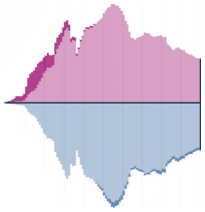


Professur für Bevölkerungswissenschaften
Chair of Population Studies

Discussion Papers

No. 2/2011

8

Henriette Engelhardt & Stefanie Roppelt

Zur Validität des Vergleichs subjektiver Daten: Eine Vignetten-Hopit-Analyse selbstberichteter Mobilität mit SHARE

Otto-Friedrich-Universität Bamberg
University of Bamberg

Zur Validität des Vergleichs subjektiver Daten:

Eine Vignetten-Hopit-Analyse selbstberichteter Mobilität mit SHARE[1]

Henriette Engelhardt und Stefanie Roppelt

Zusammenfassung

Bei einem Vergleich subjektiver Daten besteht immer die Möglichkeit, dass Antwortkategorien unterschiedlich ausgelegt werden, so dass eine direkte Gegenüberstellung ad hoc nicht möglich ist. Im vorliegenden Beitrag wird ein innovatives Verfahren zum Vergleich subjektiver Daten diskutiert, in welchem Vignetten und ein hierachical ordered probit (Hopit) Modell zur Schätzung der verwendeten Skalierungen verwendet werden. Am Beispiel der im Survey of Health Ageing and Retirement in Europe (SHARE) erhobenen selbstberichteten Mobilitätseinschränkungen in Verbindung mit entsprechenden Vignetten-Daten zeigen sich große Unterschiede in der berichteten Mobilität, besonders zwischen den Befragten in Polen und Tschechien. Unter Berücksichtigung tschechischer oder polnischer Schwellenwerte ändert sich im internationalen Vergleich der Anteil Personen, die angeben starke/extreme Schwierigkeiten bei der Bewegung zu haben um bis zu 12 Prozentpunkte. Ein Vergleich der subjektiven Antworten gibt damit nicht nur Unterschiede in objektiven Gesundheitszuständen wieder, sondern auch Differenzen in den Antwortskalierungen zwischen den Ländern. Daher ist es unabdingbar notwendig, die Antwortskalierungen beim Vergleich subjektiver Daten zu berücksichtigen.

[1] Der vorliegende Aufsatz basiert auf der Diplomarbeit von Stefanie Roppelt (2010) an der Professur für Bevölkerungswissenschaft der Universität Bamberg. Wir danken Christopher Schmidt für hilfreiche Kommentare.

Zur Validität des Vergleichs subjektiver Daten:

Eine Vignetten-Hopit-Analyse selbstberichteter Mobilität mit SHARE

„Die vergleichende Soziologie ist nicht etwa nur ein besonderer Zweig der Soziologie; sie ist soweit die Soziologie selbst.“ (Durkheim 1991: 216)

1 Einleitung

Die Methode des Vergleichs gehört seit ihren Anfängen zum Wesen der Sozialwissenschaften und insbesondere der Soziologie. Nur durch den Vergleich einzelner Individuen oder sozialer Gruppen können soziologische Theorien entwickelt, getestet und verbessert werden. Eine besondere Bedeutung kommt hierbei dem internationalen Vergleich zu, der im letzten Jahrzehnt nicht zuletzt durch die zunehmende Verfügbarkeit entsprechender Daten an Relevanz gewonnen hat (Kohn 1987; Gauthier 2002). Das Gros der erhobenen Daten besteht dabei aus subjektiven Einschätzungen der Befragten zu einem bestimmten Themenbereich. Ein Vergleich dieser subjektiven Daten kann mitunter zu wunderlichen und widersprüchlichen Ergebnisse führen (Murray et al. 2002; Salomon et al. 2004). Bei einer kategorialen Erhebung besteht denn auch immer die Möglichkeit, dass die Antwortkategorien von den Befragten unterschiedlich ausgelegt werden, so dass ein direkter Vergleich ad hoc nicht möglich ist.

Um diese Daten dennoch vergleichen zu können, werden in der Literatur verschiedene Verfahren vorgeschlagen. Eine neue von Gary King et al. (2004) entwickelte Methode besteht in der Verwendung sogenannter Vignetten zur Schätzung der seitens des Befragten verwendeten Skala anhand eines hierachical ordered probit (Hopit) Modells. Dieses Verfahren findet zunehmend Verwendung u.a. in Studien zum Vergleich von Gesundheit (Angelini et al. 2009; Bagodu et al. 2008; Gupta et al. 2009; Kapteyn et al. 2009; Salomon et al. 2004; Sirven et al. 2008; Soest et al. 2007). Ziel der vorliegenden Arbeit ist es zu zeigen, warum Vergleiche subjektiver Daten problematisch sind und welche Beiträge Vignette-Daten und das Hopit-Modell leisten können. Dies geschieht am Beispiel der selbstberichteten Mobilität, welche u.W. bislang nicht auf unterschiedliches Antwortverhalten im internationalen Vergleich untersucht worden ist.

Der Gang der Untersuchung gestaltet sich dabei wie folgt: Im nächsten Abschnitt werden die Problematik des Vergleichs subjektive Daten näher erläutert und verschiedene Lö-

sungsansätze präsentiert. Im dritten Abschnitt wird das Hopit-Modell zur Schätzung länderspezifischer Schwellenwerte mithilfe von Vignetten vorgestellt. Im vierten Abschnitt folgen die Präsentation der Daten und Variablen sowie ein deskriptiver Überblick über die Verteilung der selbstberichteten Mobilität und der Vignette-Daten. Im fünften Abschnitt werden die Ergebnisse des Hopit-Modells präsentiert und die Ergebnisse von Simulationen vorgestellt, die aufzeigen, zu welchen unterschiedlichen Ergebnissen man, je nach verwendeten Schwellenwerten gelangen kann. Abschließend werden die Ergebnisse noch einmal zusammenfassend dargestellt und diskutiert.

2 Vergleichbarkeit subjektiver Daten

Bestimmte Bereiche in Umfragen mit selbstberichteten Daten, wie Einstellungen, Werte und Normen sind per se subjektiv, so dass die Zielperson nur selbst Auskunft darüber geben kann. Hierzu werden die Befragten gebeten, sich selbst auf einer bestimmten Skala einzuordnen (Soest et al. 2007). Aber auch in anderen Bereichen, wie im Falle der Gesundheit ist eine subjektive Erhebung üblich, da diese im Gegensatz zur Erhebung einzelner diagnostizierter Erkrankungen (oder gar der Durchführung bestimmter medizinischer Tests) alle Aspekte des multidimensionalen Konzepts Gesundheit berücksichtigt. Die subjektive Beurteilung gibt zudem Aufschluss, inwiefern der Befragte durch gesundheitliche Probleme oder Einschränkungen belastet ist. Dies ist durch einen außenstehenden Beobachter in dieser Form nicht möglich (Sen 2002). Darüber hinaus sind selbstberichtete Daten in der Regel schnell und vergleichsweise kostengünstig zu erheben. Daher ist die Erhebung von subjektiven Daten unabdingbar.

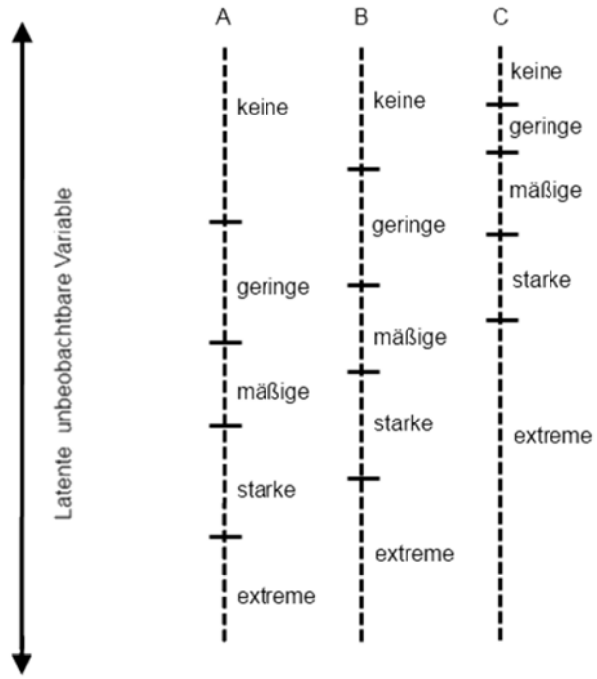
Ungeachtet der Vorteile hat das Arbeiten mit selbstberichteten Daten auch Nachteile. Selbstberichtete Daten sind zumeist subjektiv, d.h. die Fragen und Antworten werden individuell unterschiedlich interpretiert. Verschiedene Studien haben denn auch gezeigt, dass ein Vergleich ohne Bereinigung der Daten zu widersprüchlichen Ergebnissen führt (Salomon et al. 2004). Dies ist insbesondere bei international vergleichenden Studien der Fall, aber auch bei Vergleichen über soziale Gruppen oder in Längsschnittuntersuchungen (Jürges 2008).

Technisch gesprochen besteht das Problem subjektiver Daten in einer ex ante vorgenommenen Kategorisierung und dem Fehlen objektiver Maßeinheiten. Die Kategorien werden von jedem Individuum unterschiedlich interpretiert und ausgelegt (Murray et al. 2002). Bei der Erhebung kategorialer Daten wird davon ausgegangen, dass es eine latente, nicht beobachtbare Variable gibt. Deren Ausprägung wird nun vom Befragten mit den vorgegebe-

nen Antwortkategorien verglichen, und der Wert einer bestimmten Kategorie zugeordnet. Zur Auswahl dieser Kategorie verwendet der Befragte sogenannte Schwellenwerte – Werte, die definieren, welcher Kategorie sich eine Person zuordnet.

Die drei Personen (A, B, C) in Abbildung 1 haben jeweils unterschiedliche Schwellenwerte, was dazu führt, dass sich die Antworten trotz identischem Wert der latenten Variable unterscheiden. Bei systematischen individuellen Unterschieden in den Schwellenwerten sind Vergleiche der Antworten nicht mehr ohne weiteres möglich. Die Messungen spiegeln dann nicht mehr identische Werte der latenten Variablen wieder (Tandon et al. 2002). Gary King et al. (2004) bezeichnen dies in Anlehnung an die psychologische Testtheorie als *differential item functioning* (DIF).

