SOEP The German Socio-Economic Panel study

976 2018

SOEPpapers
on Multidisciplinary Panel Data Research

SOEP — The German Socio-Economic Panel study at DIW Berlin 976-2018

Wie gut funktioniert das Onomastik-Verfahren? Ein Test am Beispiel des SOEP-Datensatzes

Elisabeth Liebau, Andreas Humpert, Klaus Schneiderheinze

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: soeppapers@diw.de

Wie gut funktioniert das Onomastik-Verfahren? Ein Test am Beispiel des SOEP-Datensatzes

Elisabeth Liebau
SOEP/DIW Berlin
eliebau@diw.de

Andreas Humpert
Klaus Schneiderheinze
Humpert & Schneiderheinze GbR
hs@stichproben.de

Abstract:
In Deutschland kommt dem Onomastik-Verfahren von Humpert und Schneiderheinze für die Ziehung einer umfassenden Stichprobe von Personen mit Migrationshintergrund eine bedeutende Rolle zu. Das Verfahren leitet mit einer gewissen Wahrscheinlichkeit die regionale Herkunft einer Person von ihrem Namen ab. In diesem Beitrag wird anhand verschiedener Gütekriterien der Frage nachgegangen, wie gut das Onomastik-Verfahren funktioniert und ob es möglicherweise bei bestimmten Zuwanderergruppen zu einer verzerrten Datengrundlage führen kann. Der vorliegende Verfahrenstest auf Grundlage des SOEPs hat ergeben, dass das Onomastik-Verfahren insgesamt betrachtet bei 14 Prozent der Fälle eine falsche Zuordnung trifft. Die überwiegende Mehrheit dieser falschen Zuordnungen kommt bei der Gruppe von Personen mit Migrationshintergrund zustande. Bei der korrekten Identifikation einzelner Herkunftsländer weisen Japan (100%), Deutschland bzw. kein Migrationshintergrund (98%) und die Türkei (97%) die höchsten Trefferquoten auf. Bei Herkunftsländern wie der Schweiz (2%) und Österreich (2%) werden die geringsten Trefferquoten gemessen. Durch den Einsatz des Onomastik-Verfahrens wird das Ausmaß des Integrationsfortschrittes für Zuwanderer und ihrer Nachkommen aus Herkunftsländern mit bedeutendem (Spät-)Aussiedlerzustrom unterschätzt. Ursache hierfür dürfte die nicht vollständige Identifizierung der Teilgruppe der (Spät)Aussiedler sein, die insbesondere hinsichtlich Sprache, Identifikation mit Deutschland und sozialer Einbettung besser integriert sind. Bei den anderen untersuchten Ländern wird hingegen die soziale und strukturelle Integrationsdimension durch den Einsatz des Onomastik-Verfahrens überschätzt. Die Suche nach den Gründen für falsche Zuordnungen oder unvollständige Identifizierung von Teilgruppen legt offen, dass für die Hälfte aller Fehler ein vollständig deutschsprachiger Name der Person mit Migrationshintergrund (durch deutschsprachiges Ausland oder deutsche Minderheiten) eine plausible Erklärung darstellt.

Inhalt

Tabellen

Abbildungen

1. Einleitung – Zur Bedeutung des Onomastik-Verfahrens für die Stichprobenziehung von Zuwanderern und deren Nachkommen in Deutschland

