möglich. Es existieren allerdings auch einige Nachteile bei der konfirmatorischen Faktorenanalyse, die allerdings nicht in Zusammenhang mit der Untersuchung der Äquivalenz in der vorliegenden Arbeit stehen. So ist zumeist eine aufwändige Vorgehensweise notwendig, um zu validen Ergebnissen zu gelangen. Zudem setzt die Durchführung dieser Methode relativ große Stichproben voraus. Sollten die Anforderungen bezüglich der Stichprobengröße nicht erfüllt sein, müsste auf andere weniger zuverlässige Methoden zurückgegriffen werden.

Die Problematik der Identifizierung der Ursachen von Verzerrungen

Bei der Durchführung der konfirmatorischen Faktorenanalyse wurde festgestellt, dass bei zwei Items ein *non uniform Item bias* vorliegt. Es ist jedoch nicht möglich die Ursachen zu identifizieren, warum das Item „Country's cultural life undermined or enriched by immigrants" in Österreich und das Item „allow immigrants of same race/ethnic group as majority" in Ostdeutschland eine zu den restlichen Analyseeinheiten unterschiedliche Bedeutung für die Ausprägung der jeweiligen Konstrukte aufweisen. Dieser Umstand verdeutlicht, dass es leichter ist, Verzerrungen zu entdecken als deren Ursachen zu bestimmen. Hierfür existieren mehrere Gründe. Die Ursachen von Fehlerquellen sind in diesem Zusammenhang als Drittvariablen zu aufzufassen, die Auswirkungen auf die Antworten der Befragten aufweisen. Es müsste letztlich eine statistische Kontrolle über etwaige Drittvariablen vorgenommen werden, um Verzerrungen von Vergleichen auszuschließen beziehungsweise die Ursachen für deren Auftreten zu entdecken. Dies ist durch den Umstand, dass zumeist eine hohe Anzahl möglicher Drittvariablen existiert, die sich mit zunehmender Anzahl und Heterogenität der Länder, die in eine Analyse aufgenommen werden vervielfacht, kein leichtes Unterfangen. Um eine statistische Kontrolle von Drittvariablen im Rahmen einer Sekundäranalyse eines bestehenden Datensatzes zu ermöglichen, wäre es notwendig, dass diese als Variablen im Erhebungsinstrument enthalten sind. Zudem ist es notwendig, dass zumindest einige theoretische Konzepte mittels mehrerer Indikatoren gemessen werden, da nur so die in dieser Arbeit vorgestellten statistischen Verfahren angewandt werden können. Um die Ursachen für entdeckte Verzerrungen auf der Ebene der Items und des Instruments zu identifizieren, müssen diese schon im

Studiendesign berücksichtigt werden. Das Erhebungsinstrument sollte daher so gestaltet werden, dass die Ursachen von Verzerrungen besser beziehungsweise überhaupt zu identifizieren sind. Wenn eine Reihe von relevanten Drittvariablen im Datensatz enthalten sind, können diese zum Beispiel in ein konfirmatorisches Faktorenmodell integriert werden. Somit kann der Einfluss dieser Störeinflüsse auf länderspezifische Unterschiede bemessen und die Ursachen für Verzerrungen von Vergleichen bestimmt werden. Je mehr alternative Erklärungen auf diese Weise ausgeschlossen werden, desto sicherer ist der Befund, dass es sich um tatsächlich existierende Länderunterschiede und nicht um Messartefakte handelt. Wenn sich zeigt, dass keine oder nur ein geringes Ausmaß an Äquivalenz besteht, sollte jedoch nicht gänzlich auf Vergleiche zwischen den betreffenden Ländern verzichtet werden. Denn auch der Befund, dass ein Vergleich verzerrt ist, kann eine Erkenntnis darstellen und auf interessante Länderunterschiede hinweisen. So kann zum Beispiel der Befund, dass in den betrachteten Ländern gewisse Items eine unterschiedliche Bedeutung für ein Konstrukt aufweisen, auf verschiedene Kontexteinflüsse hinweisen, die Gegenstand weiteren Untersuchungen sein sollten, um die Ursachen hierfür herauszufinden.

Das Ziel einer jeden international vergleichenden Umfragestudie sollte die Erreichung einer möglichst hohen Ebene der Äquivalenz sein. Diese Diplomarbeit hat jedoch gezeigt, dass dies kein leichtes Vorhaben darstellt, da eine Vielzahl an möglichen Fehlerquellen existiert, die die Ergebnisse eines internationalen Vergleichs verzerren können. Es wurde deutlich, dass die Anwendung von statistischen Methoden nur einen Teilbeitrag zur Untersuchung der erreichten Ebene der Äquivalenz im Rahmen einer Sekundäranalyse leisten kann. Um Verzerrungen von Vergleichen wirklich zu minimieren, ist zudem auch ein adäquates Studiendesign und die Anwendung diverser *a priori* Techniken, wie zum Beispiel Pretests und Methodenexperimente, vonnöten. Nur so kann letztlich eine Minimierung der Verzerrungen beziehungsweise eine Maximierung der Äquivalenz der Daten sichergestellt werden.

8. Literaturverzeichnis

Adamsons, K./Buehler, C. (2007): „Mothering versus Fathering versus Parenting: Measurement Equivalence in Parenting Measures", in: *Parenting: Science and Practice 7, 3*, S.271-303.

Ahola, A. (2004): „Cognitive Model of the Question-Answering Process and Development of Pretesting", in: Prüfer, P./Rexroth, M./Fowler, F.J.Jr. (Hgg.): *Quest 2003: Proceedings of the 4th Conference on Question Evaluation Standards*, Zuma-Nachrichten: Nachrichten Spezial Band 9, Mannheim: ZUMA, S. 26-33.

Alford, R. (1963): *Party and Society*, Chicago: University of Chicago Press.

Atteslander, P. (2003): *Methoden der empirischen Sozialforschung*, 10. *neu bearbeitete* und überarbeitete Auflage, Berlin: Walter de Gruyter.

Backhaus, K./Erichson, B./Plinke, W./Weiber, R. (2006): *Multivariate Analysemethoden*, 11. überarbeitete Auflage, Berlin: Springer.

Beatty, P. (2004): „Paradigms of Cognitive Interviewing Practice, and Their Implications for Developing Standards of Best Practice", in: Prüfer, P./Rexroth, M./Fowler, F.J.Jr. (Hgg.): *Quest 2003: Proceedings of the 4th Conference on Question Evaluation Standards*, Zuma-Nachrichten: Nachrichten Spezial Band 9, Mannheim: ZUMA, S. 8-25.

Billiet, J. (2003): „Cross-Cultural Equivalence with Structural Equation Modeling", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 247-263.

Billiet, J./Koch, A./Philippens, M. (2007): „Understanding and improving response rates", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 113-138.

Blair, J./Piccinino, L. (2005): „The Development and Testing of Instruments for Cross-Cultural and Multi-Cultural Surveys", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 13-30.

Borg, I. (1998): „A Facet-Theoretical Approach to Item Equivalency", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence*, Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.145- 158.

Bortz, J./Döring, N. (1995): *Forschungsmethoden und Evaluation*, 2. vollständig überarbeitete und aktualisierte Auflage, Berlin: Springer.

Braun, M./Scott, J. (1998): „Multidimensional Scaling and Equivalence", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence,* Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.129- 144.

Braun, M. (2000): „Evaluation der Äquivalenz eines gemeinsamen Satzes an Indikatoren in der interkulturell vergleichenden Sozialforschung", *ZUMA How-to-Reihe, 3,* Mannheim: ZUMA.

Braun, M. (2003a) „Communication and Social Cognition", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 57-67.

Braun, M. (2003b): „Errors in Comparative Survey Research: An Overview", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 137-142.

Braun, M./Mohler, P. Ph. (2003): „Background Variables", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 101-115.

Braun, M./Harkness, J.A. (2005): „Text and Context: Challenges to Comparabiltiy in Survey Questions", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research,* Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 95-107.

Brosius, F. (1998): *SPSS 8,* Hamburg: International Thomson Publishing.

Bühner, M. (2004): *Einführung in die Test- und Fragebogenkonstruktion,* München: Pearson Studium.

Currle, E (2004): *Migration in Europa- Daten und Hintergründe,* Stuttgart: Lucius & Lucius.

Couper, M.P./Leeuw, E.D. (2003): „Nonresponse in Cross-Cultural and Cross-National Surveys", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 157-176.

Dean, E./Caspar, R./McAvinchey, G./Reed, L./Quiroz, R. (2005): „Developing a Low-Cost Technique for Parallel Cross-Cultural Instrument Development: The Question Appraisal System (QAS-04)", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research,* Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 31-46.

De Wit, H./ Billiet, J. (1995): „The MTMM design: back to the Founding Fathers" in: Saris, W.E./Münnich A. (Hgg.): *The Multitratit-Multimethod Approach to Evaluate Measurement Instruments,* Budapest: Eötvös University Press,S. 39-59.

Diekmann, A. (2004): *Empirische Sozialforschung- Grundlagen, Methoden Anwendungen,* König, B. (Hg.), 12. Auflage, Reinbek: Rowohlt Taschenbuch Verlag

Durkheim, E. (1976): *Die Regeln der soziologische Methode*, König, R. (Hg.), 4. revidierte Auflage, Neuwied: Herman Luchterhand Verlag.

Esposito, J.L. (2004): „With Regard to the Design of Major Statistical Surveys: Are We Waiting Too Long to Evaluate Substantive Questionnaire Content?", in: Prüfer, P./Rexroth, M./Fowler, F.J.Jr. (Hgg.): *Quest 2003: Proceedings of the 4th Conference on Question Evaluation Standards*, Zuma-Nachrichten: Nachrichten Spezial Band 9, Mannheim: ZUMA, S. 161 -171.

Esser, H. (1986): „Können Befragte lügen? Zum Konzept des „wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung", *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 38,2, S. 314-336.

Fowler, F.J.Jr. (2004): „More on the Value of Split Ballots", in: Prüfer, P./Rexroth, M./Fowler, F.J.Jr. (Hgg.): *Quest 2003: Proceedings of the 4th Conference on Question Evaluation Standards*, Zuma-Nachrichten: Nachrichten Spezial Band 9, Mannheim: ZUMA,, S. 43-51.

Fontaine, J. (2003): „Multidimensional Scaling", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 235-246.

Groves, R.M. (1989): *Survey Errors and Survey Costs*, New York: John Wiley and Sons.

Hadler M. (2007): *Soziale Ungleichheit im internationalen Vergleich: ihre Wahrnehmung, ihre Auswirkung und ihre Determinanten*, Wien: Lit-Verlag (Austria: Forschung und Wissenschaft: Soziologie: Bd. 4).

Hantrais, L./Mangen, S. (Hgg.) (1998): „Method and management of cross-national social research.", in: *Cross-national research methods in the social sciences*, London: Pinter, S.1-12.