Abbildung 1: Latente Variable und subjektive Skalierung der Antwortkategorien

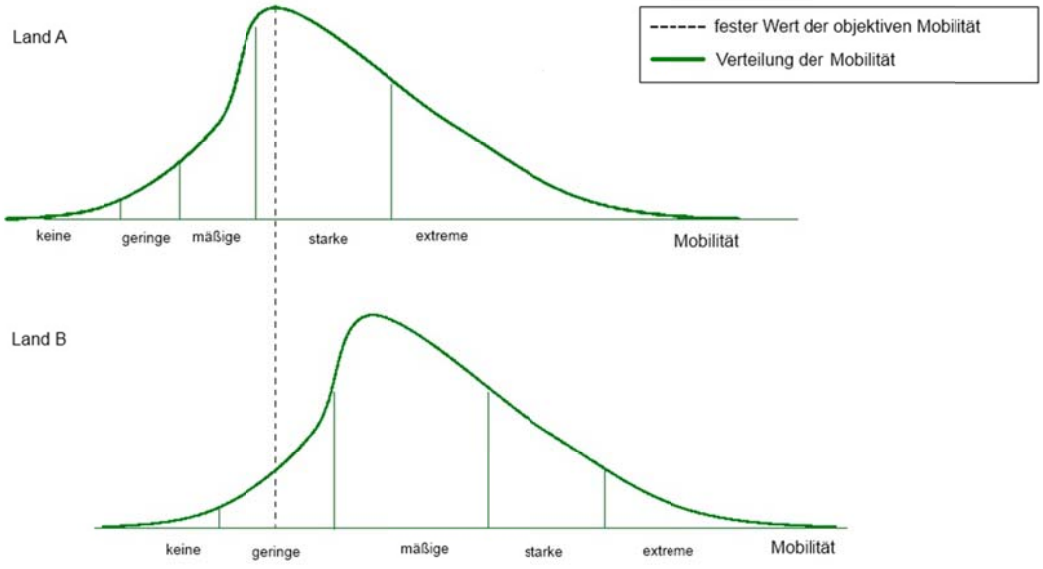


Quelle: In Anlehnung an Murray et al. (2002: 423)

Werden nun Gruppen miteinander verglichen, deren Mitglieder sich systematisch hinsichtlich der individuellen Skalierung unterscheiden, hat dies eine systematische Verzerrung der

resultierenden Verteilungen auf der Aggregatebene zur Folge. In Abbildung 2 sind die grundlegende Problematik und die daraus resultierenden Fehler am Beispiel der selbstberichteten Mobilität in zwei Ländern schematisch dargestellt. Die Verteilung in Land A ist im Vergleich zu Land B etwas nach links verschoben. Die fünf-stufigen Skalen beschreiben, inwieweit die betreffenden Personen Schwierigkeiten in ihrer Mobilität angegeben haben.

Abbildung 2: Unterschiede in den Schwellenwerten am Beispiel der Verteilung von Mobilität in zwei Populationen



Quelle: In Anlehnung an Kapteyn (2010: 35)

Die Schwellenwerte unterscheiden sich in den beiden Ländern deutlich, da in Land B die Abstände zwischen den Stufen gleichmäßiger verteilt sind als in Land A. Obwohl die Personen in Land A im Schnitt weniger Schwierigkeiten mit Mobilität haben, ist ihre Sichtweise negativer als die der Personen aus Land B. Mit der gestrichelten Linie wird ein bestimmter Wert der objektiven Mobilität angegeben. In Land A würde die Person bei diesem Wert antworten, dass sie *starke* Schwierigkeiten hat. In Land B würde eine Person trotz gleichen objektiven Einschränkungen subjektiv antworten, dass sie *geringe* Schwierigkeiten hat. Auf Basis der selbstberichteten Mobilität müsste man dann davon ausgehen, dass

die Personen in Land A mehr Schwierigkeiten haben als in Land B, was das genaue Gegenteil der wahren Mobilität wäre (Kapteyn 2010).

Die Ursachen für die systematische Verzerrung der Schwellenwerte sind unterschiedlich. So ist die individuelle Wahrnehmung der latenten Variablen davon beeinflusst, welche Erfahrungen eine Person bislang gemacht hat, in welchem Umfeld sie lebt und in welcher Gesellschaft sie groß geworden ist (Sen 2002). Weiterhin kann es Unterschiede in der Definition der abgefragten Aspekte zwischen einzelnen Ländern geben und Unterschiede in den Politiken, die diese Definition beeinflussen (Kapteyn 2010). So weist ein Vergleich der selbstberichteten Arbeitsunfähigkeit in den USA und Europa auf Unterschiede im Verständnis von Arbeitsunfähigkeit hin (Kapteyn et al. 2009), welche auf von voneinander abweichenden rechtlichen Definitionen von Arbeitsunfähigkeit beruhen. Die Befragten übernehmen diese Definition, da sie durch die Arbeitsunfähigkeitsversicherung direkt davon betroffen sind (Kapteyn et al. 2009). Daraus ergeben sich Unterschiede der Schwellenwerte, die von den Befragten verwendeten werden.

Nicht zuletzt kann der unterschiedliche Sprachgebrauch bei der Benennung der Kategorie relevant sein. Jürges (2008) weist darauf hin, dass im Fall der Frage nach der Gesundheit die Antwort „excellent“ eine durchaus übliche Antwort im angelsächsischen Raum ist. In Deutschland jedoch wird die Antwort „ausgezeichnet“ als Übertreibung empfunden und ist unüblich. Die unterschiedlichen Ursachen der Verzerrung subjektiver Daten zeigen, dass es notwendig ist, diese zu untersuchen, bevor international vergleichend gearbeitet werden kann. Denn ohne eine Bereinigung der Daten ist nicht zu unterscheiden, welcher Teil der Antworten objektive Unterschiede widerspiegelt (Item Bias) und welcher Teil Unterschiede im Antwortverhalten (Item Impact) (Vonkova 2010). Dies führt dazu, dass die Ergebnisse internationaler Vergleiche subjektiver Daten verzerrt sind und zu falschen Schlüssen führen können.

Auch im Falle der selbstberichteten Mobilität besteht die Gefahr der Nichtvergleichbarkeit der Daten, sowohl im Vergleich über die Zeit, über soziale Gruppen hinweg als auch im regionalen Vergleich. Die Ursachen für die unterschiedliche Skalierung beim internationalen Vergleich kann – neben der unterschiedlichen Semantik – in den Erfahrungen der Individuen mit den verschiedenen Gesundheitssystemen liegen. Die europäischen Länder unterscheiden sich hinsichtlich der Organisation und Ausgaben ihrer Gesundheitssysteme erheblich (Wendt 2003). Daher muss davon ausgegangen werden, dass auch die Erfahrungen der Individuen in Europa unterschiedlich sind. Um Vergleichbarkeit zu schaffen ist eine Korrektur der selbstberichteten Mobilität im Vorfeld notwendig.

Skalierung der latenten Variablen

Unterschiede in der Skalierung der latenten Variablen können u.a. durch die Bildung homogener Gruppen sowie anhand Vignette-Fragen aufgedeckt werden. Gemeinsam ist beiden Verfahren, dass der Wert der latenten Variablen fixiert wird. So werden beispielsweise Gruppen mit dem gleichen latenten Gesundheitszustand miteinander verglichen. Antworten diese in verschiedenen Kategorien, können so Unterschiede in den Schwellenwerten zwischen Individuen oder Populationen identifiziert werden. Die so festgestellten Unterschiede in den Antwortkategorien zwischen Individuen oder Populationen werden dann genutzt, um die Schwellenwerte zu identifizieren (Murray et al. 2002). Durch das Festlegen des Wertes der latenten Variablen geben die unterschiedlichen Antworten Aufschluss über die Unterschiede in den von den Befragten benutzten Schwellenwerten. Da der Wert der latenten Variablen aufgrund einer fehlenden direkten Beobachtung jedoch nicht einfach festgesetzt werden kann, müssen für dessen Festlegung andere Möglichkeiten genutzt werden. Eine Möglichkeit besteht darin, homogene Gruppen zu bilden; eine weitere ist die Verwendung von Vignette-Daten.

Homogene Gruppen

Um die latente Variable zu fixieren, können vergleichbare homogene Gruppen in verschiedenen Populationen gesucht und die subjektiven Antworten dieser Gruppen verglichen werden. Durch die Homogenität der Gruppen soll die latente Variable fixiert werden, wodurch die Unterschiede im Antwortverhalten auf Unterschiede in den Schwellenwerten zurückzuführen sind. Dabei ist von entscheidender Bedeutung, dass die Gruppen nicht anhand von Merkmalen identifiziert werden, welche die latente Variable messen. Für die Messung von Gesundheit schlagen Murray et al. (2002) z.B. vor, Veränderungen des Gesundheitszustandes oder Merkmale des Lebensstils oder des Berufslebens zu verwenden. Hierbei wird ein Nachteil des Vorgehens deutlich: Es ist schwierig festzustellen, ob Unterschiede in den Antwortkategorien auf Item Bias oder auf Item Impact zurückzuführen sind. Ein weiteres Problem besteht in der großen Menge an Daten, die benötigt wird, um sämtliche Variationen der Schwellenwerte in allen Antwortkategorien zu identifizieren.

Vignette-Fragen

Bei Vignette-Fragebögen ist es üblich, den Befragten zunächst die eigentlich interessierende Frage zu stellen, welche subjektiv beantwortet wird. Ein Beispiel dafür aus SHARE wäre: „Alles in allem gesehen, wie starke Schwierigkeiten hatten Sie während der letzten 30 Tage, sich zu bewegen?“ mit den Antwortkategorien: keine, geringe, mäßige, starke, extreme. Dabei können die Befragten die Skalierung der Antwortvorgaben individuell unter-