Bereits ein Fünftel der Bevölkerung Deutschlands wies im Jahr 2014 einen Migrationshintergrund auf (Statistisches Bundesamt 2016). Zusätzlich haben die Zuwanderung und der Wanderungsüberschuss ab dem Jahr 2015 deutlich zugenommen (Angenendt et al. 2017), so dass man auch zukünftig von einem hohen Bevölkerungsanteil mit Migrationshintergrund in Deutschland ausgehen kann. Nach wie vor werden in verschiedenen gesellschaftlichen Bereichen Unterschiede zwischen der Mehrheitsbevölkerung und Personen mit Migrationshintergrund festgestellt (siehe z.B. Bundesregierung 2016), für deren Erforschung geeignete Datengrundlagen bereitstehen sollten. Da sich mit Hilfe der amtlichen Statistik nicht alle interessierenden Teilgruppen der Population mit Migrationshintergrund identifizieren sowie relevante Fragestellungen adressieren lassen, besteht auch weiterhin Bedarf an Umfragedaten zu dieser spezifischen Bevölkerungsgruppe. Die Bereitstellung möglichst effizienter, kostengünstiger und abbildungstreuer Stichproben ist deshalb nach wie vor bedeutsam. In diesem Zusammenhang kommt insbesondere für die Identifizierung deutscher Staatsangehöriger mit Migrationshintergrund (55 Prozent aller Personen mit Migrationshintergrund nach Statistischem Bundesamt 2016) namenbasierten Verfahren eine bedeutende Rolle zu (Schnell et al. 2013). Namenbasierte Verfahren stellen selbst kein Stichprobenverfahren dar, sondern sind ein Werkzeug zur sprachlichen Analyse von Personennamen. Aufgrund dieser Eigenschaft werden sie zur Unterstützung verschiedener Stichprobenverfahren eingesetzt.

Im Gegensatz zu skandinavischen Ländern[1] ist die Stichprobenziehung von Personen mit Migrationshintergrund in Deutschland alles andere als einfach. Einen guten Überblick zu Auswahlrahmen und Verfahren in Deutschland vermittelt Salentin (2014). Die entscheidende Schwierigkeit besteht in den Identifizierungsmöglichkeiten der interessierenden Zielpopulation in den zur Verfügung stehenden Auswahlrahmen, die sowohl die Qualität als auch die Kosten der Stichprobenbildung beeinflussen.

Zum einen können vorhandene Merkmale in den Auswahlgrundlagen (wie z.B. das Merkmal Staatsangehörigkeit oder der Geburtsort im Einwohnermeldeamt) genutzt werden. Jedoch stehen oftmals Merkmale wie die Nationalität, das Geburtsland oder auch das Zuzugsjahr nach Deutschland für die Auswahlgrundlage nicht zur Verfügung, sind nicht in ihrem gesamten Umfang zugänglich oder reichen auch nicht immer für die vollständige Identifizierung der interessierenden Population aus. Für den Fall, dass die für die Stichprobenziehung benötigten Merkmale in den jeweiligen

[1] In skandinavischen Ländern verfügt jeder Einwohner über eine Identifikationsnummer, die es unter anderem ermöglicht, Informationen von Kindern, Eltern und Großeltern zu z.B. Staatsangehörigkeit oder Geburtsland zusammenzuspielen (siehe für Dänemark Thygesen et al: 2011). Damit kann bis in die dritte Generation hinein der Migrationshintergrund einer Person umfassend in Registerdaten ermittelt und diese Information zur Stichprobenziehung genutzt werden.

Auswahlgrundlagen gar nicht oder nicht vollständig zur Verfügung stehen, können diese zum anderen sowohl über das Screening-Verfahren als auch über namenbasierte Verfahren grundsätzlich oder ergänzend erhoben werden.

Beim Screening-Verfahren werden die relevanten Merkmale für die Bestimmung des Migrationshintergrundes direkt erfragt.[2] Bei namenbasierten Verfahren wird hingegen der Migrationshintergrund mit einer gewissen Trefferwahrscheinlichkeit vom Namen einer Person abgeleitet.

Das Screening-Verfahren hat den großen Nachteil, dass nur von befragungsbereiten Personen, die eine selektive Gruppe darstellen können (Groves 2006), die relevanten Merkmale erhoben werden können. Darüber hinaus entstehen in Abhängigkeit der genauen Art der Kontaktdaten (Telefonnummer oder Wohnanschrift) schon für die Ermittlung der relevanten Merkmale für die Stichprobenziehung, insbesondere bei anzunehmender geringer Inzidenz der spezifischen Zielgruppe wie bspw. in Deutschland geborene Kinder von Zuwanderern aus Vietnam, enorme Kosten.