Harkness, J.A./Schoua-Glusber, A. (1998): „Questionnaires in Translation", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence*, Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.87- 128.

Harkness, J.A. (2003): „Questionnaire Translation", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 35-56.

Harkness, J.A./Van de Vijver, F.J.R./Johnson, T.P. (2003): „Questionnaire Design in Comparative Research", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 19-34.

Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (2003a): „Comparative Research", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 3-16.

Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.) (2003b): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology).

Harkness, J.A. (2007): „Improving the comparability of translations", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 79-94.

Haslinger. A./Kytir J (2006): „Stichprobendesign, Stichprobenziehung und Hochrechnung des Mikrozensus ab 2004", in: *Statistische Nachrichten, 6*, S. 510-518.

Häder, S. (2000): „Telefonstichproben", *ZUMA How-to-Reihe, 6,* Mannheim: ZUMA.

Häder, S./ Gabler, S. (2003): „Sampling and Estimation", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 117-134.

Häder, S./Lynn, P. (2007): „How representative can a multi-nation survey be?", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S.33-52.

Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.) (2005): „Methodological Aspects in Cross-National Research: Foreword", in: *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 5-10.

Hoffmeyer-Zlotnik, J.H.P./Warner, U. (2008): *Privater Haushalt: Konzepte und ihre Operationalisierung in nationalen und internationalen sozialwissenschaftlichen Umfragen*, Mannheim: Forschung Raum und Gesellschaft.

House,R./ Javidan,M./ Hanges,P./ Dorfman,P. (2002): „Understanding cultures and implicit leadership theories across the globe: an introduction to project GLOBE", in: *Journal of world Business,37,* S.3-10.

Johnson, T.P./O'Rourke, D./Chavez, N./Sudman, S./Warneke, R./Lacey, L. (1997): „Social Cognition and Responses to Survey Questions among Culturally Diverse Populations" in: Lyberg, C./Biemer, P./Collin, M./Dippo, C./deLeeuw, E/Schwartz, N./Trewin, D. (Hgg.): *Survey Measurement and Process Quality*, New York: Wiley, S. 87-114.

Johnson, T.P (1998): „Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence,* Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.1- 39.

Johnson, T.P. (2003): „Glossary", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods,* Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 347-357.

Johnson, T.P./Van de Vijver, F.J.R. (2003): „Social Desirability in Cross-Cultural Research", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 195-204.

Johnson, T.P./Cho, Y.I./Holbrook, A./O'Rourke, D./Warnecke, R./Chávez, N. (2005): „Cultural Variability in the Effects of Question Design Features on Respondent Comprehension",in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 65-78.

Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.) (2007): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications.

Jowell, R./Kaase, M./Fitzgerald, R./Gillian, E. (2007): „The European Social Survey as a measurement model", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 1-31.

Kim, J.O./ Müller, C.W. (1978): *Factor analysis. Statistical methods and practical issues*, Newbury Park: Sage.

Kleiner, P./Pan, Y. (2006): „Cross-Cultural Communication and the Telephone Survey Interview" in: Harkness, J.A. (Hg.): *Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, Zuma-Nachrichten: Nachrichten Spezial Band 12, Mannheim: ZUMA, S. 81-90.

Koch, A./Blohm, M. (2006): „Fieldwork Details in the European Social Survey 2002/2003" in: Harkness, J. A. (Hg.): *Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, Zuma-Nachrichten: Nachrichten Spezial Band 12, Mannheim: ZUMA, S. 21-52.

Kohn, M.L. (1987): „Cross-national Research as an Analytic Strategy.", in: *American Sociological Review 52,6*, S.713-731.

Kolsrud, K./Skjåk, K.K. (2005): „Harmonising background variables in the European Social Survey", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 163-181.

Kolsrud, K./Skjåk, K.K./Henrichsen, B. (2007): „Free and immediate acces to data", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 139-156.

Lambert, P.S. (2005): „Ethnicity and the Comparative Analysis of Contemporary Survey Data", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 259-278.

Lass, J. (1997): „Telephone ownership - a cause of sampling bias in Europe?", in: Saris, W.E./ Kaase, M. (Hgg.): *Eurobarometer: Measurement Instruments for Opinions in Europe*, Zuma-Nachrichten: Spezial Band 2, Mannheim: ZUMA, S.45-63.

Leeuw, E.D./Hox, J.J. (2003): „The Use of Meta-Analysis in Cross-National Studies", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 329-345.

Likert, R. (1932): „A Technique for the Measurements of Attitudes", in: Woodworth, R.S. (Hg.): *Archives of Psychology, 140*, New York, S. 1-55.

Lobe, B./Livingstone, S./Haddon, L. (2007): *Researching Children's Experiences Online across Countries: Issues and Problems in Methodology*, London: EU Kids Online.

Lobe, B./Livingstone, S./ Olafsson, K./Simões, J.A. (2008): *Best Practice Research Guide: How to research children and online technologies in comparative perspective*, London: EU Kids Online.

Lynn, P./Japec, L./Lyberg L. (2006): „What's So Special About Cross-National Surveys?" in: Harkness, J.A. (Hg.): *Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, Zuma-Nachrichten: Nachrichten Spezial Band 12, Mannheim: ZUMA, S. 7-20.

MacInnes, J. (2006): „Category and Comparison Across What Kind of Frontier?" in: Harkness, J.A. (Hg.): *Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, Zuma-Nachrichten: Nachrichten Spezial Band 12, Mannheim: ZUMA, S. 101-114.

Mayerl, J. (2009): *Kognitive Grundlagen sozialen Verhaltens – Framing, Einstellungen, Rationalität*, Wiesbaden: VS Verlag für Sozialwissenschaften.

Miller, K./Willis, G./Eason, C./Moses, L./Canfield, B. (2005): „Interpreting the Results of Cross-Cultural Cognitive Interviews: A Mixed-Method Approach", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 79-92.

Mohler, P.Ph./Smith, T.W./Harkness, J.A. (1998): „Respondents' Ratings of Expressions from Response Scales: A Two-Country, Two-Language Investigation on Equivalence and Translation", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence*, Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.159- 184.

Mohler, P.Ph./Uher, R. (2003): „Documenting Comparative Surveys for Secondary Analysis", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 311-325.

Mohler, P. (2007): „What is being learned from the ESS?", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 157-168.

Moors, G. (2008): „Exploring the effect of a middle response category on response style in attitude measurement", in: *Quality and Quantity, Vol. 42, Ausgabe 6,* S. 779-794.

Moore, B.Jr. (1966): *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*, Boston: Beacon.

Neller, K. (2005): „Kooperation und Verweigerung: Eine Non-Response-Studie", in: Wolf, C. (Hg.): *ZUMA-Nachrichten, 57*, S. 9-36.

Newton, K./Montero, J.R. (2007): „Patterns of political and social participation in Europe", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 205-238.

Norris, P./Davis, J. (2007): „A continental divide? Social Capital in the US and Europe", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 239-264.

Prüfer, P./Rexroth, M. (2005): „Kognitive Interviews", *ZUMA How-to-Reihe, 15*, Mannheim: ZUMA.

Przeworski, A./Teune, H. (1970): *The Logic of Comparative Social Inquiry*, New York: Wiley-Interscience.

Ragin, C./Zaret, D. (1983): „Theory and Method of Comparative Research Two Strategies", in*: Social Forces, 61,3,* S. 731-754.

Ragin, C.C. (1987): *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*, Berkeley: University of California Press.

Ragin, C.C. (1994): *Constructing Social Research: The Unity and Diversity of Method*, Thousand Oaks: Pine Forge Press.

Rammstedt, B. (2004): „Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung", *ZUMA How-to-Reihe, 12*, Mannheim: ZUMA.

Reeskens, T./Hooghe, M. (2008): „Cross Cultural measurement of generalized trust. Eviodence from the European Social Survey (2002 und 2004)", in: *Social Indicators Research, 85, 3*, S. 515-532.

Reinecke, J. (2005): *Strukturgleichungsmodelle in den Sozialwissenschaften*, München: Oldenbourg Verlag.

Reuband, K.H. (2002): „Frageformen, themenspezifische Sensibilitäten und Antwortmuster: Wie Fragen in Statementform und in Form dichotomer Antwortvorgaben Antwortmuster beeinflussen“, in: Zentralarchiv für empirische Sozialforschung (Hg.): *ZA Information, 51*, S. 82-99.

Rokkan, S.(1993): „Cross-cultural, cross-societal and cross-national research”, in: *Historical Social Research, 18(No. 2 = No. 66),* S. 6-54.

Rosar, U. (2004): „Ethnozentrosmus und Immigration“, in: Van Deth, J.W.(Hg.): *Deutschland in Europa. Ergebnisse des European Social Survey 2002 – 2003*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 77-102.

Ross; C.E./Mirowsky, J. (1984): „Socially-Desirable Response and Acquiescence in a Cross-Cultural Survey of Mental Health”, in: *Journal of Health and Social Behavior, 25*, S.189-197.

Rother, N. (2005): „Measuring Attitudes Towards Immigration Across Countries with the ESS: Potential Problems of Equivalence”, in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 109-126.

Saris, W.E. (1995): „Designs and models for quality assessment of survey measures” in: Saris, W.E./Münnich A. (Hgg.): *The Multitratit-Multimethod Approach to Evaluate Measurement Instruments*, Budapest: Eötvös University Press,S. 9-37.

Saris, W.E./Münnich A. (1995): *The Multitratit-Multimethod Approach to Evaluate Measurement Instruments*, Budapest: Eötvös University Press.

Saris, W.E./Scherpenzeel, A.C. (1996): „Methodological procedures for comparative research”, in: Saris, W.E./Veenhoven, R./ Scherpenzeel, A.C./Bunting, B. (Hgg.): *A Comparative Study of Satisfaction with Life in Europe*, Budapest: Eötvös University Press, S. 49- 76.

Saris, W.E./Kaase, M. (Hgg.) (1997a): *Eurobarometer:Measurement Instruments for Opinions in Europe*, Zuma-Nachrichten: Nachrichten Spezial Band 2, Mannheim: ZUMA.

Saris, W.E./Kaase, M. (1997b): „Summary and Discussion“, in: Saris, W.E./Kaase, M. (Hgg.): *Eurobarometer:Measurement Instruments for Opinions in Europe*, Zuma-Nachrichten: Nachrichten Spezial Band 2, Mannheim: ZUMA, S.142-154.

Saris, W.E. (1998): „The Effects of Measurement Error in Cross-Cultural Research”, in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence,* Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.67- 86.

Saris, W.E. (2003a): „Multitrait – Multimethod Studies”, in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 265-274.

Saris, W.E. (2003b): „Response Function Equality", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 275-288.

Saris, W.E./Gallhofer, I. (2007a): „Can questions travel succesfully?", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 53-78.

Saris, W.E./Gallhofer I. (Hgg.) (2007b): *Design, evaluation, and analysis of questionnaires for survey research*, Hoboken: Wiley-Interscience (Wiley series in survey methodology).