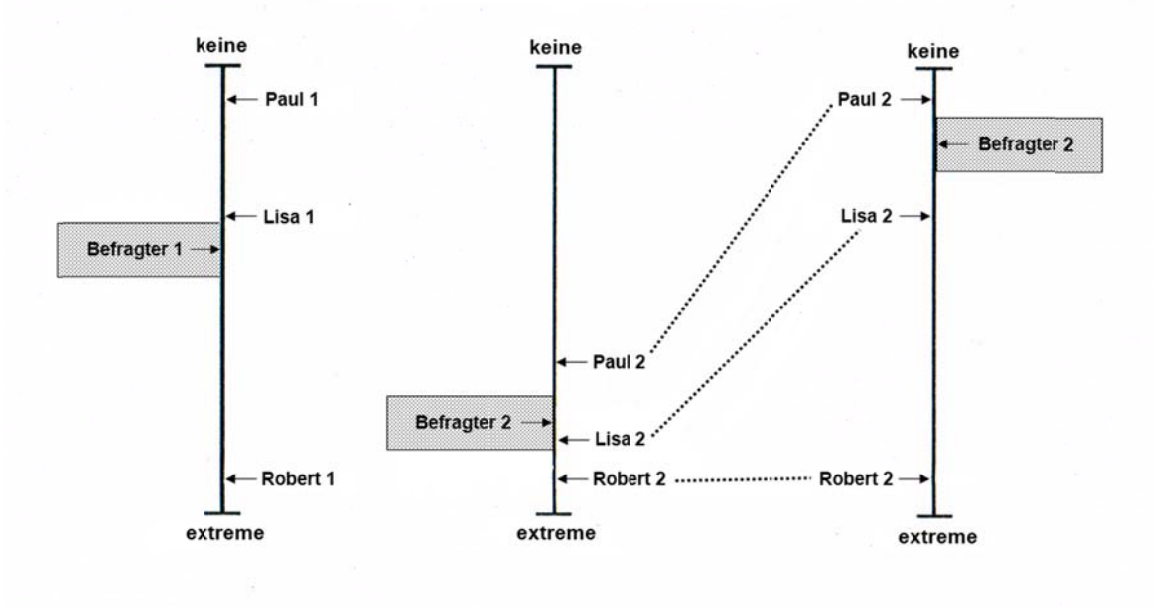
schiedlich auslegen. Im Anschluss werden dann die Vignette-Fragen gestellt. In diesen werden den befragten Personen kurze Text vorgelegt, die Personen beschreiben, die unterschiedlich starke Ausprägungen in der zu messenden Dimension haben, also z.B. unterschiedlich starke Einschränkungen der Mobilität. Dies kann dem Befragten wie folgt dargestellt werden:

„Robert kann Entfernungen bis zu 200 Metern ohne Probleme gehen, fühlt sich aber nach einem Kilometer oder dem Steigen von mehr als einem Treppenabsatz müde. Er hat keine Probleme bei täglichen Verrichtungen, wie Lebensmittel vom Einkaufen nach Hause tragen. Wie starke Schwierigkeiten hat Robert Ihrer Meinung nach, sich zu bewegen?“ Die Antwortkategorien entsprechen denen der Frage nach der individuellen Mobilitäteinschränkung: keine, geringe, mäßige, starke, extreme.

Es ist üblich Name und Geschlecht der in der Vignette beschriebenen Person zufällig zu wechseln (King et al. 2004). Der Gesundheitszustand, der in der Vignette beschrieben wird, ist in allen Ländern der gleiche, wobei auf einheitliche Übersetzung zu achten ist. Dadurch sind Unterschiede in der Einordnung der Mobilität in verschiedenen Ländern folglich auf DIF zurückzuführen.

In Abbildung 3 ist dargestellt, wie Vignetten genutzt werden, um Differenzen zwischen Antwortskalierungen aufzudecken. Auf der Skala keine bis extreme Schwierigkeiten sind links sowohl die Antwort von Befragten 1 abgetragen, als auch dessen Einordnung dreier fiktiver Vignette-Personen. In der Mitte findet sich das Gleiche für einen zweiten Befragten. Dieser ordnet sich so ein, dass er im Vergleich zum Befragten 1 größere Schwierigkeiten angibt. Vergleicht man aber die Einordnung der beiden Befragten in Bezug auf die Vignette-Fragen, so gibt Befragter 2 geringere Schwierigkeiten an. Deutlich wird, dass sich die Antwortskalierung der beiden Befragten stark voneinander unterscheidet, denn Befragter 2 ordnet die drei Vignette-Personen sehr viel enger an und näher an der Kategorie „extrem“ als Befragter 1. Dies geschieht, obwohl der gesundheitliche Zustand der Vignette-Personen für beide Befragte der gleiche ist. Rechts ist die Einordnung von Befragten 2 abgetragen, unter Anpassung der Antwortskalierung von Befragten 1 angepasst wurde. Erst jetzt sind die Antworten der beiden Befragten vergleichbar.

Abbildung 3: Vergleich der selbstberichteten Mobilität zweier Befragten und deren Antworten auf Vignette-Fragen



Quelle: In Anlehnung an Kapteyn (2010: 35)

Vergleichbarkeit durch einheitliche Skalen

Um beim Vorliegen von DIF einheitliche Skalen zu erstellen, können zwei unterschiedliche Strategien verfolgt werden. Zum Einen können mehrere Items zur Abbildung einer Dimension verwendet werden. Ein Beispiel dafür ist die Item Response Theorie. Zum Anderen kann Vergleichbarkeit unter Verwendung externer Informationen geschaffen werden. Diese Methode lässt sich mit Hilfe von objektiven Messungen und einem Hopit-Modell umsetzen (Murray et al. 2002).

Item Response Theorie

Die Item Response Theorie (IRT) ist eine psychologische Testtheorie. Genaugenommen ist die IRT nicht eine einzelne Theorie, sondern umfasst eine Familie von formalen, mathematischen, probabilistischen Messmodellen, welche postulieren, dass dem beobachtbaren Testverhalten (manifeste Variable) eine Fähigkeit, Eigenschaft bzw. Disposition (latente

Variable) zugrunde liegt, die das Testverhalten steuert. Das Herzstück der IRT stellt die Modellierung des Itemantwortverhaltens durch eine mathematische non-lineare Funktion, welche Item Response Function genannt wird, dar. Es werden Modelle, welche einen, zwei bzw. drei Itemparameter postulieren, unterschieden. Die einparametrischen Modelle beschreiben das Antwortverhalten mit Hilfe von einem einzigen Itemparameter, welcher die Positionierung eines Items auf der latenten Variablen bestimmt. Zu diesen Modellen gehört u.a. das bekannte eindimensionale Rasch-Modell. Zweiparametrische Modelle sind komplexere Modelle und nutzen neben dem Lokationsparameter einen zweiten Itemparameter, den Steigungsparameter, zur Spezifizierung der Beziehung zwischen dem beobachtbaren Antwortverhalten und der latenten Variable. Und schließlich wird in dreiparametrischen Modellen zusätzlich zu den beiden genannten Itemparametern ein Rateparameter konzipiert, welcher besonders bei der Modellierung des Antwortverhaltens in Tests, in denen Testpersonen möglicherweise die richtige Antwort raten können (z. B. Leistungstest), eine Rolle spielt. Modelle, welche sowohl zwei- als auch dreiparametrisch spezifiziert werden können, sind z. B. das Graded Response Modell und das Generalized Partial Credit Modell. Zur Beschreibung der einzelnen Modelle sei hier auf Embretson & Reise (2000) und Rost (2004) verwiesen. Durch die Schätzung der genannten Parameter können DIF aufgedeckt und einheitlichen Skalen konstruiert werden.

Im Rahmen des Problems der internationalen Vergleichbarkeit von selbstberichteten Daten ist die IRT trotz der verschiedenen Modellvarianten nur bedingt geeignet. Verzerrungen können nur aufgedeckt werden, wenn in einer Itembatterie die zu messende Dimension durch mindestens ein Item auch unverzerrt abgebildet wird. Sind alle Items systematisch verzerrt oder wurde gar nur ein Item erhoben (wie in unserem Fall), ist die IRT ohne zusätzliche Informationen nicht in der Lage DIF aufzudecken, und kann somit auch keine einheitlich Skalen schaffen (Angoff 1993; Penfield & Camilli 2007).

Objektive Messungen und Hopit-Modell

Eine mögliche Lösung des Problems der IRT besteht darin, zusätzliche Informationen in die Modelle aufzunehmen. Hiermit lassen sich systematische Verzerrungen, die bei allen Items vorliegen, erkennen. Eine Quelle für zusätzliche externe Informationen bieten objektive Messungen. Damit sind kleinere Tests oder medizinische Untersuchungen gemeint, die während einer Befragung durchgeführt werden können. Eine Voraussetzung der Nutzung dieser Messungen ist, dass der Test eine Dimension (vollständig) abbildet (Murray et al. 2002). Des Weiteren muss die Messung in verschiedenen Situationen durchführbar sein, ohne dass die Ergebnisse systematisch verzerrt werden (Murray et al. 2002), was insbesondere durch Unterschiede in den Befragungssituationen auftreten könnte. In einigen Studien

werden medizinische Messungen bereits durchgeführt. So finden sich im Survey of Health Ageing and Retirement in Europe (SHARE) beispielsweise die Messung der Greifkraft mit Hilfe eines Dynamometers (Mackenbach et al. 2005; Hank et al. 2009b) und die Messung der Gehgeschwindigkeit bei Personen über 75 Jahren. Medizinische Tests bzw. Messungen werden auch in anderen Studien, z.B. im Sozio-ökonomischen Panel und in English Longitudinal Study of Ageing, durchgeführt (Hank et al. 2009a).

Die durch derartige Untersuchungen und Messungen erhaltenen Informationen werden dann in einem sogenannten *hierachical ordered probit (Hopit*)-Modell verwendet (Tandon et al. 2002). Dieses unterscheidet sich von einem geordneten Probit-Modell darin, dass die Schwellenwerte der kategorialen abhängigen Variablen in Abhängigkeit von zusätzlichen Kovariaten modelliert werden. In diesem Fall fließen die Informationen der objektiven Messungen in die Modellierung der Schwellenwerte mit ein. Grenzen dieser Anwendung liegen in der Verfügbarkeit mindestens eines objektiven Maßes. Neben den objektiven Messungen können auch die Vignetten-Daten (alleine oder auch in Kombination mit objektiven Daten) zur Schätzung der individuellen Schwellenwerte herangezogen werden. Vignetten können denn auch in Bereiche anwendbar sein, die einer objektiven Messung nur schwer zugänglich sind.

3 Hopit-Modell

Die statistische Modellierung der Schwellenwerte bei ordinal skalierten Daten unter Verwendung von Vignetten anhand eines Hopit-Modells wurde erstmals von King et al. (2004) vorgeschlagen. Das Verfahren beruht auf zwei Annahmen.

Die erste Annahme bezieht sich auf das Antwortverhalten. Die Befragten beurteilen zum einen ihre eigene Situation, zum anderen auch die Situation der fiktiven Personen in den Vignetten. Die Annahme lautet, dass die Befragten bei der Beurteilung die gleiche Antwortskala verwenden (*Response Consistency*). Daraus folgt, dass sowohl die Antworten der selbstberichteten Fragen als auch die Antworten zu den Vignette-Fragen mit DIF behaftet sind, und zwar für jeden Befragten in annähernd der gleichen Art und Weise (King et al. 2004). Verletzt ist diese Annahme beispielsweise dann, wenn sich der Befragte den Personen, die in den Vignetten beschrieben sind, über- oder unterlegen fühlt und deshalb eine andere Skala anlegt. Da dies je nach untersuchtem Themengebiet durchaus möglich erscheint wird diese Annahme besonders kritisiert (Soest et al. 2007; Gupta et al. 2009; Vonkova & Hullegie 2010). Im Rahmen des Themas Mobilität scheint die Annahme der

Response Consistency allerdings plausibel, wie die Ergebnisse von Vonková & Hullegie (2010) zeigen.