Namenbasierte Verfahren können unabhängig von der individuellen Befragungsbereitschaft auch für eine große Anzahl an Fällen relativ preiswert, aber eben nur mit einer gewissen Trefferwahrscheinlichkeit, den Migrationshintergrund eines Namensträgers bestimmen. Grundsätzlich bestehen bei namenbasierten Verfahren zwei Fehlermöglichkeiten. Bei falsch positiven Fällen wird ein spezifischer Migrationshintergrund seitens des Verfahrens zugeschrieben, obwohl dieser nicht besteht. Bei falsch negativen Fällen wird hingegen der spezifisch interessierende Migrationshintergrund vom namenbasierten Verfahren nicht erkannt. Während der erste Fehler lediglich die Brutto- und folglich die Nettostichprobe um die falschen Fälle verringert, und dabei Screeningkosten erzeugt (die falschen Fälle müssen im Erhebungsverlauf ermittelt und ausgeschlossen werden), reduziert der zweite Fehler die Auswahlgrundlage um dazugehörige Elemente. Dies kann ein Problem sein, wenn die Anzahl der fehlenden Elemente groß ist und/oder sich die fehlenden Elemente von den enthaltenen systematisch unterscheiden.

Für die Ziehung von repräsentativen Migrantenstichproben stehen in den meisten Auswahlgrundlagen nicht alle benötigten Merkmale zur Verfügung. Hier ist insbesondere an das Merkmal des Geburtslandes der Eltern, welches für die Identifizierung der 2. Zuwanderergeneration unabdingbar ist, zu denken. Da das Screening-Verfahren offensichtlich selektiv und kostenintensiv ist, kommt namenbasierten Verfahren für die Stichprobenziehung von Personen mit Migrationshintergrund in Deutschland potentiell eine tragende Rolle zu. Nach Schnell et al. hat sich in Deutschland das von Humpert und Schneiderheinze (2000) entwickelte Onomastik-Verfahren als Standardverfahren für sozialwissenschaftliche Stichprobenziehungen bei Migranten etabliert (Schnell 2011: 290). Zum tatsächlichen Funktionieren des Onomastik-Verfahrens ist jedoch bislang nur wenig bekannt (Leicht et al. 2005, Liebau 2011, Kruse/Dollmann 2017[3]), weshalb in diesem Beitrag die Güte dieses spezifischen

[2] So wurde beispielsweise im Rahmen von bevölkerungsrepräsentativen Mehrthemenumfragen (face-to-face wie telefonischen Busumfragen beim Feldinstitut Infratest) mit Hilfe der dort eingesetzten Screening-Frage: „Gibt es unter den erwachsenen Haushaltsmitgliedern in diesem Haushalt jemanden, der vor zehn Jahren – also im Jahr 1984 – noch nicht in der Bundesrepublik gelebt hat, sondern im Ausland oder in der ehemaligen DDR?“ die interessierende Zielpopulation für das SOEP-Teilsample D ermittelt (Schupp/Wagner 1995: 17).

[3] Leicht et al. stellen für die Zielpopulation der selbständigen Aussiedler fest, dass gut die Hälfte der seitens des Onomastik-Verfahrens als Aussiedler gekennzeichneten Personen nicht dieser ethnischen Gruppe

namenbasierten Verfahrens auf der Datengrundlage des Sozio-oekonomischen Panels systematischer untersucht werden soll.