Scherpenzeel, A.C. (1995): „Design of Meta Analysis" in: Saris, W.E./ Münnich A. (Hgg.): *The Multitratit-Multimethod Approach to Evaluate Measurement Instruments*, Budapest: Eötvös University Press, S. 187- 206.

Scheuch, E.K. (1993a): „The cross-cultural use of sample surveys: problems of comparability", in: *Historical Social Research, 18(No. 2 = No. 66)*, S. 104-138.

Scheuch, E.K. (1993b): „Theoretical implications of comparative survey research: why the wheel of cross-cultural methodology keeps on being reinvented", in: *Historical Social Research, 18(No. 2 = No. 66)*, S. 172-195.

Schnell, R./Hill, P.B/Esser E. (2008): *Methoden der empirischen Sozialforschung*, 8. unveränderte Auflage, München: Oldenbourg.

Scholz, E. (2005): „Harmonisation of survey data in the International Social survey Programme (ISSP)", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research*, Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 183-201.

Schröder, W.H. (1982): „Cross-national comparative research: some practical remarks", in: *Historical Social Research, 24*, S.111-121.

Schubert, P./Greil, A. (1997): „ Sample design and consequences", in: Saris, W.E./ Kaase, M. (Hgg.): *Eurobarometer: Measurement Instruments for Opinions in Europe*, Zuma-Nachrichten: Spezial Band 2, Mannheim: ZUMA, S.24-31.

Schwarz, N. (2003): „Culture-Sensitive Context Effects: A Challenge for Cross-Cultural Surveys", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 93-100.

Schwartz, S.H. (2007): „Value orientations: measurements, antecedents", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 169-204.

Singer, E./Hoewyk, J.v./Maher, M.P. (1998): „Does the Payment of Incentives create Expectation Effects", in: Koch, A./Porst, R. (Hgg.): *Nonresponse in Survey Research*, Zuma-Nachrichten: Nachrichten Spezial Band 4, Mannheim: ZUMA, S. 229 -238.

Skjåk, K.K./Harkness, J. A. (2003): „Data Collection Methods", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 179-194.

Smith, T.W. (2003): „Developing Comparable Questions in Cross-National Surveys", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 69-91.

Stoop, I. (2007): „If it bleads, it leads: the impact of media reported events", in: Jowell, R./Roberts, C./Fitzgerald, R./Gillian, E. (Hgg.): *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, Los Angeles: Sage Publications, S. 95-112.

Sudman, S. /Bradburn, N.M./Schwarz, N. (1996): *Thinking about Answers*, San Francisco: Jossey-Bass Publications.

Tannenbaum, E./Mochmann, E. (1996): „Toward a European Database for Comparative Social Research", in: *Historical Social Research, 21(No. 2 = No. 78)*, S. 118-125.

Treibel, A. (1990): *Migration in modernen Gesellschaften: soziale Folgen von Einwanderung, Gastarbeit und Flucht*, Weinheim: Juventa Verlag.

Van de Vijver, F.J.R./Leung, K. (1997): „Methods and data analysis of comparative research", in: Berry, J.W./Poortinga, Y.H./Pandey, J. (Hgg.): *Handbook of Cross-Cultural Psychology*, 2 .Auflage, Vol. 1, Boston: Allyn & Bacon, S.257-300.

Van de Vijver F.J.R. (1998): „Towards a Theory of Bias and Equivalence", in: Harkness, J.A. (Hg.): *Cross-cultural survey equivalence*, Zuma-Nachrichten: Spezial Band 3, Mannheim: ZUMA, S.41- 66.

Van de Vijver, F.J.R. (2003a): „Bias and Equivalence: Cross-Cultural Perspectives", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 143-155.

Van de Vijver, F.J.R. (2003b): „Bias and Substantive Analyses", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 207-233.

Van Deth, J.W. (2003): „Using Published Survey Data", in: Harkness, J.A./Van de Vijver, F.J.R./Mohler, P.Ph. (Hgg.): *Cross-Cultural Survey Methods*, Hoboken: Wiley-Interscience (Wiley series in survey methodology), S. 291-309.

Van Deth, J.W.(Hg.) (2004): *Deutschland in Europa. Ergebnisse des European Social Survey 2002 – 2003*, Wiesbaden: VS Verlag für Sozialwissenschaften.

Van Herk, H. (2000): *Equivalence in a Cross-National Context: Methodological and Empirical Issues in Marketing Research*, Tilburg: Tilburg University Press.

Verba, S. (1993): „The uses of survey research in the study of comparative politics: issues and strategies", in: *Historical Social Research, 18(No. 2 = No. 66)*, S. 55-103.

Warnecke, R./ Johnson, T.P./Chavez, N./ Sudman, S./ O'Rourke, D./Lacey, L./Horm, J. (1997): „Improving Question Wording in Surveys of Culturally Diverse Populations", in: *Annals of Epidemiology, 7*, S.334-342.

Warner, U./Hoffmeyer-Zlotnik, J.H.P. (2006): „Discussion of the Income Measure in the European Social Survey: A Proposal of Revised Survey Questions About the ‚Total Net Household Income'"in: Harkness, J.A. (Hg.): *Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI),* Zuma-Nachrichten: Nachrichten Spezial Band 12, Mannheim: ZUMA, S. 53-66.

Weins, C. (2008): „Möglichkeiten und Grenzen des internationalen Vergleichs fremdenfeindlicher Vorurteile", *Sozialwissenschaftlicher Fachinformationsdienst (soFid): Methoden und Instrumente der Sozialwissenschaften, 1*, S. 25-43.

Weiss, H./Reinprecht, C. (2004): *Nation und Toleranz? Empirische Studien zu nationalen Identitäten in Österreich,* Wien: Braunmüller.

Wolf, C. (2005): „Measuring religious affiliation and religiosity in Europe", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research,* Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 279-295.

Zaletel, M./Vehovar, V. (1998): „The Stability of Nonresponse Rates According to Socio- Demographic Categories", in: Koch, A./Porst, R. (Hgg.): *Nonresponse in Survey Research,* Zuma-Nachrichten: Nachrichten Spezial Band 4, Mannheim: ZUMA, S. 75 - 84.

Zucha, V. (2002): „Überprüfung der internationalen Vergleichbarkeit von Indikatoren zur Messung von Arbeitsorientierungen mittels konfirmatorischer Faktorenanalyse",in: Dutter, R. (Hg.): *Festschrift 50 Jahre Österreichische Statistische Gesellschaft,* Wien: Österreichische Statistische Gesellschaft, S. 171-192.

Zucha, V. (2004): *Interkulturelle Vergleichbarkeit von Einstellungsfragen zu sozialer Ungleichheit. Methoden der Überprüfung und Ursachen mangelnder Äquivalenz.* Wien: Dissertation.

Zucha, V. (2005): „The level of equivalence in the ISSP 1999 and its implications on further analysis", in: Hoffmeyer-Zlotnik, J.H.P./Harkness, J.A (Hgg.): *Methodological Aspects in Cross-National Research,* Zuma-Nachrichten: Nachrichten Spezial Band 11, Mannheim: ZUMA, S. 127-147.

8.1. Internetquellen

Baltes-Götz, B. (2008): „Analyse von Strukturgleichungsmodellen mit Amos 16.0“, Universitäts-Rechenzentrum Trier, www.uni-trier.de/fileadmin/urt/doku/amos/v16/amos16.pdf (22.09.09).

Billiet, J (o.J.): „Questions about National, Subnational and Ethnic Identity”, in: European Social Survey: „Questionnaire Development Report“, S. 384-418, http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=62&Itemid=96 (22.09.09).

Billiet, J./Welkenhuysen-Gybels, J. (2004): „Assessing Cross-National Construct Equivalence in the ESS: The case of religious involvement”, Paper prepared for presentation at the European Conference on Quality and Methodology in Official Statistics. Mainz: 24-26 May 2004, http://www.s3ri.soton.ac.uk/qmss/documents/BilletMainzpapermeasurementESSdefinite.pdf (22.09.09).

Brown, G./Micklewright, J./Schnepf, S.V./Waldmann, R. (2005): „Cross-National Surveys of Learning Achievement: How Robust are the Findings?”, IZA Discussion Paper No. 1652. http://ssrn.com/abstract=758327. (22.09.09).

Bundeskanzleramt Rechtsinformationssystem (2009): „Bundesgesetz vom 7. Juli 1976 über die Rechtsstellung von Volksgruppen in Österreich (Volksgruppengesetz)“, http://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10000602 (22.09.09).

Bundeszentrale für politische Bildung (o.J.): „Zahlen und Fakten- Die soziale Situation in Deutschland“, http://www.bpb.de/files/HXVH9I.pdf (22.09.09).

Carolin Reißlandt (2005): „Migration in Ost- und Westdeutschland von 1955 bis 2004“, Bundeszentrale für politische Bildung, http://www.bpb.de/themen/8Q83M7,0,0,Migration_in_Ost_und_Westdeutschland_von_1955_bis_2004.html (22.09.09).

European Social Survey (=ESS) (o.J.a): „Chapter III: The Questionnaire-Part 1“, http://www.europeansocialsurvey.org/index.php?option=com_docman&task=cat_view&gid=95&Itemid=80 (22.09.09).

European Social Survey (=ESS) (o.J.b): „Chapter IV: Translation“, http://www.europeansocialsurvey.org/index.php?option=com_docman&task=cat_view&gid=95&Itemid=80 (22.09.09).

European Social Survey (=ESS) (o.J.c): „ESS3 - Main questionnaire”, http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=questionaires (22.09.09).

European Social Survey (=ESS) (o.J.d): „Weighting European Social Survey Data",http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation (22.09.09).

European Social Survey Deutschland (=ESS Deutschland) (o.J.): „deutsche Fragebogenversion der dritten Welle des European Social Surveys, http://www.europeansocialsurvey.de/dokumentation/dritte.htm (22.09.09).

ESS Central Coordinating Team (2008): „European Social Survey Round 3. Measuring social and political change in Europe. Publishable Final Activity Report" Fitzgerald R./Widdop, S. (Hgg.), http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=547&itemid=80 (22.09.09).

Eurostat (Hg.) (2006) : Die ausländische Bevölkerung in den Mitgliedstaaten der EU, in: Statistik kurz gefasst- Bevölkerung und Soziale Bedingungen, 8, http://www.eds-destatis.de/de/downloads/sif/nk_06_08.pdf (22.09.09).

German Social Science Infrastructure Services (=GESIS) (2009): „Tabular History of International Comparative Survey Research Projects", http://www.gesis.org/en/services/data/portals-links/comparative-survey-projects/?0 (22.09.09).

Institut für Arbeitsmarkt- und Berufsforschung (2008): „IAB- Kurzbericht 6/2008", http://doku.iab.de/kurzgraf/2008/kbfolien06081.pdf (22.09.09).