Die zweite Annahme besagt, dass alle Befragten die Situationen in den Vignetten bis auf einen Zufallsfehler in der gleichen Weise wahrnehmen (*Vignette Equivalence*). Mit Hilfe der Vignetten soll bei jedem Befragten das Gleiche auf der gleichen Skala gemessen werden. Aus dieser Annahme folgt, dass Unterschiede zwischen den Befragten nur durch die Antwortskalierung zustande kommen. Grundlegend ist zudem davon auszugehen, dass die zu messende Variable in allen untersuchten Kulturen existiert und eine übereinstimmende Bedeutung hat.

Das statistische Modell besteht aus zwei Komponenten, einem Teil der die selbstberichtete Antwort modelliert und einem zweiten, der die Vignette-Daten einbindet. Grundlage für beide Teile ist ein geordnetes Probit-Modell, da die abhängige Variable (z.B. Beurteilung der Mobilität) jeweils als kategoriale Variable erhoben wird. Zusätzlich dazu werden die Schwellenwerte in Abhängigkeit von Ländern und individuellen Charakteristiken modelliert. Formal sieht die Modellierung nach King et al. (2004) wie folgt aus:

Y_i^* sei die subjektiv wahrgenommene Mobilität von Person *i*, für die gilt:

$$Y_i^* = X_i'\beta + \varepsilon_i \qquad (1)$$

Hierbei bezeichnet X_i erklärende Variablen mit fixen Effekten β und $\varepsilon_i \sim N(0,1)$ einen residualen Fehlerterm. Y_i^* ist nun nicht direkt beobachtbar, sondern nur die ordinale Variable Y_i, welche die Antworten der Befragten wiedergibt. Welche Antwortkategorie $k = 1, \dots, K$ der Befragte nennt, ist zum einen von der individuell wahrgenommenen Mobilität und zum anderen von den Schwellenwerten τ_i^k abhängig. Es gilt:

$$Y_i = k, \qquad \text{wenn} \quad \tau_i^{k-1} < Y_i^* < \tau_i^K \qquad (2)$$

mit $-\infty = \tau_i^0 < \tau_i^1 < \cdots < \tau_i^K = \infty$.

Im Unterschied zum klassischen ordinalen Probit-Modell werden die Schwellenwerte dann in Abhängigkeit von Kovariaten V modelliert, wobei $V = X$ sein kann, aber nicht muss:

$$\tau_i^1 = \gamma_1 V_i, \qquad \tau_i^K = \tau_i^{k-1} + \exp\{\gamma^k V_i\} \qquad (3)$$

γ_k bezeichnet die unbekannten Parameter der Antwortkategorie *k*. Der Umstand, dass τ_i^k für die Befragten *i* unterschiedlich sein kann entspricht DIF. Nur durch Verwendung der

selbstberichteten Mobilität könnten β und γ nicht getrennt identifiziert werden. Hierzu sind zusätzliche Informationen aus den Vignetten notwendig.

Sei Z_{ij}^* die vom Befragten *i* wahrgenommene Mobilität, welche in der Vignette *j* beschrieben wird. Dann ist

$$Z_{ij}^* = \theta_j + u_{ij}, \tag{4}$$

wobei θ_j die wahre Mobilität der Person in der Vignette-Frage ist und diese von dem Befragten nur mit einem Zufallsfehler u_{ij} wahrgenommen wird. Dieser ist normalverteilt mit $u_{ij} \sim N(0, \sigma^2)$. Zusätzlich wird angenommen, dass ε_i, u_{ij} und X unabhängig sind. Da auch Z_{ij}^* eine latente Variable ist, die bei der Antwort zu Grunde gelegt wird, gilt:

$$Z_{ij} = k, \text{ wenn } \tau_{i1}^{k-1} < Z_{ij}^* < \tau_i^k. \tag{5}$$

Die Schwellenwerte τ_{ij}^k sind dann durch die gleichen Parameter γ_k und die erklärenden Variablen V_i in (3) bestimmt. Es gilt:

$$\tau_{i1}^1 = \gamma_1^1 V_i, \quad \tau_{i1}^k = \tau_{i1}^{k-1} + \exp\{\gamma_1^k V_i\}. \tag{6}$$

Die Annahme der *Response Consistency* impliziert, dass τ_{ij}^k in den Gleichungen (3) und (6) gleich sind; *Vignette Equivalence* besagt, dass Z_{ij}^* und X_i unabhängig sind. Mit Hilfe dieser Annahmen können β und γ einzeln identifiziert werden. Anhand der Vignette-Gleichung lassen sich die Parameter γ und τ und mittels der selbstberichteten Mobilität der Parameter β identifizieren. Dafür genügt eine einzelne Vignette-Frage.

Üblicherweise werden die beiden Schritte simultan mit Hilfe von Maximum-Likelihood geschätzt (King et al. 2004; Kapteyn 2010). Hierzu wird jede Beobachtung im Datensatz verdoppelt, so dass für die Variable Mobilitätseinschränkung jeweils eine Angabe aus der Vignette und eine selbstberichtete Angabe vorliegen. Die Dummy-Variablen „Vignette“ und „Selbstberichtet“ geben an, aus welcher Frage die jeweilige Angabe stammt. Damit sich die zu schätzenden Koeffizienten der Kovariablen nur auf die Angaben der selbstberichteten Gesundheit beziehen, werden die Kovariablen mit der Dummy-Variable „Selbstberichtet“ multipliziert (für die technischen Details siehe Rabe-Hesketh & Skrondal 2002). Die Berechnungen des präsentierten Modells wurden mithilfe des für Stata entwickelten Programms gllamm Programm durchgeführt (Rabe-Hesketh & Skrondal 2008).

4 Daten und Variablen

Für die Berechnung des Modells wurden die Daten der zweiten Welle des Survey of Health, Ageing and Retirement in Europe (SHARE, Release 2.3.0) verwendet.[2] SHARE ist eine Längsschnittstudie, die erstmals ab 2004 in elf europäischen Ländern und Israel durchgeführt wurde. Die zweite Befragungswelle fand ab Herbst 2006 bis Frühjahr 2007 in 14 europäischen Ländern statt. Dabei wurden repräsentative Daten von über 50 jährigen Personen und allen Haushaltsmitgliedern erhoben. Bei der Befragung mussten die Haushaltsmitglieder das Alter 50 noch nicht erreicht haben. Zusätzlich zum Hauptfragebogen wurde in beiden Welle von SHARE in einigen Ländern von einem Teil der Befragten ein Vignette-Fragebogen ausgefüllt. In der zweiten Welle haben Deutschland, Niederlande, Frankreich, Belgien, Schweden, Dänemark, Tschechien, Polen, Italien, Spanien und Griechenland am Vignette-Fragebogen teilgenommen.

In der zweiten Welle wurden zwei unterschiedliche Versionen des Vignette-Fragebogens verwendet, die Zuordnung erfolgt bezogen auf das Alter der Personen. Personen 65 Jahre und älter erhielten Version C des Fragebogens und Personen unter 65 Jahren erhielten Version B. Die Versionen unterscheiden sich hinsichtlich der Reihenfolge der Fragen und des Geschlechts der Personen in den Vignette-Fragen. In der zweiten Welle ist nur noch eine Vignette-Frage zur Mobilität enthalten (SHARE 2010).[3]

Für die folgenden Analysen wurde die Stichprobe des Vignette-Samples (7731 Fälle) auf Personen über 50 Jahre beschränkt und alle Fälle mit fehlenden Daten ausgeschlossen. Die für die Analysen verwendete Stichprobe beinhaltet insgesamt 6978 Fälle aus elf Ländern. Deutschland ist mit 15,38% am stärksten vertreten und Frankreich liefert mit 4,97% den kleinsten Anteil. Die geringe Fallzahl in Frankreich ist auf Unterschiede in der Erhebung zurück zuführen, da hier die Personen, die in der zweiten Welle neu aufgenommen wurden, nicht an den Vignette-Fragebögen teilgenommen haben (SHARE 2010). Der Anteil der Männer liegt in der Stichprobe bei 45,73% und im Schnitt sind die Befragten 64 Jahre alt (vgl. Tabelle 1).

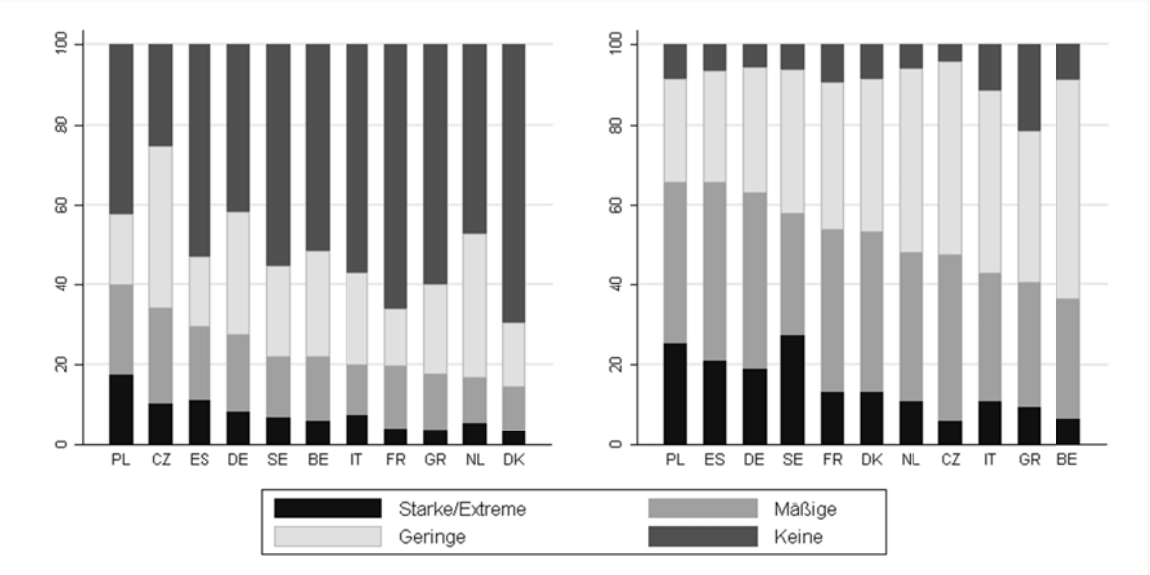
---

[2] Detaillierte Informationen über SHARE finden sich unter www.share-project.org. Die SHARE-Datenerhebung wurde hauptsächlich durch das 5. und 6. Forschungsrahmenprogramm der Europäischen Union finanziert (Projekte QLK6-CT-2001-00360; RII-CT-2006-062193; CIT5-CT-2005-028857). Weitere Finanzmittel wurden vom US National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01, OGHA 04-064, R21 AG025169), sowie nationalen Geldgebern zur Verfügung gestellt.