Kern der Analyse ist dabei die Fragestellung, ob und inwieweit dem Verfahren eine hinreichend sichere Identifizierung von Personen mit und ohne Migrationshintergrund und damit der Sicherstellung einer möglichst abbildungstreuen Stichprobe gelingt. Dafür wird zunächst das Onomastik-Verfahren im Detail beschrieben (Abschnitt 2). In Abschnitt 3 wird das Onomastik-Verfahren hinsichtlich verschiedener Gütemaße überprüft. Hier steht zunächst der Grad der Leistungsfähigkeit für Personen mit und ohne Migrationshintergrund im Mittelpunkt. Die Effizienz und das Selektivitätspotenzial des Verfahrens werden anschließend für eine Vielzahl an Herkunftslandgruppen überprüft. Der Forschungsstand beschränkt sich bislang entweder auf das Gütemaß Effizienz (Leicht et al. 2005) oder bei komplexeren Gütemessungen auf sehr wenige spezifische Herkunftslandgruppen (Aussiedler und jüdische Kontingentflüchtlinge bei Liebau 2011) oder eine sehr spezifische Altersgruppe (14-Jährige bei Kruse/Dollmann 2017). Auch wird in der vorliegenden Analyse für eine spezifische Auswahl an Herkunftslandgruppen dem tatsächlichen Ausmaß von Verzerrungen auf inhaltliche Fragestellungen nachgegangen. Neben sozio-demografischen Merkmalen wird dabei ein breites Spektrum an Integrationsindikatoren betrachtet. Ebenso werden die Gründe für Fehlzuweisungen durch das Verfahren beschrieben und deren jeweilige Bedeutung eingeschätzt. Zum Abschluss werden die Ergebnisse bilanziert, Analysebeschränkungen beschrieben und sich anschließende Forschungsfragen formuliert (Abschnitt 4).

2. Eigenschaften des Onomastik-Verfahrens

Die Beurteilung der Güte eines Verfahrens auf Grundlage von empirischen Testergebnissen setzt zumindest grobe Kenntnisse zur spezifischen Konstruktion des Verfahrens voraus. Deshalb sollen in diesem Abschnitt die wesentlichsten Eigenschaften des Verfahrens in knapper Form dargestellt werden.

angehören (Leicht et al. 2005: 61). Liebau betrachtet bei Ihrem Gütetest neben Aussiedlern auch jüdische Kontingentflüchtlinge aus der ehemaligen Sowjetunion. Ca. 80 Prozent der seitens des Onomastik-Verfahrens als Aussiedler oder jüdische Kontingentflüchtlinge gekennzeichneten Haushalte gehören diesen beiden ethnischen Gruppen tatsächlich an, jedoch werden auch über 80 Prozent der Haushalte, die seitens des Onomastik-Verfahrens anderen Herkunftslandgruppen aus der ehemaligen Sowjetunion zugeordnet worden sind, durch Aussiedler oder jüdische Kontingentflüchtlinge gestellt (Liebau 2011: 98). D.h. die spezifischen Herkunftslandgruppen ‚Juden aus der früheren SU' und ‚Aussiedler aus der früheren SU' enthalten jeweils nur einen Teil der jüdischen Kontingentflüchtlinge und Aussiedler. Für die Merkmale Zuzugsjahr und Arbeitslosenrate konnten für die seitens des Onomastik-Verfahrens identifizierten Aussiedler im Vergleich zu den nicht identifizierten Aussiedlern signifikante Unterschiede festgestellt werden (ebd. 99). Kruse und Dollmann untersuchen für sechs spezifische Herkunftsländer/-regionen (Türkei, ehemalige Sowjetunion, Polen, ehemaliges Jugoslawien, westliche Länder und nicht westliche Länder) für die Population der 14-Jährigen die Effizienz und das Selektivitätspotential des Onomastik-Verfahrens. Dabei stellen sie Herkunftsland- und Generationenunterschiede fest (Kruse/Dollmann 2017: 442). So funktioniert das Onomastik-Verfahren für Zuwanderer und deren Nachkommen aus der Türkei und Staaten des ehemaligen Jugoslawiens am besten, auch wird die erste Generation treffsicherer identifiziert.