Jowell, R. and the Central Coordinating Team (2007): „European Social Survey 2006/2007:Technical Report", Version 3.2, London: Centre for Comparative Social Surveys, City University. http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation (22.09.09).[76].

Langer, W. (1999a): „Einführung in die Grundlagen der explorativen Faktorenanalyse", http://www.soziologie.uni-halle.de/langer/lisrel/index.html (22.09.09).

Langer, W. (1999b): „Praktische Durchführung der explorativen Faktorenanalyse", http://www.soziologie.uni-halle.de/langer/lisrel/skripten/faktxeno.pdf (22.09.09).

Lynn, P. (2001): „Developing Quality Standards for Cross-National Survey Research: Five Approaches", *Working Papers of the Institute for Social and Economic Research, paper 2001-21*, Colchester: University of Essex auf http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2001-21.pdf (22.09.09).

[76] Bestehend aus dem Hauptdokument „ESS3-2006 Documentation Report" und 5 Anhängen die als seperate Dokumente unter diesem Pfad erhältlich sind: „ESS3 – Appendix A1, Population statistics", „ESS3 – Appendix A2, Classifications and coding standards", „ESS3 – Appendix A3, Variables and questions", „ESS3 – Appendix A4, Variable lists", „ESS3 – Appendix A5, Other country specific documentation".

Norwegian Social Science Data Service (NSD) (2006): „ESS3 2006 Data Protocol", Vers. 1.4, http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation (22.09.09).

O'Shea, R./Bryson, C./Jowell, R. (o.J.): „Comparative Attitudinal Research in Europe",http://www.europeansocialsurvey.org/index.php?option=com_docman&Itemid=&task=doc_download&gid=1 (22.09.09).

Philippens, M./Billiet, J. (o.J.): „Nonresponse in Cross-national Surveys: Results of the European Social Survey", http://www.ined.fr/fichier/t_rendezvous/126/rendezvous_fichier_philippens.pdf (22.09.09).

Smith, T.W. (2004): „Methods for Assessing and Calibrating Response Scales", http://www.srl.uic.edu/shethsudman/presentations/smith.PDF (22.09.09).

Smith, T.W. (2007): „Survey Non-Response Procedures in Cross-National Perspective: The ISSP Non-Response Survey", http://w4.ub.uni-konstanz.de/srm/article/viewFile/50/49 (22.09.09).

Symons, K/Matsuo, H./Beullens, K./Billiet J. (2008): „Response Based Quality Assessment in the ESS- Round 3", Leuven: Center of Sociological Research, http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation (22.09.09).

Temme, D./Hildebrandt, L. (2006): „Probleme der Validierung mit Strukturgleichungsmodellen", SFB Discussion Paper 82, Berlin: Humboldt-Universität, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2006-082.pdf (22.09.09).

Temme, D./Hildebrandt, L. (2008): „Gruppenvergleiche bei hypothetischen Konstrukten – Die Prüfung der Übereinstimmung von Messmodellen mit der Strukturgleichungsmethodik", SFB Discussion Paper 42, Berlin: Humboldt-Universität, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2008-042.pdf (22.09.09).

8.2. Übermittelte Quellen

Wiener Institut für Sozialwissenschaftliche Dokumentation und Methodik (=WISDOM): „österreichische Fragebogenversion der dritten Welle des Europen Social Surveys", übermittelt durch WISDOM.

9. Anhang

ANHANGSVERZEICHNIS

ad Kapitel 5.3.1 Häufigkeitsauszählungen

Tabelle A.1: Häufigkeitsauszählung „Allow immigrants of same race/ethnic group as majority"

in %	ESS	GB	Ost-D	West-D	Ö
Allow many	23,5	11,3	17,0	24,2	18,4
Allow some	44,2	48,1	45,9	48,2	49,3
Allow a few	24,3	31,3	26,5	22,0	27,3
Allow none	8,0	9,4	10,6	5,6	5,0
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Österreich haben zirka 49 % der befragten Personen die Einstellung,, dass es einigen Immigranten die derselben Volksgruppe oder ethnischen Gruppe wie die Mehrheit der Österreicher angehören, erlaubt sein sollte nach Österreich zu kommen, um hier zu leben.

Tabelle A.2: Häufigkeitsauszählung „Allow immigrants of different race/ethnic group as majority"

in %	ESS	GB	Ost-D	West-D	Ö
Allow many	13,0	8,0	7,8	10,7	8,7
Allow some	37,5	40,8	37,0	41,9	33,7
Allow a few	34,4	36,5	36,1	35,3	45,3
Allow none	15,1	14,7	19,2	12,1	12,4
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Österreich haben zirka 45 % der befragten Personen die Einstellung,, dass es wenigen Immigranten die einer anderen Volksgruppe oder ethnischen Gruppe wie die Mehrheit der Österreicher angehören, erlaubt sein sollte nach Österreich zu kommen, um hier zu leben.

Tabelle A.3: Häufigkeitsauszählung „Allow immigrants from poorer countries outside Europe"

in %	ESS	GB	Ost-D	West-D	Ö
Allow many	12,5	8,0	7,2	11,2	9,4
Allow some	34,9	36,8	30,4	37,0	35,7
Allow a few	34,8	37,8	38,3	37,0	42,7
Allow none	17,8	17,3	24,0	14,9	12,2
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Großbritannien haben zirka 45 % der befragten Personen die Einstellung,, dass es wenigen Immigranten die aus ärmeren Ländern außerhalb Europas stammen, erlaubt sein sollte nach Großbritannien zu kommen, um hier zu leben.

Tabelle A.4: Häufigkeitsauszählung „Immigration bad or good for country's economy"

in %	ESS	GB	Ost-D	West-D	Ö
Good for economy	3,3	2,2	4,2	2,0	4,4
1	3,0	2,3	1,3	2,6	3,8
2	9,7	8,7	7,6	8,7	10,0
3	12,6	10,6	8,2	11,6	13,3
4	11,0	10,4	8,6	12,7	11,1
5	23,5	20,8	26,0	24,6	25,4
6	9,2	11,3	8,8	11,2	8,9
7	9,8	10,9	10,1	10,5	8,4
8	7,5	9,5	9,5	7,3	5,8
9	4,4	5,0	6,1	3,5	4,2
Bad for economy	6,0	8,5	9,4	5,4	4,8
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Ostdeutschland sind zirka 9 % der befragten Personen der Meinung, dass es schlecht für die Wirtschaft Deutschlands ist, wenn Zuwanderer nach Deutschland zu kommen.

Tabelle A.5: Häufigkeitsauszählung „Country's cultural life undermined or enriched by immigrants"

in %	ESS	GB	Ost-D	West-D	Ö
Cult. Life enriched	5,6	3,4	6,9	5,3	4,7
1	4,9	3,2	3,4	4,6	3,2
2	13,4	9,4	13,3	14,5	8,2
3	14,5	12,4	14,5	14,5	10,7
4	10,8	10,5	10,3	11,9	8,2
5	20,3	17,0	24,4	22,7	24,3
6	7,7	12,2	7,2	8,5	9,5
7	8,4	11,6	5,4	8,1	10,9
8	6,0	8,8	6,2	4,4	8,0
9	3,6	4,6	3,6	1,9	6,3
Cult. life undermined	4,6	6,8	4,6	3,5	6,0
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Westdeutschland sind zirka 5 % der befragten Personen der Meinung, dass das kulturelle Leben Deutschlands durch Zuwanderer bereichert wird.

Tabelle A.6: Häufigkeitsauszählung „Country is made a worse or a better place by immigrants"

in %	ESS	GB	Ost-D	West-D	Ö
Better place to live	2,6	2,4	2,2	2,1	1,9
1	2,5	2,5	1,2	1,6	1,3
2	7,3	7,3	4,9	5,9	4,1
3	9,9	8,4	5,5	7,6	6,1
4	10,1	9,1	6,6	10,9	6,5
5	30,3	23,7	33,9	31,9	32,1
6	10,6	13,3	9,9	13,7	13,6
7	9,9	11,4	12,0	11,1	12,1
8	7,1	9,4	9,7	6,8	8,8
9	4,1	4,6	5,1	3,7	6,2
Worse place to live	5,5	7,9	8,8	4,6	7,2
Gesamt	100 %	100 %	100 %	100 %	100 %

Anmerkung: Gewichtung mit Designgewicht, Angabe der gültigen Antweilswerte.
Lesebeispiel: In Österreich sind zirka 32 % der befragten Personen der Meinung, dass Österreich durch die Zuwanderer weder ein besserer noch ein schlechterer Ort zum Leben wird.

ad Kapitel 5.3.2 Verteilungen der Items

Tabelle A.7: Schiefe der Items

	ESS	GB	Ost-D	West-D	Ö
allow immigrants of same race/ethnic group as majority	0,38	0,26	0,32	0,44	0,26
allow immigrants of different race/ethnic group as majority	0,04	0,09	-0,03	0,11	-0,16
allow immigrants from poorer countries outside Europe	-0,28	0,01	-0,20	-0,01	-0,10
immigration bad or good for country's economy	0,16	0,06	-0,06	0,18	0,20
country's cultural life undermined or enriched by immigrants	0,29	0,05	0,31	0,32	-0,02
country is made a worse or a better place by immigrants	0,11	-0,03	-0,05	0,04	-0,07

Anmerkung: Gewichtung mit Designgewicht.
Lesebeispiel: In Österreich hat die Verteilung des Items „allow immigrants of same race/ethnic group as majority" einen Schiefekoeffizienten von 0,26 und ist somit rechtsschief beziehungsweise linkssteil.