[3] Die Dokumentation ist jeweils für das aktuelle Release online verfügbar. Die Dokumentation für Release 2.3.0 ist auf Anfrage bei den Autoren erhältlich.

Die selbstberichtete Mobilität wurde in SHARE mit einer fünfstufigen Skala erfasst. Da die letzte Kategorie nicht stark besetzt ist, wurde die vierte und fünfte Kategorie zusammengefasst. Zudem wurde die Polung beider Variablen geändert, um eine intuitive Interpretation der Schätzkoeffizienten zu ermöglichen. In Abbildung 4 sind die länderspezifischen Verteilungen der Antworten auf die Frage nach Schwierigkeiten mit der eigenen Mobilität dargestellt. Da die Alters- und Geschlechtsverteilung über die Länder variiert wurden die Verteilungen nach Alter und Geschlecht standardisiert (Preston et al. 2001). Die Länder sind nach dem Anteil der Befragten geordnet, die angegeben haben *starke/extreme* oder *mäßige* Probleme zu haben. Standardisiert nach Alter und Geschlecht haben die Befragten in Polen die größten Schwierigkeiten mit der Mobilität und in Dänemark die geringsten. Der höchste Anteil der Kategorie *starke/extreme* oder *mäßige* Schwierigkeiten wird in Polen mit knapp 40% erreicht.

Abbildung 4: Länderspezifische Verteilungen der selbstberichteten Mobilitätsprobleme, standardisiert nach Alter und Geschlecht (links) und Mobilitätsprobleme in den den Vignetten (rechts)



Quelle: SHARE Release 2.3.0; eigene Berechnungen

Im Vergleich dazu ist der Anteil der Befragten, die angeben, dass die in der Vignette beschriebene Person *starke/extreme* oder *mäßige* Schwierigkeiten hat, in allen Ländern deutlich höher. Auch hier erreicht Polen mit ca. 65% den höchsten Wert (vgl. Abbildung 4). Die

Mobilität der Person in der Vignette-Frage wird in allen Ländern mit größeren Schwierigkeiten verbunden und der Anteil der Befragten, die angeben, dass die Person *keine* Schwierigkeiten hat ist deutlich geringer. Der Anteil verschiebt sich zu Gunsten der mittleren Kategorien.

In das Hopit-Modell wurden zur Schätzung der Parameter sowie der Schwellenwerte verschiedene Variablen aufgenommen, die zur Erklärung der Mobilität von Bedeutung sind. Dabei wurden auch sogenannte objektive Gesundheitsmaße aufgenommen. Wie erwähnt, können solche Variablen benutzt werden, um einheitliche Antwortskalierungen zu schaffen. In der vorliegenden Arbeit dienen sie dazu, die Erklärungskraft des Modells zu verbessern. Wie sich bei den Analysen zeigen wird, haben solche objektiven Maße keine Variation in den Schwellenwerten, denn durch ihre Objektivität gibt es keine unterschiedlichen Antwortmuster.

Ergänzend wurden soziodemographische Variablen berücksichtigt. Diese sind das Alter der Befragten, das Geschlecht sowie der Bildungsgrad in der ISCED Codierung. Weiterhin wurde mittels Dummy-Variablen berücksichtigt, ob der Befragte erwerbstätig (abhängig beschäftigt oder Selbstständig), erwerbsunfähig (wegen Krankheit oder Behinderung), oder nicht erwerbstätig (arbeitslos oder Rentner) ist. Die Referenzkategorie bilden Befragte, die angaben, Hausmann bzw. -frau zu sein oder eine andere Erwerbssituation haben, z.B. von Vermögen leben. Die gesundheitsrelevanten Variablen beinhalten Fragen über Aktivitäten des täglichen Lebens, BMI und gesundheitliche Probleme. Außerdem wurde berücksichtigt, ob die Person alleine lebt, Probleme bei der Verrichtung täglicher Aufgaben hat oder chronische Krankheiten hat. Aus der Liste der Aktivitäten des täglichen Lebens (activities of daily living, kurz ADL) wurden sieben ausgewählt, die für die Analysen relevant sind. Berücksichtigung fanden: 1. sich anziehen, einschließlich Socken und Schuhe, 2.durch einen Raum gehen, 3. baden oder duschen, 4. ins Bett legen oder aus dem Bett aufstehen, 5. Benutzen der Toilette, 6. einkaufen von Lebensmitteln und 7. arbeiten im Haus oder im Garten.

Zur Erfassung gesundheitlicher Probleme beschränken wir uns auf die mobilitätsrelevanten Beeinträchtigungen. Die entsprechende Dummy-Variable nimmt den Wert 1 an, wenn eine oder mehr der folgenden gesundheitlichen Beeinträchtigungen vorliegen: Schmerzen im Rücken, im Knie, in der Hüfte oder in einem anderen Gelenk, Herzprobleme oder Angina Pectoris, Schmerzen in der Brust bei körperlicher Betätigung, Atemnot, geschwollene Beine, Hinfallen, Angst davor hinzufallen, Schwindel, Ohnmacht, kurzzeitige Bewusstlosigkeit, Ermüdung und Erschöpfung.

5 Ergebnisse

Die Ergebnisse des Hopit-Modells finden sich in Tabelle 1. In der ersten Spalte sind die Schätzwerte für Koeffizienten aus Gleichung (1) aufgelistet und in den Spalten zwei bis vier die Schätzungen der Schwellenwertgleichungen. Die Schätzkoeffizienten der Grundgleichung entsprechen den Erwartungen und sind größtenteils signifikant. Nicht signifikante Effekte im Vergleich zu Hausfrauen und -männern finden sich für erwerbstätige Personen sowie für Rentner und Arbeitslose. Auch für allein lebende Personen findet sich kein signifikanter Effekt im Vergleich zu Personen, die mit einer oder mehreren Personen zusammenleben. Alle anderen Kovariaten weisen signifikante Effekte auf.

Die in das Modell aufgenommenen Variablen zur Gesundheit sind in der Grundgleichung alle signifikant. Jedoch gibt es kaum signifikante Effekte in den Schwellenwertgleichungen. Dies ist darauf zurückzuführen, dass die verwendeten Variablen als annähernd objektive Maße der Gesundheit bzw. Mobilität gesehen werden können. Daher ist hier nicht mit unterschiedlichem Antwortverhalten zwischen Gruppen zu rechnen.

Im Einzelnen stellen sich die Effekte der erklärenden Variablen wie folgt dar: Befragte, die angegeben haben, in den letzten sechs Monaten in ihren Aktivitäten eingeschränkt gewesen zu sein, geben größere Schwierigkeiten mit der Mobilität an als Befragte, die nicht eingeschränkt waren oder sind. Ebenso haben Personen, die langjährige Erkrankungen haben, größere Schwierigkeiten mit der Mobilität, als Personen ohne langjährige Erkrankungen. Das Vorliegen von Einschränkungen der Mobilität sowie gesundheitlicher Probleme und das Vorliegen von Problemen mit ADL wirken sich ebenso negativ auf die selbstberichtete Mobilität aus. Auch Menschen mit Übergewicht und Adipositas haben mehr Probleme mit Mobilität als normalgewichtige Personen, sowie Personen mit Untergewicht im Vergleich zu Personen mit Normalgewicht. Für das Alter ergibt sich ein negativer Einfluss auf die selbstberichtete Mobilität, d.h. mit zunehmendem Alter steigt die Wahrnehmung der Mobilitätseinschränkung. Männer berichten im Vergleich zu Frauen von weniger Schwierigkeiten mit der Mobilität. Hier überraschen die nicht signifikanten Werte aus den Schwellenwertgleichungen, da in der Literatur darauf hingewiesen wird, dass sich Frauen und Männer aufgrund von Erziehung, hinsichtlich des Berichtens von gesundheitlichen Einschränkungen, unterscheiden (Jürges 2008). Der Effekt in der Mobilitätsgleichung ist jedoch mit dieser Einschätzung konsistent. Weiterhin geben sowohl Personen mit geringer als auch Personen mit mittlerer Bildung mehr Probleme mit der Mobilität an, als Personen mit hoher Bildung. Personen, die aufgrund von Berufsunfähigkeit nicht erwerbstätig sind, berichten größere Schwierigkeiten mit der Mobilität zu haben, als Hausfrauen bzw. -männer.

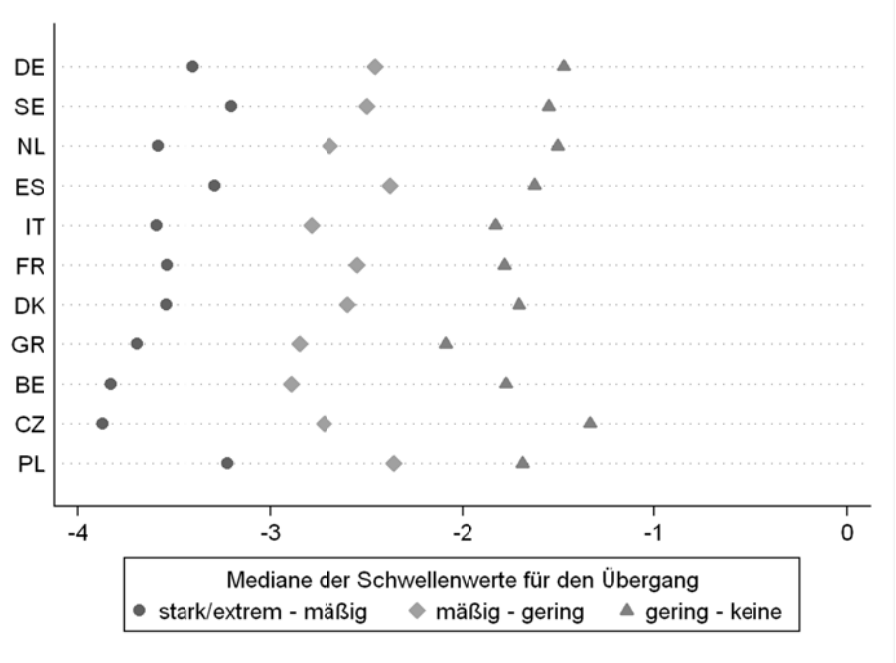
Tabelle 1: Hopit-Modell für selbstberichte Mobilität