Tabelle A.8: Wölbung der Items

	ESS	GB	Ost-D	West-D	Ö
allow immigrants of same race/ethnic group as majority	-0,54	-0,38	-0,55	-0,30	-0,35
allow immigrants of different race/ethnic group as majority	-0,78	-0,63	-0,78	-0,58	-0,45
allow immigrants from poorer countries outside Europe	-0,85	-0,71	-0,79	-0,71	-0,51
immigration bad or good for country's economy	-0,51	-0,66	-0,55	-0,33	-0,37
country's cultural life undermined or enriched by immigrants	-0,53	-0,69	-0,38	-0,24	-0,59
country is made a worse or a better place by immigrants	-0,18	-0,45	-0,12	0,16	-0,01

Anmerkung: Gewichtung mit Designgewicht.
Lesebeispiel: In Österreich hat die Verteilungskurve des Items „allow immigrants of same race/ethnic group as majority" einen Wölbungskoeffizienten von -0,35 und ist somit flachgipfliger als im Fall einer Normalverteilung des Items.

ad Kapitel 5.3.4 Analyse der fehlenden Werte

Tabelle A.9: Korrelation der gesamten Anzahl der fehlenden Werte bezüglich der zu prüfenden Skalen und sozialstatistischen Merkmalen

	ESS	GB	Ost-D	West-D	Ö
Alter	-0,08**	-0,04*	0,06*	-0,03	-0,05**
Bildung	-0,1**	-0,07**	-0,03	-0,1**	-0,06**
Einkommen	-0,173**	-0,01	-0,03	-0,02	0
Geschlecht	0	0	0	0	0

Anmerkung: **= Korrelation ist auf dem 0,01 Niveau signifikant (einseitig), *= Korrelation ist auf dem 0,05 Niveau signifikant (einseitig); Korrelationen mit Geburtsjahr, Bildung und Einkommen mit Spearmans Rho, Assoziation mit Geschlecht mit asymmetrischen Lambda; für letzteres kann kein Signifikanzniveau bestimmt werden, da asymptotischer Standardfehler gleich Null ist; Gewichtung mit Designgewicht, paarweiser Fallausschluss.
Lesebeispiel: In West-Deutschland beträgt der Korrelationskoeffizient zwischen der Anzahl der fehlenden Werte und dem Merkmal höchste erreichte Bildung 0,1. Dieses Ergebnis ist hochsignifikant.

ad Kapitel 6 Korrelationsbasierte Verfahren zur Überprüfung der erreichten Ebene der Äquivalenz

Tabelle A.10: Interkorrelationen der Items im ESS-Gesamtdatensatz

Itemkorrelationen	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	1	0,68	0,58	0,39	0,37	0,39
allow immigrants of different race/ethnic group as majority	0,70	1	0,78	0,49	0,49	0,51
allow immigrants from poorer countries outside Europe	0,60	0,78	1	0,47	0,47	0,49
immigration bad or good for country's economy	0,41	0,50	0,49	1	0,59	0,63
country's cultural life undermined or enriched by immigrants	0,39	0,50	0,48	0,61	1	0,67
country is made a worse or a better place by immigrants	0,41	0,52	0,50	0,65	0,69	1

Anmerkung: unterhalb der Diagonale Korrelation nach Pearson, oberhalb Spearmans Rho (alle bivariaten Korrelationen hochsignifikant), listenweiser Fallausschluss, gewichtet mit Designgewicht.
Lesebeispiel: Das erste und das zweite Item weisen einen Korrelationskoeffizienten von 0,7 nach Pearson und ein Spearmans Rho von 0,68 auf.

Tabelle A.11: Interkorrelationen der Items im Datensatz Großbritanniens

Itemkorrelationen	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	1	0,77	0,65	0,45	0,44	0,49
allow immigrants of different race/ethnic group as majority	0,78	1	0,78	0,55	0,54	0,59
allow immigrants from poorer countries outside Europe	0,67	0,78	1	0,53	0,53	0,58
immigration bad or good for country's economy	0,47	0,56	0,54	1	0,65	0,72
country's cultural life undermined or enriched by immigrants	0,46	0,54	0,53	0,67	1	0,77
country is made a worse or a better place by immigrants	0,50	0,59	0,57	0,74	0,78	1

Anmerkung: unterhalb der Diagonale Korrelation nach Pearson, oberhalb Spearmans Rho (alle bivariaten Korrelationen hochsignifikant), listenweiser Fallausschluss, gewichtet mit Designgewicht.
Lesebeispiel: Das erste und das zweite Item weisen einen Korrelationskoeffizienten von 0,78 nach Pearson und ein Spearmans Rho von 0,77 auf.

Tabelle A.12: Interkorrelationen der Items im ostdeutschen Datensatz

Itemkorrelationen	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	1	0,72	0,60	0,51	0,47	0,52
allow immigrants of different race/ethnic group as majority	0,73	1	0,74	0,53	0,51	0,56
allow immigrants from poorer countries outside Europe	0,61	0,74	1	0,46	0,46	0,52
immigration bad or good for country's economy	0,52	0,55	0,47	1	0,58	0,65
country's cultural life undermined or enriched by immigrants	0,49	0,54	0,48	0,60	1	0,65
country is made a worse or a better place by immigrants	0,52	0,58	0,53	0,67	0,69	1

Anmerkung: unterhalb der Diagonale Korrelation nach Pearson, oberhalb Spearmans Rho (alle bivariaten Korrelationen hochsignifikant), listenweiser Fallausschluss, gewichtet mit Designgewicht.
Lesebeispiel: Das erste und das zweite Item weisen einen Korrelationskoeffizienten von 0,73 nach Pearson und ein Spearmans Rho von 0,72 auf.

Tabelle A.13: Interkorrelationen der Items im westdeutschen Datensatz

Itemkorrelationen	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	1	0,67	0,58	0,46	0,40	0,48
allow immigrants of different race/ethnic group as majority	0,69	1	0,77	0,50	0,46	0,56
allow immigrants from poorer countries outside Europe	0,59	0,78	1	0,46	0,44	0,54
immigration bad or good for country's economy	0,48	0,51	0,48	1	0,52	0,60
country's cultural life undermined or enriched by immigrants	0,42	0,49	0,47	0,55	1	0,64
country is made a worse or a better place by immigrants	0,48	0,57	0,55	0,61	0,67	1

Anmerkung: unterhalb der Diagonale Korrelation nach Pearson, oberhalb Spearmans Rho (alle bivariaten Korrelationen hochsignifikant), listenweiser Fallausschluss, gewichtet mit Designgewicht.
Lesebeispiel: Das erste und das zweite Item weisen einen Korrelationskoeffizienten von 0,68 nach Pearson und ein Spearmans Rho von 0,67 auf.

Tabelle A.14: Interkorrelationen der Items im österreichischen Datensatz

Itemkorrelationen	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	1	0,67	0,61	0,45	0,46	0,46
allow immigrants of different race/ethnic group as majority	0,70	1	0,78	0,51	0,57	0,58
allow immigrants from poorer countries outside Europe	0,63	0,80	1	0,47	0,53	0,53
immigration bad or good for country's economy	0,47	0,54	0,51	1	0,62	0,63
country's cultural life undermined or enriched by immigrants	0,49	0,60	0,56	0,64	1	0,70
country is made a worse or a better place by immigrants	0,48	0,60	0,56	0,65	0,71	1

Anmerkung: unterhalb der Diagonale Korrelation nach Pearson, oberhalb Spearmans Rho (alle bivariaten Korrelationen hochsignifikant), listenweiser Fallausschluss, gewichtet mit Designgewicht.
Lesebeispiel: Das erste und das zweite Item weisen einen Korrelationskoeffizienten von 0,7 nach Pearson und ein Spearmans Rho von 0,67 auf.

ad Kapitel 6.3 Überprüfung der funktionalen Äquivalenz mittels einer Hauptachsenanalyse mit Oblimin-Rotation

Tabelle A.15: Anti-Image Korrelationen der Items im Datensatz mit den 4 Analyseeinheiten

	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	0,89	-0,42	-0,12	-0,09	-0,04	-0,01
allow immigrants of different race/ethnic group as majority	-0,42	0,81	-0,53	-0,05	-0,06	-0,11
allow immigrants from poorer countries outside Europe	-0,12	-0,53	0,85	-0,06	-0,04	-0,09
immigration bad or good for country's economy	-0,09	-0,05	-0,06	0,91	-0,20	-0,33
country's cultural life undermined or enriched by immigrants	-0,04	-0,06	-0,04	-0,20	0,86	-0,47
country is made a worse or a better place by immigrants	-0,01	-0,11	-0,09	-0,33	-0,47	0,84

Anmerkung: listenweiser Fallausschluss, gewichtet mit Designgewicht; in der Diagonale befindet sich das Maß der Stichprobeneignung für das jeweilige Item.
Lesebeispiel: Das erste und das zweite Item weisen einen negativen partiellen Korrelationskoeffizienten von 0,42 auf.

Tabelle A.16: Anti-Image Korrelationen der Items im Datensatz Großbritanniens

	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	0,85	-0,53	-0,13	0,00	0,00	-0,04
allow immigrants of different race/ethnic group as majority	-0,53	0,81	-0,48	-0,10	-0,05	-0,05
allow immigrants from poorer countries outside Europe	-0,13	-0,48	0,88	-0,06	-0,06	-0,08
immigration bad or good for country's economy	0,00	-0,10	-0,06	0,90	-0,18	-0,41
country's cultural life undermined or enriched by immigrants	0,00	-0,05	-0,06	-0,18	0,86	-0,52
country is made a worse or a better place by immigrants	-0,04	-0,05	-0,08	-0,41	-0,52	0,82

Anmerkung: listenweiser Fallausschluss, gewichtet mit Designgewicht; in der Diagonale befindet sich das Maß der Stichprobeneignung für das jeweilige Item.
Lesebeispiel: Das erste und das zweite Item weisen einen negativen partiellen Korrelationskoeffizienten von 0,53 auf.

Tabelle A.17: Anti-Image Korrelationen der Items im ostdeutschen Datensatz

	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	0,88	-0,44	-0,12	-0,13	-0,05	-0,05
allow immigrants of different race/ethnic group as majority	-0,44	0,82	-0,49	-0,09	-0,09	-0,07
allow immigrants from poorer countries outside Europe	-0,12	-0,49	0,86	0,00	-0,03	-0,11
immigration bad or good for country's economy	-0,13	-0,09	0,00	0,89	-0,20	-0,35
country's cultural life undermined or enriched by immigrants	-0,05	-0,09	-0,03	-0,20	0,88	-0,41
country is made a worse or a better place by immigrants	-0,05	-0,07	-0,11	-0,35	-0,41	0,85

Anmerkung: listenweiser Fallausschluss, gewichtet mit Designgewicht; in der Diagonale befindet sich das Maß der Stichprobeneignung für das jeweilige Item.
Lesebeispiel: Das erste und das zweite Item weisen einen negativen partiellen Korrelationskoeffizienten von 0,44 auf.

Tabelle A.18: Anti-Image Korrelationen der Items im westdeutschen Datensatz

	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	0,89	-0,39	-0,09	-0,13	-0,03	-0,04
allow immigrants of different race/ethnic group as majority	-0,39	0,79	-0,57	-0,07	-0,04	-0,11
allow immigrants from poorer countries outside Europe	-0,09	-0,57	0,83	-0,05	-0,05	-0,11
immigration bad or good for country's economy	-0,13	-0,07	-0,05	0,91	-0,20	-0,28
country's cultural life undermined or enriched by immigrants	-0,03	-0,04	-0,05	-0,20	0,85	-0,44
country is made a worse or a better place by immigrants	-0,04	-0,11	-0,11	-0,28	-0,44	0,85

Anmerkung: listenweiser Fallausschluss, gewichtet mit Designgewicht; in der Diagonale befindet sich das Maß der Stichprobeneignung für das jeweilige Item.
Lesebeispiel: Das erste und das zweite Item weisen einen negativen partiellen Korrelationskoeffizienten von 0,39 auf.