	Mittelwerte	Selbstberichtete Mobilität		Schwellenwertgleichungen τ^1		τ^2		τ^3	
Alter	64,90	-0,017	***	-0,005	**	0,003		0,003	
		(0,003)		(0,002)		(0,002)		(0,002)	
Mann	0,46	0,122	**	0,123	***	-0,044		-0,002	
		(0,039)		(0,029)		(0,031)		(0,028)	
Allein lebend	0,18	-0,009		0,005		-0,012		-0,029	
		(0,047)		(0,036)		(0,037)		(0,034)	
Aktiv	0,57	0,527	***	-0,039		-0,015		0,039	
		(0,047)		(0,036)		(0,038)		(0,034)	
Geringe Bildung	0,27	-0,194	***	-0,030		-0,004		-0,046	
		(0,058)		(0,042)		(0,045)		(0,043)	
Mittlere Bildung	0,50	-0,208	***	-0,110	**	-0,010		0,027	
		(0,048)		(0,034)		(0,037)		(0,035)	
Krank	0,49	-0,170	***	0,000		0,032		-0,023	
		(0,045)		(0,034)		(0,036)		(0,032)	
Erwerbstätig	0,29	0,035		0,002		0,015		-0,019	
		(0,071)		(0,052)		(0,058)		(0,051)	
Rentner/arbeitslos	0,55	0,007		-0,049		0,018		-0,007	
		(0,062)		(0,047)		(0,051)		(0,045)	
Erwerbsunfähig	0,04	-0,541	***	-0,108		0,027		-0,038	
		(0,102)		(0,078)		(0,080)		(0,081)	
ADL	0,13	-0,574	***	0,107	*	-0,045		-0,057	
		(0,055)		(0,043)		(0,041)		(0,045)	
Eingeschränkte Mobilität	0,46	-0,381	***	-0,071	*	0,043		0,091	**
		(0,045)		(0,034)		(0,036)		(0,031)	
Gesundheitliche Probleme	0,61	-0,296	***	-0,011		0,070	*	0,045	
		(0,045)		(0,032)		(0,036)		(0,031)	
Untergewicht	0,01	-0,535	**	-0,187		0,031		-0,020	
		(0,183)		(0,155)		(0,153)		(0,137)	
Übergewicht	0,44	-0,109	**	0,002		-0,012		0,036	
		(0,041)		(0,030)		(0,033)		(0,029)	
Adipositas	0,19	-0,299	***	-0,034		0,057		-0,017	
		(0,050)		(0,038)		(0,039)		(0,038)	
Schweden	0,06	0,249	**	0,162	**	-0,283	***	0,000	
		(0,087)		(0,058)		(0,071)		(0,066)	
Niederlande	0,07	-0,039		-0,225	***	-0,017		0,235	***
		(0,084)		(0,061)		(0,067)		(0,058)	
Spanien	0,06	0,133		0,080		-0,040		-0,229	**
		(0,091)		(0,062)		(0,068)		(0,076)	
Italien	0,09	-0,016		-0,212	***	-0,150	*	0,002	
		(0,079)		(0,058)		(0,066)		(0,056)	
Frankreich	0,05	0,327	***	-0,163	*	0,040		-0,204	**
		(0,096)		(0,069)		(0,071)		(0,075)	
Dänemark	0,13	0,402	***	-0,177	***	0,020		-0,062	
		(0,071)		(0,050)		(0,053)		(0,053)	
Griechenland	0,07	-0,244	**	-0,324	***	-0,095		-0,221	***
		(0,084)		(0,064)		(0,072)		(0,063)	
Belgien	0,12	-0,175	*	-0,445	***	0,003		0,150	**
		(0,069)		(0,056)		(0,058)		(0,048)	
Tschechien	0,12	-0,318	***	-0,461	***	0,175	***	0,359	***
		(0,067)		(0,055)		(0,051)		(0,047)	
Polen	0,08	0,253	**	0,158	**	-0,120	*	-0,333	***
		(0,081)		(0,055)		(0,061)		(0,070)	
Vignette Dummy		-2,512	***						
		(0,186)							
Konstante				-2,756	***	-0,371	*	-0,305	*
				(0,197)		(0,148)		(0,136)	

Anmerkung: Standardfehler in Klammern; * $p < 0,05$, ** $p < 0,01$, *** $p < 0,001$.
Quelle: SHARE Release 2.3.0; eigene Berechnungen

Aus Tabelle 1 wird darüber hinaus ersichtlich, dass sich bis auf die Niederlande, Italien und Spanien alle Länder signifikant von Deutschland in der Mobilitätsgleichung unterscheiden. Dabei berichten die Dänen, Franzosen, Polen und Schweden weniger Schwierigkeiten mit der Mobilität als die Deutschen, die tschechischen, griechischen und belgischen Befragten dagegen mehr. Darüber hinaus zeigt der geschätzte Wert der Vignette-Dummy, dass die Befragten die Mobilität der Person in der Vignetten-Frage schlechter einschätzen als die eigene. Dies wurde bereits im Vergleich der beiden Verteilungen in Abbildung 4 deutlich.

Abbildung 5: Mediane der geschätzten Schwellenwerte getrennt nach Ländern



Quelle: SHARE Release 2.3.0; eigene Berechnungen

In Abbildung 5 sind die länderspezifischen Mediane der geschätzten Schwellenwerte dargestellt. Hier werden große Unterschiede zwischen den in das Modell aufgenommenen Ländern deutlich. Besonders auffällig sind die Schwellenwerte der tschechischen Befragten. Für sie finden sich der niedrigste erste Schwellenwert und der höchste dritte Schwellenwert. Dies bedeutet, dass Tschechen im Vergleich zu Befragten aus anderen Ländern viel früher bereits mit *mäßig* antworten, gleichzeitig aber trotz guter Mobilität noch mit *gering* antworten, statt anzugeben, *keine* Schwierigkeiten zu haben. Tschechische Befragte

neigen dazu, die beiden mittleren Kategorien zu wählen. Im Gegensatz dazu sind die polnischen Befragten zu sehen. Hier liegen die Schwellenwerte deutlich näher zusammen, so dass hier die Befragten mehr zu den Randkategorien neigen. Überspitzt dargestellt könnte man sagen, dass polnische Befragte zu Extremen neigen, was ihr Antwortverhalten angeht. In der Stichprobe ist der niedrigste Schwellenwert des Übergangs von Kategorie *gering* zu *keine* in Griechenland zu finden. Trotz Problemen mit der Mobilität geben griechische Befragte also noch an *keine* Schwierigkeiten zu haben. Im Vergleich dazu ist der größte Schwellenwert für diesen Übergang in Tschechien zu finden, ebenso wie der kleinste für den Übergang von der Kategorie *starke/extreme* zu *mäßige* Schwierigkeiten. Besonders ähnlich sind sich die Schwellenwerte in Frankreich und Dänemark und die größten Unterschiede scheinen zwischen den tschechischen und den polnischen Befragten zu bestehen. Warum diese Unterschiede bestehen und worin ihre Ursache zu sehen ist, ist nicht Thema dieses Artikels, aber bedarf sicherlich einer genaueren Betrachtung.

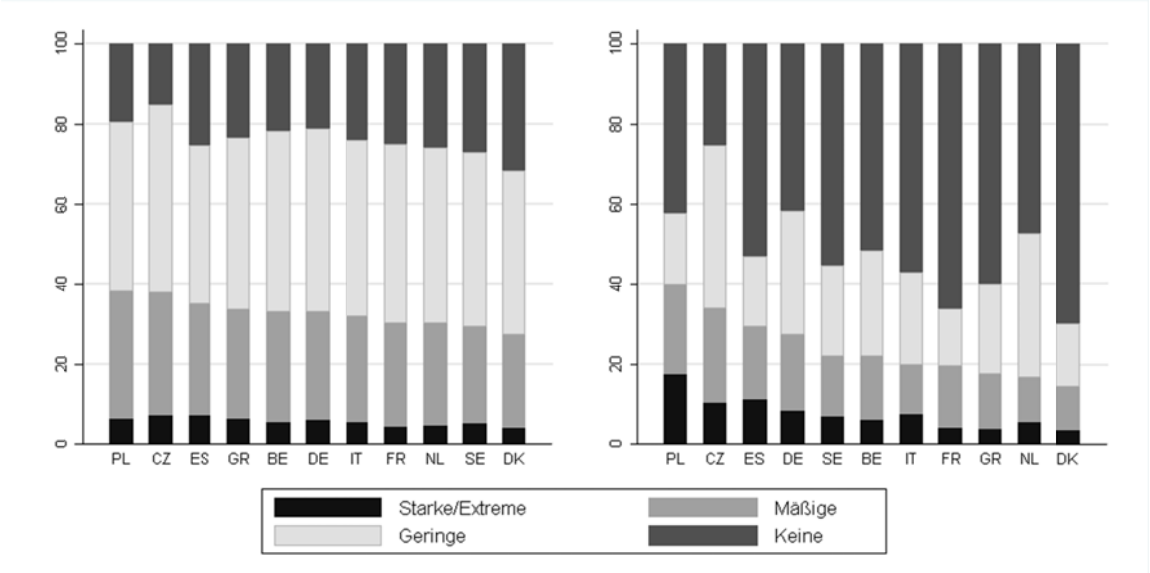
Im Weiteren sollen die Schwellenwerte von Tschechien und Polen auf die anderen Länder angewandt werden. Dieses Vorgehen gibt Aufschluss über die Auswirkungen der Unterschiede im länderspezifischen Antwortverhalten.

Simulation

Für die in Abbildung 6 dargestellten Simulationsergebnisse wurden jedem Land statt der eigenen Schwellenwerte die Schwellenwerte der tschechischen Befragten zugewiesen. Damit entfällt das länderspezifische Antwortmuster, so als ob jeder Befragte die gleiche Skalierung der Antwortkategorien wie die tschechischen Befragten verwenden würde. Verglichen werden soll dies mit der länderspezifischen Verteilung der selbstberichteten Mobilität.

Auffällig ist zunächst, dass in der Simulation der Anteil der Personen, die angegeben haben, *keine* Schwierigkeiten mit ihrer Mobilität zu haben, deutlich geringer ist. Liegt der Anteil für die selbstberichtete Mobilität ohne Anpassungen in Dänemark bei ca. 70%, so sinkt er für die simulierten Werte auf ca. 30%. Unter der Annahme, dass alle Befragten die Antwortskalierungen der tschechischen Befragten anwenden, ist es damit aber dennoch der größte Anteil unter den untersuchten Ländern. Auch sinkt der Anteil der Personen, die angeben *starke/extreme* Schwierigkeiten zu haben. Besonders deutlich wird dies bei den Befragten in Polen. Hier sinkt der entsprechende Anteil von ca. 18% auf knapp über 5%.