Tabelle A.19: Anti-Image Korrelationen der Items im österreichischen Datensatz

	1	2	3	4	5	6
allow immigrants of same race/ethnic group as majority	0,91	-0,37	-0,14	-0,09	-0,04	-0,01
allow immigrants of different race/ethnic group as majority	-0,37	0,82	-0,55	-0,06	-0,11	-0,12
allow immigrants from poorer countries outside Europe	-0,14	-0,55	0,85	-0,04	-0,07	-0,08
immigration bad or good for country's economy	-0,09	-0,06	-0,04	0,91	-0,26	-0,30
country's cultural life undermined or enriched by immigrants	-0,04	-0,11	-0,07	-0,26	0,87	-0,42
country is made a worse or a better place by immigrants	-0,01	-0,12	-0,08	-0,30	-0,42	0,87

Anmerkung: listenweiser Fallausschluss, gewichtet mit Designgewicht; in der Diagonale befindet sich das Maß der Stichprobeneignung für das jeweilige Item.
Lesebeispiel: Das erste und das zweite Item weisen einen negativen partiellen Korrelationskoeffizienten von 0,37 auf.

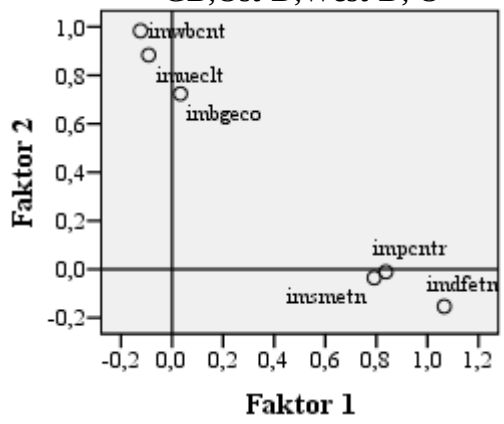
Tabelle A.20: Kommunalitäten der Komponenten nach Faktorenextraktion

	D,GB,Ö	GB	Ost-D	West-D	Ö
allow immigrants of same race/ethnic group as majority	0,58	0,67	0,61	0,53	0,56
allow immigrants of different race/ethnic group as majority	0,89	0,91	0,89	0,91	0,88
allow immigrants from poorer countries outside Europe	0,69	0,68	0,62	0,67	0,72
immigration bad or good for country's economy	0,56	0,64	0,59	0,52	0,58
country's cultural life undermined or enriched by immigrants	0,66	0,70	0,62	0,61	0,7
country is made a worse or a better place by immigrants	0,79	0,86	0,76	0,73	0,73

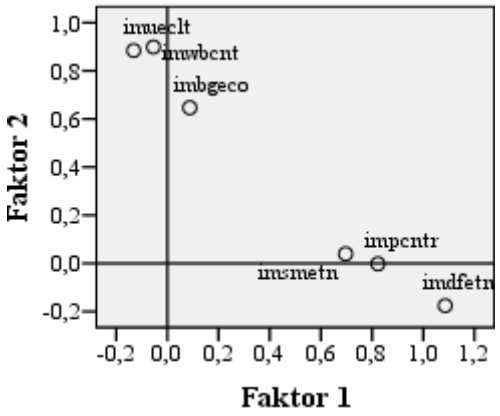
Anmerkung: Gewichtung mit Designgewicht; Extraktionsmethode: Hauptachsenanalyse, listenweiser Fallausschluss, es wurden Faktoren extrahiert, die einen Eigenwert ≥0,5 aufweisen.
Lesebeispiel: Das erste Item weist in Österreich eine Kommunalität von 0,5 auf, was bedeutet, dass etwas mehr als die Hälfte der Varianz dieses Items durch die zwei extrahierten Faktoren erklärt wird.

Abbildungen A.1-5: Faktordiagramme im gedrehten Faktorenbereich (Delta =0,2)

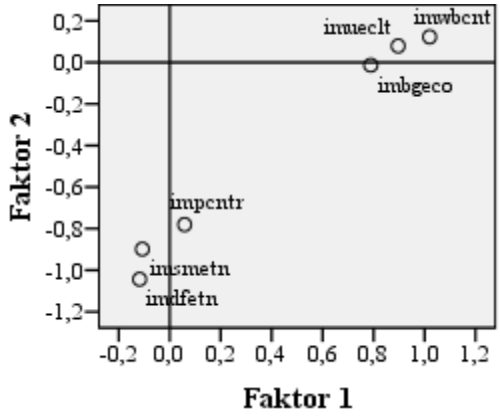
GB,Ost-D,West-D, Ö



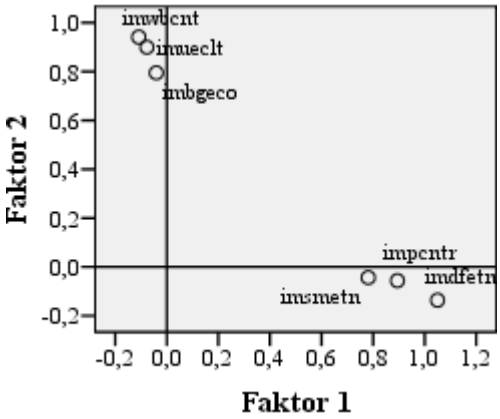
Westdeutschland



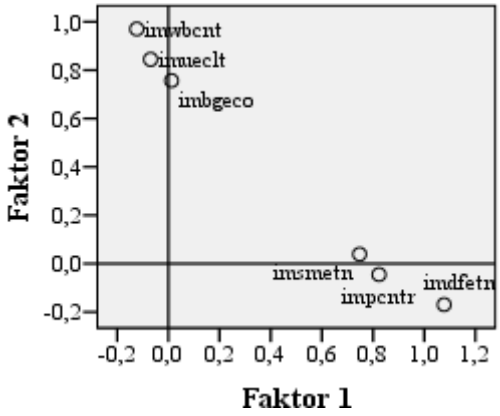
Großbritannien



Österreich



Ostdeutschland

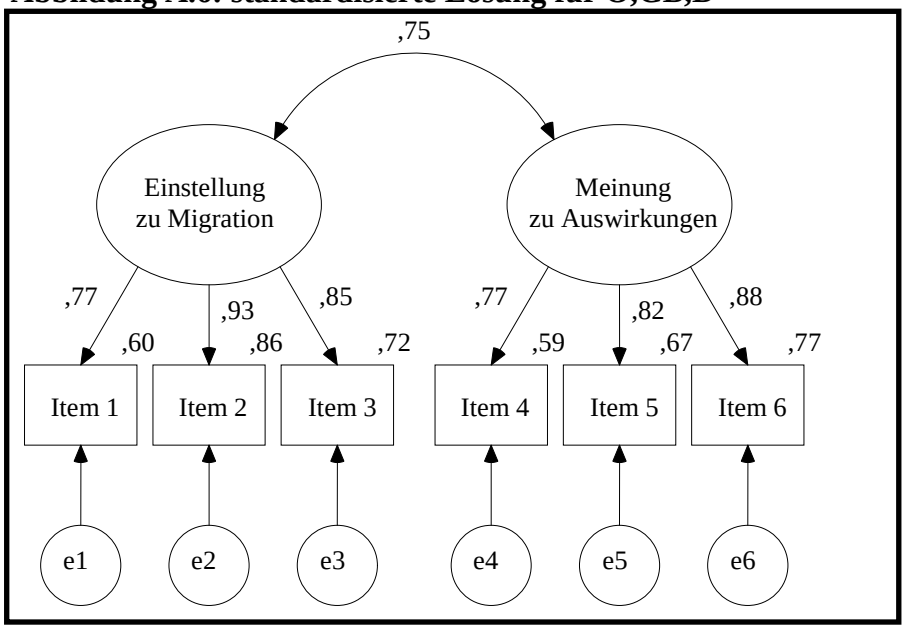


Legende:

- Item 1: imsmetn
- Item 2: imdfetn
- Item 3: impcntr
- Item 4: imbgeco
- Item 5: imueclt
- Item 6: imwbcnt

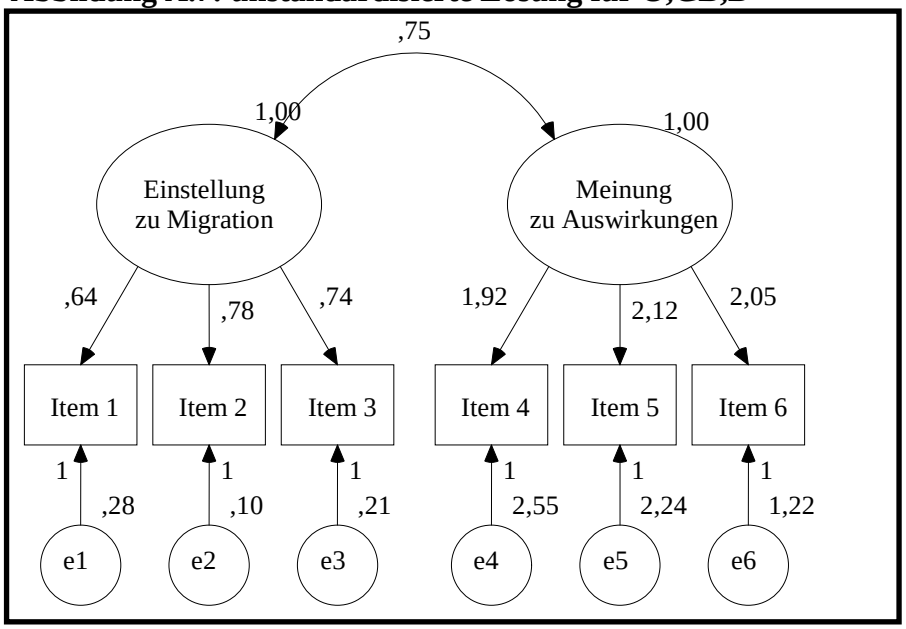
ad Kapitel 6.4 Überprüfung des Ausmaßes der funktionalen Äquivalenz mittels einer konfirmatorischen Faktorenanalyse

Abbildung A.6: standardisierte Lösung für Ö,GB,D



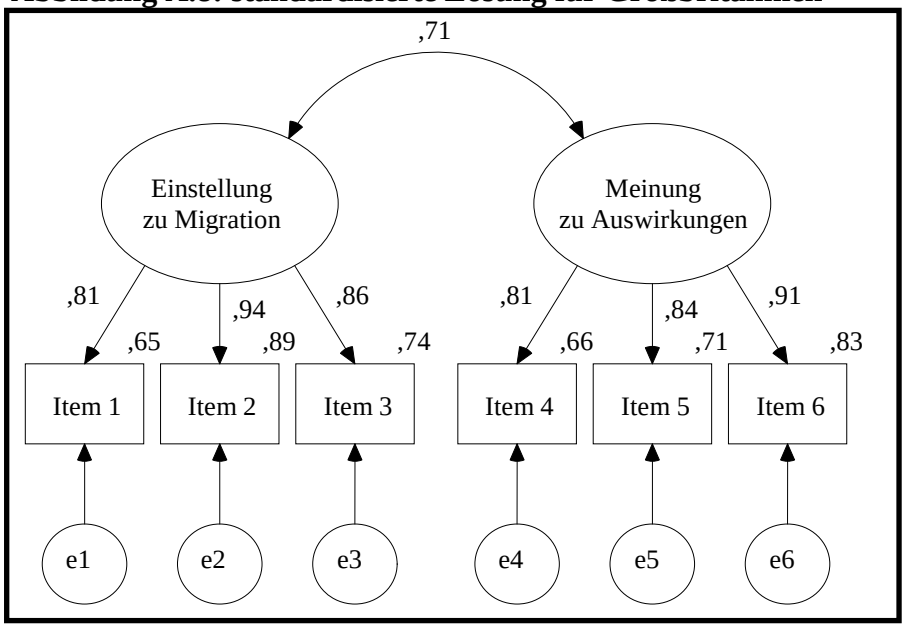
Lesebeispiel: Das Item 2 weist eine Faktorladung von 0,93 auf den Faktor Einstellung zu Migration auf.