Abbildung 6: Verteilung der simulierten Werte selbstberichteter Mobilität mit tschechischen Schwellenwerten (links) und im Vergleich die Verteilung der selbstberichteten Mobilität, standardisiert nach Alter und Geschlecht (rechts)



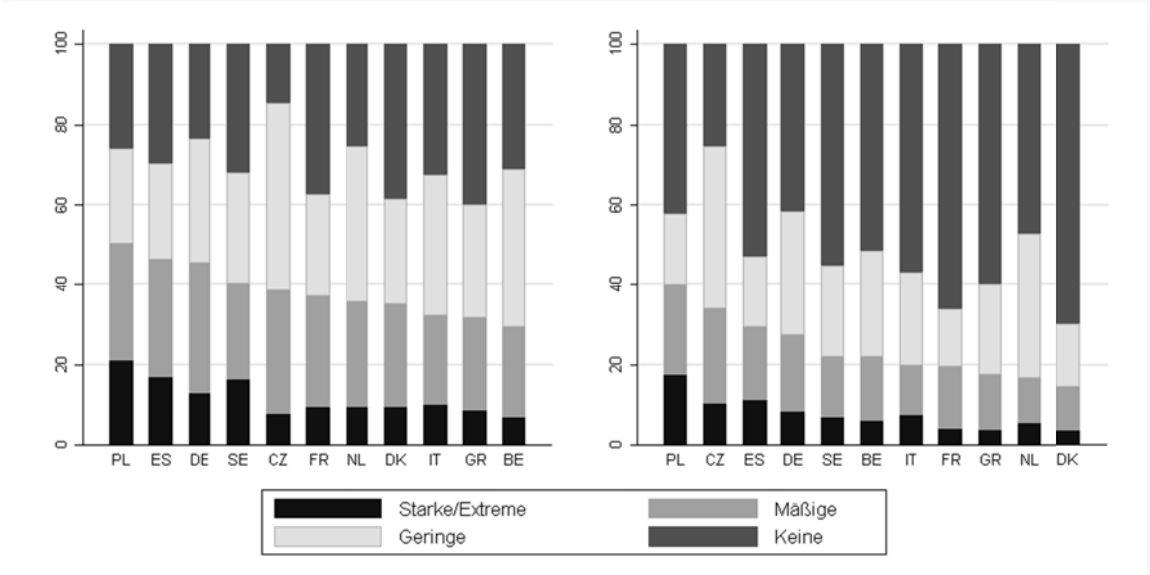
Quelle: SHARE Release 2.3.0; eigene Berechnungen

In allen Ländern steigt dagegen durch die Anwendung der tschechischen Schwellenwerte der Anteil der Personen die angegeben haben *mäßige* oder *geringe* Schwierigkeiten zu haben. Ursache dafür ist der sehr niedrige Schwellenwert für den Übergang von der Kategorie *starke/extreme* zu *mäßige* und der sehr hohe Schwellenwert für den Wechsel von Kategorie *geringe* zu *keine*. Dadurch ändert sich auch die Reihenfolge der Länder. In beiden Grafiken sind diese nach der Größe des Anteils in den beiden Kategorien *starke/extreme* und *mäßige* sortiert. Schweden sinkt hier in der Rangfolge von Platz fünf für die rohen Werte auf Platz 10, da besonders der Anteil in der Kategorie *mäßig* enorm zu nimmt.

Die Unterschiede zwischen den Ländern nehmen mit den simulierten Werten jedoch deutlich ab. Dies hat zwei Gründe: Zum einen fällt durch die Verwendung einheitlicher Schwellenwerte, in diesem Fall der tschechischen, die Variation weg, die allein auf Unterschiede im Antwortverhalten zurückzuführen ist. Dies ist Ziel des verwendeten Modells gewesen. Zum anderen sind in das Modell objektive Maße eingeflossen, welche ebenfalls die Variation der selbstberichteten Mobilität zwischen den Ländern verringern (vgl. Abbildung 8).

Die Verwendung der tschechischen Schwellenwerte führt dazu, dass in allen Ländern der Anteil der Personen mit *keinen* Schwierigkeiten sinkt. Die Beurteilung wird in allen Ländern vermehrt in die mittleren Kategorien verlagert. Ohne Anpassungen würde davon ausgegangen werden, dass der Großteil der befragten Personen *keine* Schwierigkeiten mit der Mobilität hat. Dies ändert sich durch die Verwendung der tschechischen Schwellenwerte, da hier die gleichen Befragten *geringe*, teilweise sogar *mäßige* Schwierigkeiten angegeben hätten.

Abbildung 7: Verteilung der simulierten Werte der selbstberichteten Mobilität mit jeweils landbezogenem Schwellenwert (links) und im Vergleich die Verteilung der selbstberichteten Mobilität, standardisiert nach Alter und Geschlecht (rechts)



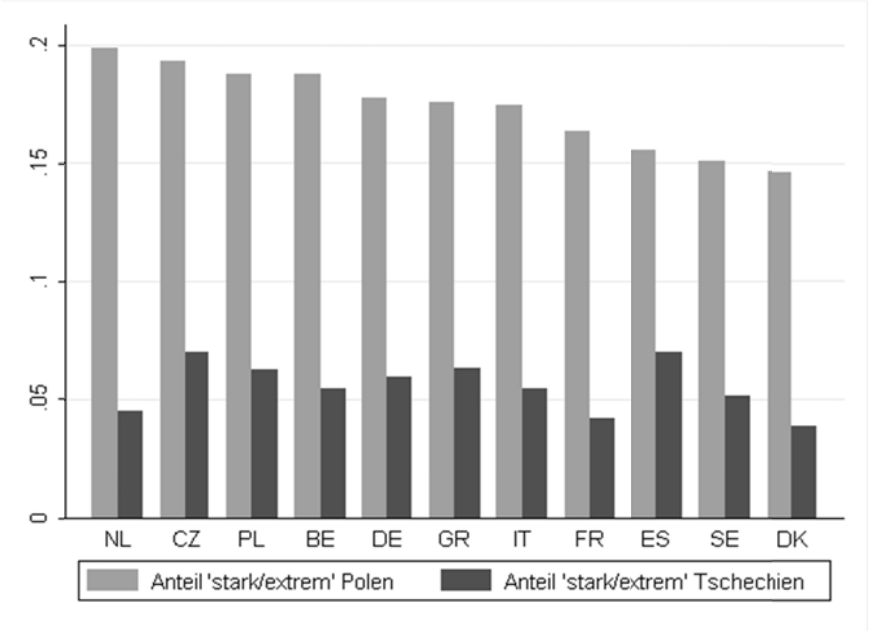
Quelle: SHARE Release 2.3.0; eigene Berechnungen

Kontrafaktische Simulation: polnische Schwellenwerte

Im Folgenden soll verglichen werden, welche Auswirkung die unterschiedliche Verwendung von Schwellenwerten hat. In Abbildung 9 sind die Anteile der Kategorie *starke/extreme* Schwierigkeiten für jedes Land abgebildet, sowohl unter Verwendung der polnischen als auch der tschechischen Schwellenwerte.

Werden die polnischen Schwellenwerte angewandt, liegt der Anteil der Personen die angeben, *starke/extreme* Schwierigkeiten mit ihrer Mobilität zu haben, deutlich höher als bei der Verwendung der tschechischen Schwellenwerte.

Abbildung 8: Simulierter Anteil Personen mit starken/extremen Mobilitätsproblemen bei polnischen und tschechischen Schwellwerten



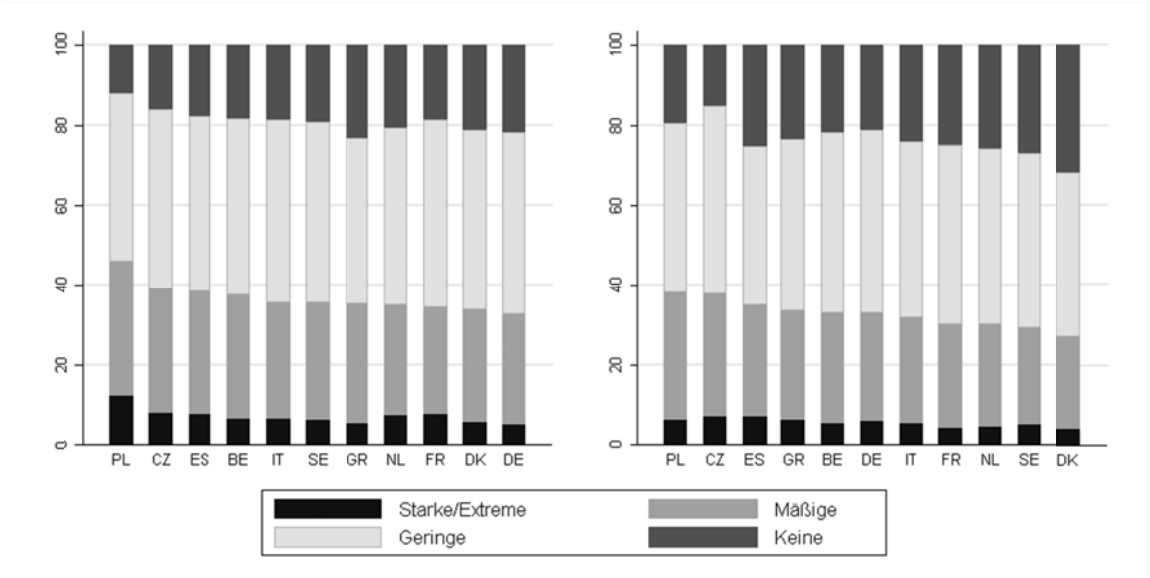
Quelle: SHARE Release 2.3.0; eigene Berechnungen

Für die tschechischen Schwellenwerte liegt der Anteil für alle Länder unter 0,07, für die polnischen Schwellenwerte über 0,13. Die Unterschiede, die durch Verwendung einheitlicher Schwellenwerte deutlich werden zeigen auf, dass ein Vergleich der selbstberichteten Mobilität ohne Anpassungen zu falschen Schlüssen führen würde. Denn trotz gleichem objektivem Mobilitätswert geben Befragte unter Verwendung der tschechischen Schwellenwerte an, *mäßige* Schwierigkeiten zu haben, während unter Verwendung polnischer Schwellenwerte die gleichen Befragten *starke/extreme* Schwierigkeiten angeben würden. Daher führen Vergleiche der Daten ohne Anpassung dazu, dass in einzelnen Ländern die Schwierigkeiten der Befragten mit der Mobilität unter- in anderen jedoch überschätzt würden. Eine Anpassung der verwendeten Antwortskalierung ist daher dringend notwendig.

Ebenso wird in Abbildung 9 deutlich, dass je nach verwendeter Skalierung die Schlussfolgerungen unterschiedlich ausfallen. Bei Verwendung der tschechischen Schwellenwerte hätte Spanien den größten Anteil an Personen mit *starken/extremen* Schwierigkeiten. Wird die polnische Antwortskalierung verwendet, so hätten die Niederlande den größten Anteil. Wie bereits erwähnt, unterscheidet sich auch der Anteil in den Länder deutlich, je nach verwendeten Schwellenwerten. Zusammenfassend muss also festgehalten werden, dass bei der Verwendung simulierter Daten mit einheitlichen Schwellenwerten für internationale Vergleiche, genau darauf geachtet werden sollte, ob dies nicht zu falschen Schlussfolgerungen führen könnte.

Durch die Simulation wird deutlich, dass ein erheblicher Teil der Unterschiede der Mobilität zwischen den Ländern auf die verwendeten Antwortskalierungen der Befragten zurückzuführen ist. Die Ungleichheiten zwischen den Ländern wären ohne Anpassung der Skalierung deutlich überschätzt. Die Variation, welche nach Anpassung der Schwellenwerte zwischen den Ländern bestehen bleibt, ist vermutlich z.T. auf institutionelle Unterschiede in den Gesundheitssystemen zurückzuführen. Werden in der Simulation nicht nur die Schwellenwerte eines Landes, sondern auch dessen Parameter auf alle anderen Länder angewendet werden die Abweichungen zwischen den Ländern noch geringer (vgl. Abbildung 9). Übrig bleibt dann die Variation, die auf unbeobachtete Heterogenität zurückzuführen ist.

Abbildung 9: Verteilung der simulierten Werte der selbstberichteten Mobilitätsprobleme mit tschechischen Schwellenwerten und Parametern (links) sowie nur mit tschechischen Schwellenwerten (rechts)



Quelle: SHARE Release 2.3.0; eigene Berechnungen

5 Zusammenfassung und Diskussion

Das Anliegen dieser Arbeit war, den Leser bezüglich der Probleme des Vergleichs subjektiver Daten in der empirischen Sozialforschung zu sensibilisieren. Bei einem Vergleich subjektiver Daten besteht immer die Möglichkeit, dass Antwortkategorien unterschiedlich ausgelegt werden, so dass ein direkter Vergleich ad hoc nicht möglich ist. Hierzu wurde ein Überblick über die Probleme und die Methoden des subjektiven Vergleichs gegeben und ein innovatives Verfahren vorgestellt, welches auf Vignetten basiert, die anhand eines Hopit-Modells zur Schätzung der zugrunde liegenden Skalierungen verwendeten werden.

Am Beispiel der in SHARE erhobenen selbstberichteten Mobilitätseinschränkungen in Verbindung mit entsprechenden Vignetten-Daten zeigen sich große Unterschiede in der berichteten Mobilität, besonders zwischen den Befragten in Polen und Tschechien. Unter Berücksichtigung tschechischer oder polnischer Schwellenwerte ändert sich im internationalen Vergleich der Anteil Personen, die angeben *starke/extreme* Schwierigkeiten bei der Bewegung zu haben um bis zu 12 Prozentpunkte. Der Vergleich der subjektiven Antworten

gibt damit nicht nur Unterschiede in objektiven Gesundheitszuständen wieder, sondern auch Differenzen in den Antwortskalierungen zwischen Ländern oder Personengruppen. Daher ist es unabdingbar notwendig, die Antwortskalierungen beim Vergleich subjektiver Daten zu berücksichtigen.

Dabei ist auch die Frage, warum sich die betrachteten Länder mehr oder weniger in ihren Schwellenwerten unterscheiden, von Bedeutung. In der vorliegenden Arbeit sind insbesondere die Unterschiede zwischen Tschechien und Polen auffällig, aber auch die Ähnlichkeiten in den verwendeten Schwellenwerten von Frankreich, Italien und Dänemark. Für die weitere Forschung bleibt zu klären worin die Ursachen für diese Gemeinsamkeiten und Unterschiede bestehen. Ein lohnenswerter Ansatzpunkt könnten hierbei die jeweiligen institutionellen Rahmenbedingungen der Länder sein.

Weiterhin bleibt offen, welche Anpassungen der Schwellenwerte vorgenommen werden sollten, wenn Unterschiede in den Skalierungen aufgedeckt wurden. Wie gezeigt wurde variieren die Ergebnisse in Abhängigkeit von den gewählten Schwellenwerten zur Standardisierung. Ein möglicher Ansatz könnte darin bestehen, wie im Falle von Altersstandardisierungen, Standardbevölkerungen als Basis für die Schwellenwerte zu verwenden. Da die Wahl der Standardbevölkerung jedoch a priori immer willkürlich ist, muss die Wahl der Standard-Schwellenwerte letztlich vom Erkenntnisinteresse abhängig sein. Denkbar wäre neben der Wahl der Schwellenwerte einer bestimmten Population auch die Wahl synthetischer (z.B. mittlerer) Schwellenwerte. Die Anpassung der Schwellenwerte sollte jedoch unbedingt erkenntnisorientiert und nicht ergebnisorientiert erfolgen.

Literatur

Angelini, V. / Cavapozzi, D. / Corazzini, L. / Paccagnella, O. 2009. Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases. In: "Marco Fanno" Working Paper (90). Universität von Padua.

Angoff, W.H. 1993. Perspectives on Differential Item Functioning Methodology. S. 3-23 in: P.W. Holland / H. Wainer (Hrsg.), Differential Item Functioning. Hillsdale: Lawrence Erlbaum Associates.

Durkheim, E. 1991. Die Regeln der soziologischen Methode. Herausgegeben und eingeleitet von René König, 2. Auflage. Frankfurt a.M.: Suhrkamp.

Embretson, S.E. / Reise, S.P. 2000. Item Response Theory for Psychologists. Mahwah, NJ: Erlbaum.

Gauthier, A.H. 2002. The Promises of Comparative Research. Schmollers Jahrbuch 122(1): 5–30.

Gupta, N.D. / Kristensen, N. / Pozzoli, D. 2009. External Validation of the Use of Vignettes in Cross-Country Health Studies. In: IZA Discussion Paper (3989).

Hank, K. / Jürges, H. / Schaan, B. 2009a. Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe. Methoden - Daten - Analysen 3: 97–108.

Hank, K. / Jürges, H. / Schupp, J. / Wagner, G. 2009b. Isometrische Greifkraft und sozialgerontologische Forschung. Zeitschrift für Gerontologie und Geriatrie 42: 117–126.

Jürges, H. 2008. Self-assessed health, reference levels and mortality.Applied Economics 40: 569–582.

Kapteyn, A. 2010. What can we learn from (and about) global aging? Demography 47 (Supplement): S191-S210.

Kapteyn, A. / Smith, J.P. / van Soest, A. 2009. Work Disability, Work, and Jusitfiaktion Bias in Europe and the U.S.. Labor and Population Working Paper (696).

King, G: / Murray, C.J.L. / Salomon, J.A. / Tandon, A. 2004. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. American Political Science Review 98: 191–207.

Kohn, M.L. 1987. Cross-National Research As An Analytic Strategy. American Sociological Review 52: 713–731.

Mackenbach, J. / Avendano, M. / Andersen-Ranberg, K. / Aro, A.R. 2005. Physical Health. S. 82-88 in: A. Boersch-Supan / K.H. Alcser (Hrsg.), Health, Ageing and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Murray, C.J.L. / Tandon, A. / Salomon, J.A. / Mathers, C.D. / Ritu, S. 2002. New Approaches To Enhance Cross-Popultaion Comparability Of Survey Results. S. 421-430 in C.J.L. Murray / J.A. Salomon / C.D. Mathers / A.D. Lopez (Hrsg.), Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.

Penfield, R.D. / Camilli, G. 2007. Differential Item Functioning and Item Bias. In: C.R. Rao / S. Sinharay (Hrsg.), Psychometrics. Amsterdam: Elsevier.

Preston, S.H. / Heuveline, P. / Guillor, M. 2001. Demography. Measuring and Modeling Population Processes. Oxford: Blackwell.

Rabe-Hesketh, S. / Skrondal, A. 2002. Estimating Chopit Models in gllamm. Political Efficacy Example from King et al. Online verfügbar unter: http://www.gllamm.org/examples.html (02.12.2010).

Rabe-Hesketh, S. / Skrondal, A. 2008. Multilevel and Longitudinal Modeling using Stata. 2. Aufl. College Station, Tex.: Stata Press.

Roppelt, S. 2010. Zur Validität des Vergleichs subjektiver Daten: Das Hopit-Modell am Beispiel selbstberichteter Mobilität im europäischen Vergleich. Diplomarbeit, Professur für Bevölkerungswissenschaft, Universität Bamberg.

Rost, J. 2004. Lehrbuch Testtheorie - Testkonstruktion. 2. Psychologie Lehrbuch. Bern: Huber.

Salomon, J.A. / Tandon, A. / Murray, C.J.L. 2004. Comparability of Self Rated Health: Cross Sectional Multi-Country Survey using Anchoring Vignettes. British Medical Journal (BMJ) 328: 258–261.

Sen, A. 2002. Health: Perception versus Observation. British Medical Journal (BMJ) 324: 860–861.

SHARE (2010). SHARE Guide to Release 2.3.1: Waves 1 & 2. Online verfügbar unter: http://www.share-project.org/t3/share/fileadmin/pdf_documentation/SHARE_Release_Guide_2.3.1.pdf (14.12.2010)

Sirven, N., Santos-Eggimann, B. / Spagnoli. J. 2008. Comparability of Health Care: Responsiveness in Europe Using anchoring Vignettes from SHARE.IRDES Working paper (15). Paris.

Soest, A. van / Delany, L. / Harmon, C. / Kapteyn, A. / Smith, J.P. 2007. Validating the Use of Vignettes for Subjective Threshold Scales. In: RAND Working Paper (501).

Tandon, A. / Murray, C.J.L. / Salomon, J.A. / King, G. 2002. Statistical Models for Enhancing Cross-Population Comparability. Genf: World Health Organization.

Vonková, H. / Hullegie, P. 2010. Is the Anchoring Vignettes Method sensitive to the domain and choice of the Vignette? Netspar Discussion Paper (004). Tilburg.

Wendt, C. 2003. Krankenversicherung oder Gesundheitsversorgung? Gesundheitssysteme im Vergleich. Wiesbaden: Westdt. Verl.