Abbildung A.7: unstandardisierte Lösung für Ö,GB,D



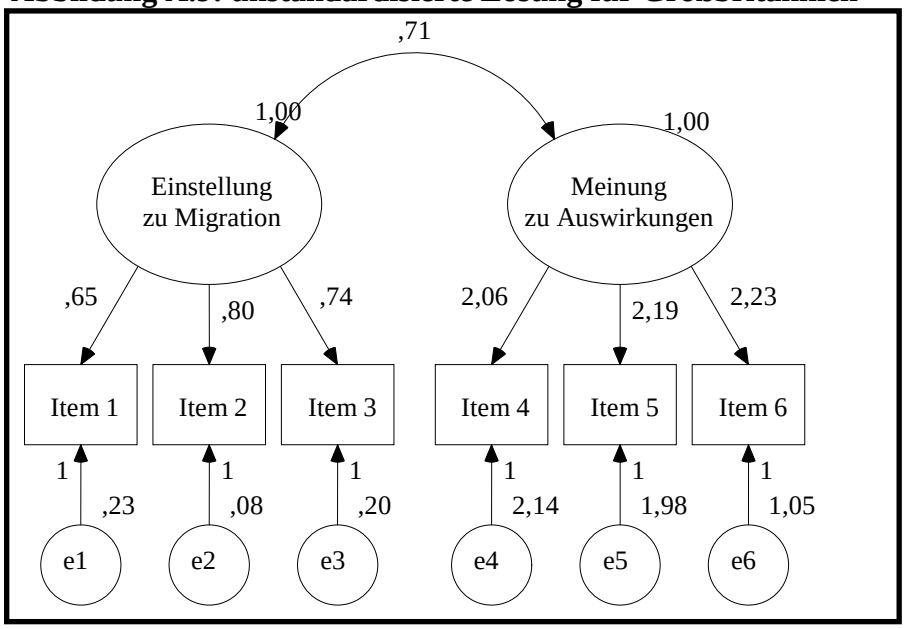
Lesebeispiel: Das Item 2 weist einen Regressionskoeffizienten von 0,78 auf den Faktor Einstellung zu Migration auf.

Abbildung A.8: standardisierte Lösung für Großbritannien



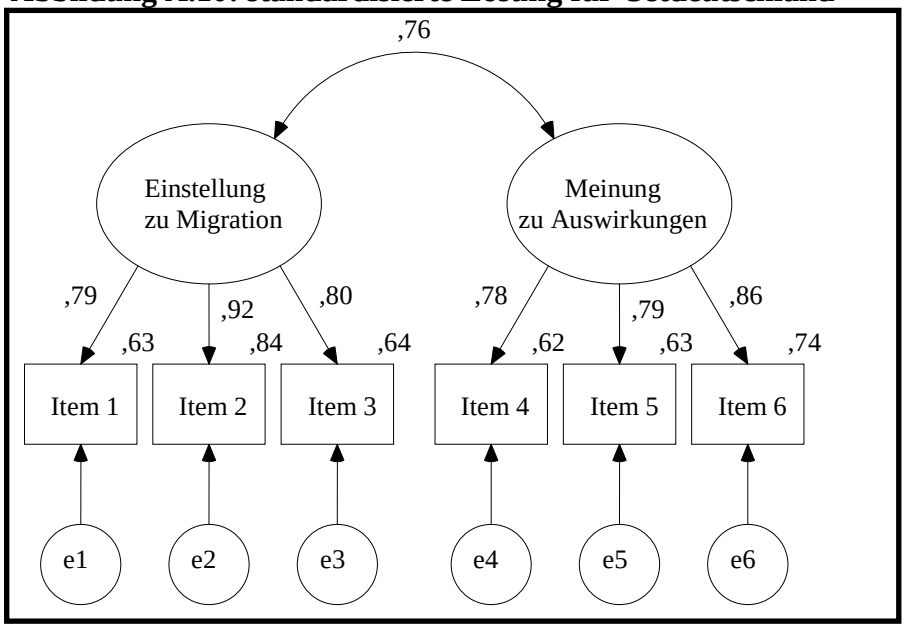
Lesebeispiel: Das Item 2 weist eine Faktorladung von 0,94 auf den Faktor Einstellung zu Migration auf.

Abbildung A.9: unstandardisierte Lösung für Großbritannien



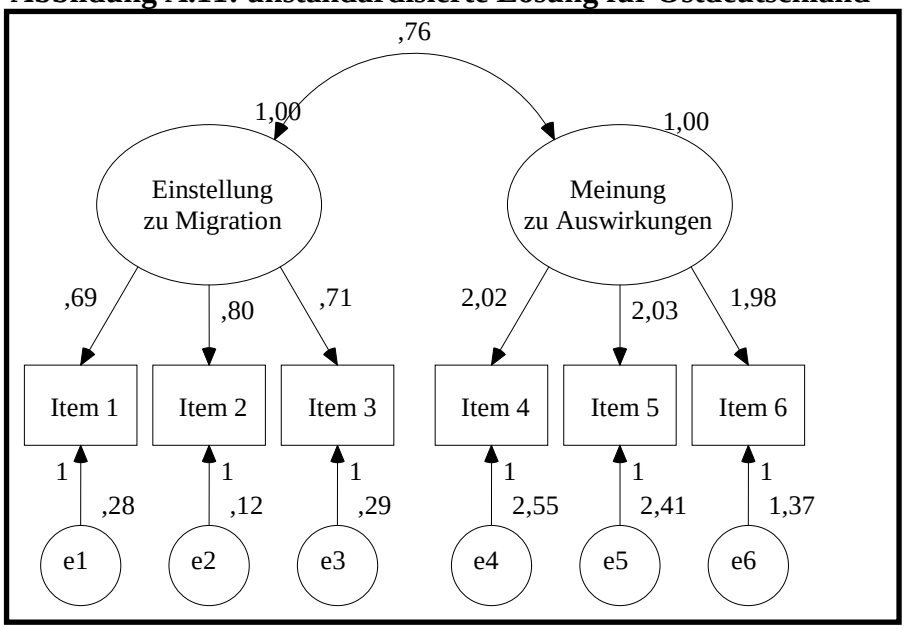
Lesebeispiel: Das Item 2 weist einen Regressionskoeffizienten von 0,8 auf den Faktor Einstellung zu Migration auf.

Abbildung A.10: standardisierte Lösung für Ostdeutschland



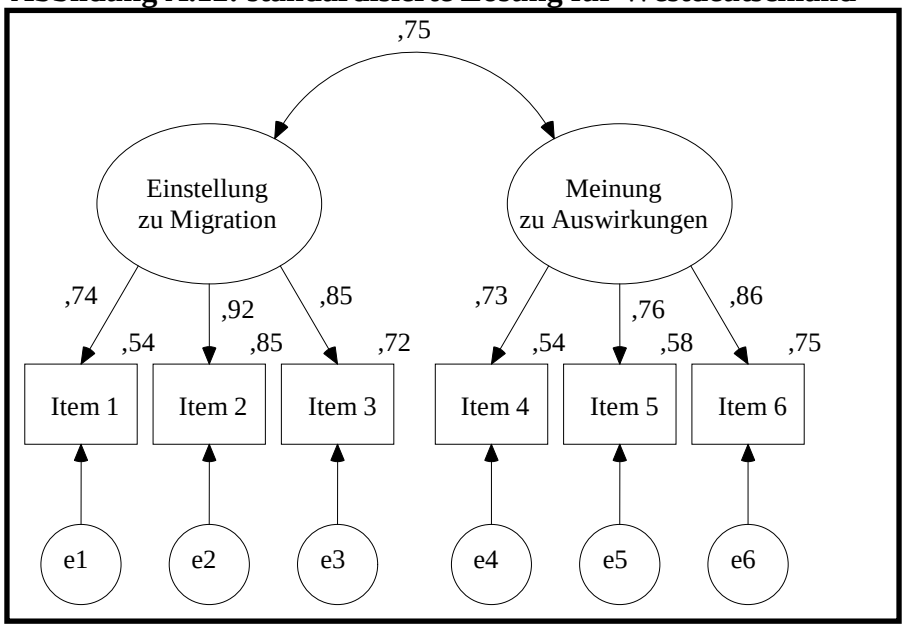
Lesebeispiel: Das Item 2 weist eine Faktorladung von 0,92 auf den Faktor Einstellung zu Migration auf.

Abbildung A.11: unstandardisierte Lösung für Ostdeutschland



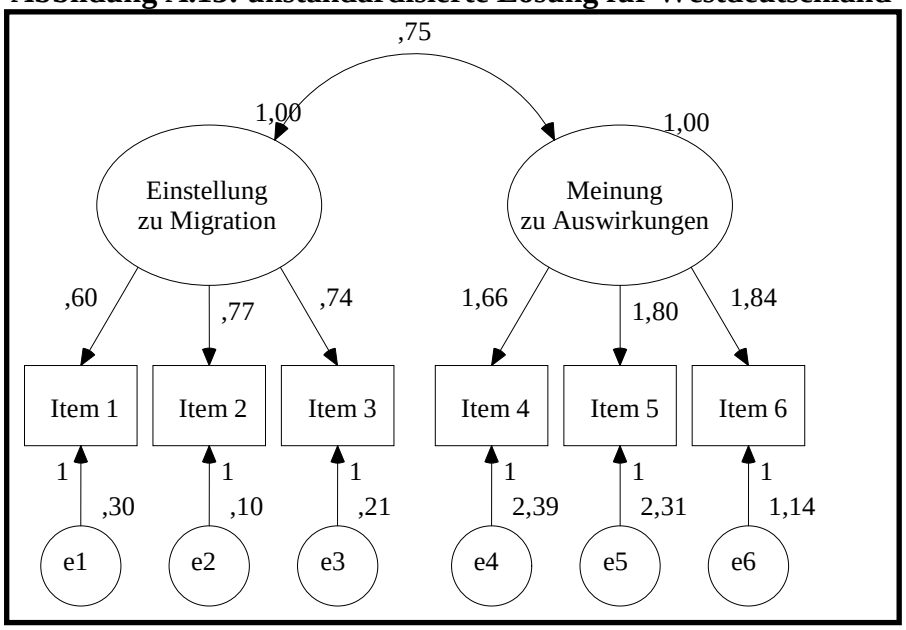
Lesebeispiel: Das Item 2 weist einen Regressionskoeffizienten von 0,8 auf den Faktor Einstellung zu Migration auf.

Abbildung A.12: standardisierte Lösung für Westdeutschland



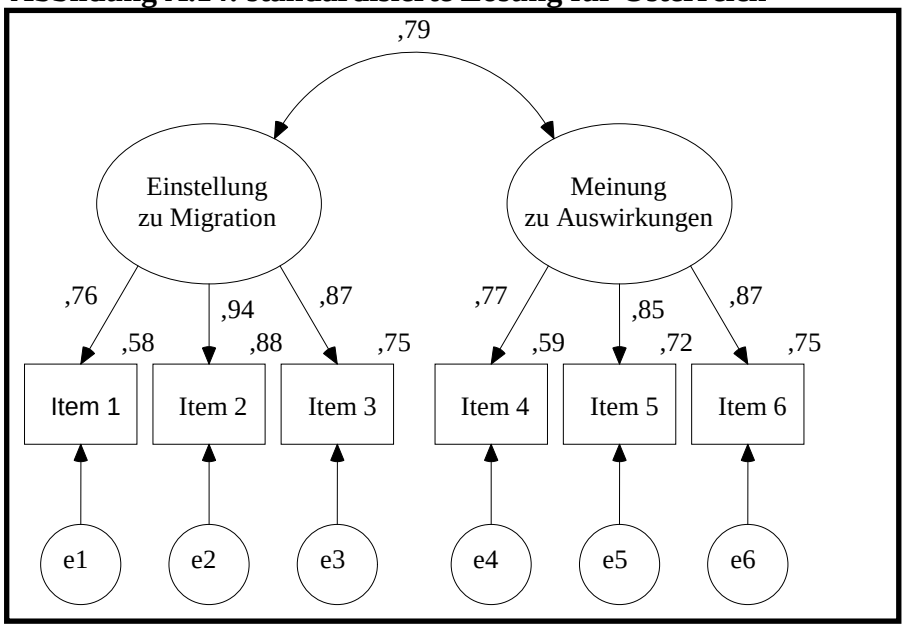
Lesebeispiel: Das Item 2 weist eine Faktorladung von 0,92 auf den Faktor Einstellung zu Migration auf.

Abbildung A.13: unstandardisierte Lösung für Westdeutschland



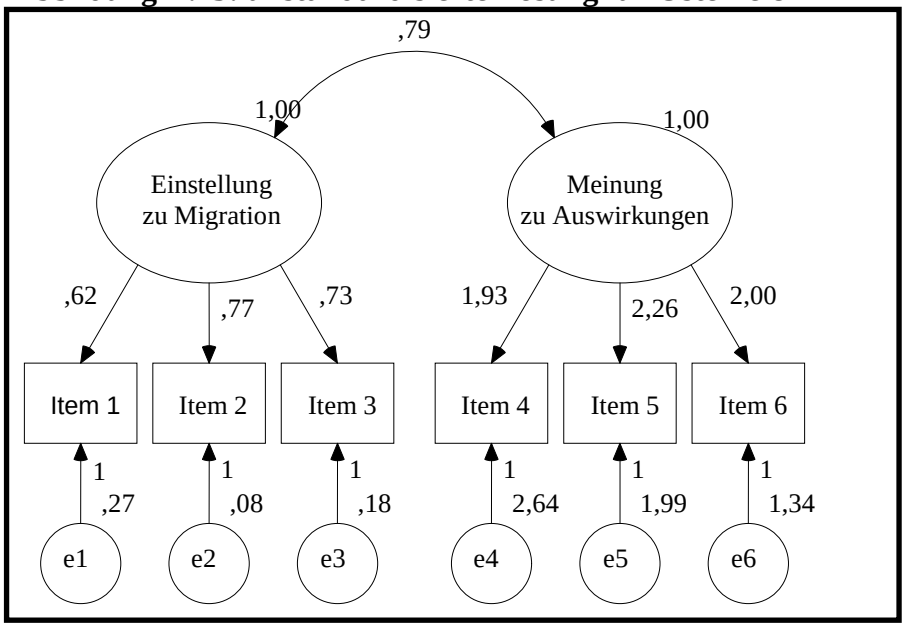
Lesebeispiel: Das Item 2 weist einen Regressionskoeffizienten von 0,77 auf den Faktor Einstellung zu Migration auf.

Abbildung A.14: standardisierte Lösung für Österreich



Lesebeispiel: Das Item 2 weist eine Faktorladung von 0,94 auf den Faktor Einstellung zu Migration auf.

Abbildung A.15: unstandardisierte Lösung für Österreich



Lesebeispiel: Das Item 2 weist einen Regressionskoeffizienten von 0,77 auf den Faktor Einstellung zu Migration auf.

Multipler Gruppenvergleich

Tabelle A.21: Gütemaße für den multiplen Gruppenvergleich zwischen Österreich und Ostdeutschland

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	2,150	,995	,987	,991	,020	-	-	-
Faktorladungen invariant	2,751	,992	,983	,984	,024	20,625	4	,000

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade und das mit den invarianten Faktorladungen 20. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 20,6; die bei einer Differenz von 4 Freiheitsgraden hochsignifikant ist.

Tabelle A.22: Gütemaße für den multiplen Gruppenvergleich zwischen Österreich und Westdeutschland

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,664	,992	,980	,983	,027	-	-	-
Faktorladungen invariant	3,594	,991	,980	,980	,026	13,262	4	,010

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade und das mit den invarianten Faktorladungen 20. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 13,3; die bei einer Differenz von 4 Freiheitsgraden signifikant ist.

Tabelle A.23: Gütemaße für den multiplen Gruppenvergleich zwischen Österreich und Großbritannien

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,896	,993	,982	,987	,026	-	-	-
Faktorladungen invariant	4,308	,991	,981	,981	,028	23,815	4	,000

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade und das mit den invarianten Faktorladungen 20. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 23,8; die bei einer Differenz von 4 Freiheitsgraden hochsignifikant ist.

Tabelle A.24: Gütemaße für den multiplen Gruppenvergleich zwischen Großbritannien und Ostdeutschland

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	4,212	,991	,975	,981	,032	-	-	-
Faktorladungen invariant	3,893	,989	,977	,978	,030	10,465	4	,033

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade und das mit den invarianten Faktorladungen 20. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 10,5; die bei einer Differenz von 4 Freiheitsgraden signifikant ist.

Tabelle A.25: Gütemaße für den multiplen Gruppenvergleich zwischen Westdeutschland und Ostdeutschland

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,979	,988	,970	,972	,034	-	-	-
Faktorladungen invariant	4,028	,985	,969	,965	,034	16,899	4	,002

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade und das mit den invarianten Faktorladungen 20. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 16,9; die bei einer Differenz von 4 Freiheitsgraden signifikant ist.

Tabelle A.26: Gütemaße für den multiplen Gruppenvergleich zwischen Großbritannien und Westdeutschland

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	5,726	,989	,970	,975	,035	-	-	-
Faktorladungen invariant	4,811	,988	,975	,975	,031	4,592	4	,332
zus. Kovarianz der latenten Var. invariant	7,319	,979	,961	,953	,040	76,725	7	,000

Anmerkung: Das erste Modell enthält 16 Freiheitsgrade, das mit den invarianten Faktorladungen 20 und das Modell bei dem zusätzlich die Kovarianz der latenten Variablen restringiert wurde 23. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant. Der Chi-Quadrat-Differenzentest bezieht sich jeweils auf das vorhergehende Modell.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 4,6; die bei einer Differenz von 4 Freiheitsgraden nicht signifikant ist.

Multipler Gruppenvergleich (partielle Invarianz der Faktorladungen)

Tabelle A.27: Gütemaße multipler Gruppenvergleich alle Analyseeinheiten (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,938	,991	,978	,982	,021	-	-	-
Faktorladungen part. invariant	3,671	,991	,979	,981	,020	13,498	6	,036

Anmerkung: Es wurden zwischen den drei Gruppen alle Regressionskoeffizienten bis auf die des ersten und fünften Items fixiert. Das erste Modell enthält 32 Freiheitsgrade und das mit den invarianten Faktorladungen 38. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 13,5; die bei einer Differenz von 6 Freiheitsgraden signifikant ist.

Abbildung A.28: Gütemaße für den multiplen Gruppenvergleich zwischen allen Analyseeinheiten (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,938	,991	,978	,982	,021	-	-	-
Faktorladungen invariant	3,735	,990	,979	,980	,020	15,907	6	,014

Anmerkung: Es wurden lediglich die Regressionskoeffizienten der Items des ersten Faktors zwischen den Gruppen fixiert. Das erste Modell enthält 32 Freiheitsgrade und das mit den invarianten Faktorladungen 38. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 15,9; die bei einer Differenz von 6 Freiheitsgraden signifikant ist.

Abbildung A.29: Gütemaße für den multiplen Gruppenvergleich zwischen allen Analyseeinheiten (partielle Invarianz)

	χ^2/df	GFI	AGFI	CFI	RMSEA	χ^2Diff	dfDiff	Sig.
Modell ohne Restriktionen	3,938	,991	,978	,982	,021	-	-	-
Faktorladungen invariant	4,125	,989	,977	,977	,021	30,757	6	,000

Anmerkung: Es wurden lediglich die Regressionskoeffizienten der Items des zweiten Faktors zwischen den Gruppen fixiert. Das erste Modell enthält 32 Freiheitsgrade und das mit den invarianten Faktorladungen 38. Der Test der Nullhypothese, dass der RMSEA-Wert tatsächlich kleiner als 0,05 ist, ist insignifikant.
Lesebeispiel: Der Chi-Quadrat-Differenzentest ergibt eine Chi-Quadrat-Differenz von 30,757; die bei einer Differenz von 6 Freiheitsgraden hochsignifikant ist.

Lebenslauf

Persönliche Daten

Name	Martin VOGLER
Geburtsdaten	14. Jänner 1982
Geburtsort	Wien
Staatsangehörigkeit	Österreich

Schulbildung/Hochschulbildung

1988 – 1992	Volksschule in Wien
1992 – 2001	Bundesrealgymnasium in Wien
WS 2002	Diplomstudium Betriebswirtschaftslehre an der Wirtschaftsuniversität Wien,
SS 2003 – WS 2009	Diplomstudium Soziologie der Rechts-, Sozial und Wirtschaftswissenschaften an der Fakultät für Human- und Sozialwissenschaften an der Universität Wien

Zum Studium passende Erfahrung

2009	Durchführung einer sekundären Marktanalyse zum Thema „Vereinbarkeit von Familie und Beruf“ im Auftrag von Accor Services Austria GmbH