Personennamen haben sich in Europa seit dem Spätmittelalter zur Gestalt einer Kombination von Vorname und Nachname entwickelt. Diese Form ist später auch in Ländern mit anderen Namenstraditionen (z.B. muslimische Staaten, Japan) übernommen und gesetzlich geregelt worden: z.B. in Japan 1870 (verpflichtend 1875), in der Türkei 1935, in Marokko 1950 und in Tunesien 1959.[4]
Vornamen sind stark durch die Ausbreitung der Religionen geprägt. Hingegen spiegelt sich in Nachnamen die Sprache ihrer Entstehungsumgebung wider. In den meisten Fällen lässt sich deshalb mit hoher Wahrscheinlichkeit vom Namen auf die Sprache schließen.
In der Namenforschung (Onomastik) wird deshalb vorrangig auf sprachwissenschaftlicher Grundlage die Bedeutung von Namen untersucht. Diesen Umstand nutzt das sogenannte Onomastik-Verfahren von Humpert & Schneiderheinze (Humpert/Schneiderheinze 2000), indem es von sprachanalytischen Erkenntnissen der Namenforschung auf spezifische Regionen bzw. Sprachräume schließt. Ein Personenname kann also mit Hilfe wissenschaftlicher Erkenntnisse der Namenforschung Rückschlüsse auf den Migrationshintergrund des Namenträgers ermöglichen.
Die Namenforschung hat im Laufe der Zeit neben zahlreichen Arbeiten zu einzelnen Namen auch umfangreiche Verzeichnisse von Namen hervorgebracht. Solche Nachnamenlexika existieren für viele europäische und in etwas geringerem Umfang für nicht-europäische Sprachen.[5] Lexika mit einem Anspruch auf eine gewisse Vollständigkeit existieren für die Niederlande[6], die Schweiz[7] und Polen[8]. In Deutschland entsteht derzeit das auf rund 200.000 Namen angelegte Digitale Familiennamenwörterbuch Deutschlands (DFD), das aber keine Vollständigkeit anstrebt.[9]
Die Namenforschung bietet jedoch nicht nur Informationen zu sprachlichen Wurzeln konkreter Namen, sondern liefert neben den Namenlexika auch Erkenntnisse zu Strukturen von Namen in verschiedenen Sprachen. Daher können auch nicht verzeichnete Nachnamen, die den größeren Teil der vorkommenden Namen ausmachen, sehr häufig einer Sprache zugeordnet werden. Wo die Grenzen der Onomastik erreicht sind, werden Namen durch statistische Verfahren (z.B. Clusteranalyse und Häufigkeitsverteilungen) untersucht und können so häufig sprachlich geklärt werden. Statistische Häufigkeiten müssen auch bei onomastisch geklärten Namen berücksichtigt werden, weil durch Migration der Menschen auch ihre Namen wandern. So sind beispielsweise viele englische Namen heute in den USA wesentlich häufiger als in Großbritannien vorhanden. Die Migrationsgeschichte ist also bei der sprachwissenschaftlichen Analyse von Personennamen von essentieller Bedeutung.

---

4 In den arabischen Staaten, in denen die europäische Namensform (Vorname Nachname) eingeführt worden ist, wurden die neuen Namen meist aus den „alten“ arabischen Namen gebildet, die aus bis zu fünf Teilen bestehen können. Die neuen Namen wurden dabei mehr oder weniger willkürlich gebildet. Unter anderem kann jeder Vorname auch ein Nachname sein.

5 Eine gekürzte Version einer Bibliographie mit über 1.000 Quellen kann auf www.stichproben.de heruntergeladen werden.

6 Nederlands Repertorium van Familienamen. 14 Bde. Assen/Amsterdam 1963-88.

7 Familiennamenbuch der Schweiz. Répertoire des noms familles Suisses. 1. Aufl., 2 Bde. Zürich 1940, 3. Aufl., 3 Bde. Zürich 1989.

8 Rymut, Kazimierz (1993): Słownik nazwisk współcześnie w Polsce używanych [Wörterbuch der in Polen gebräuchlichen Familiennamen]. 10 Bde. Kraków: Instytut Języka Polskiego. Es werden auch nicht polnische Namen mit Häufigkeitsangaben verzeichnet, die von zugewanderten Bevölkerungsgruppen getragen werden.

9 Es sollen alle (auch nicht deutsche) Familiennamen erfasst werden, die im Telefonbuch (Telekom, Stand 2005) wenigstens zehnmal vorkommen. Siehe: www.namenforschung.net/dfd/projektvorstellung. Dadurch bleibt ein großer Teil der in Deutschland existierenden Nachnamen unberücksichtigt.

Vor diesem Hintergrund ist das sogenannte Onomastik-Verfahren von Humpert & Schneiderheinze mit den folgenden Eigenschaften kontinuierlich seit 1999 fortentwickelt worden:

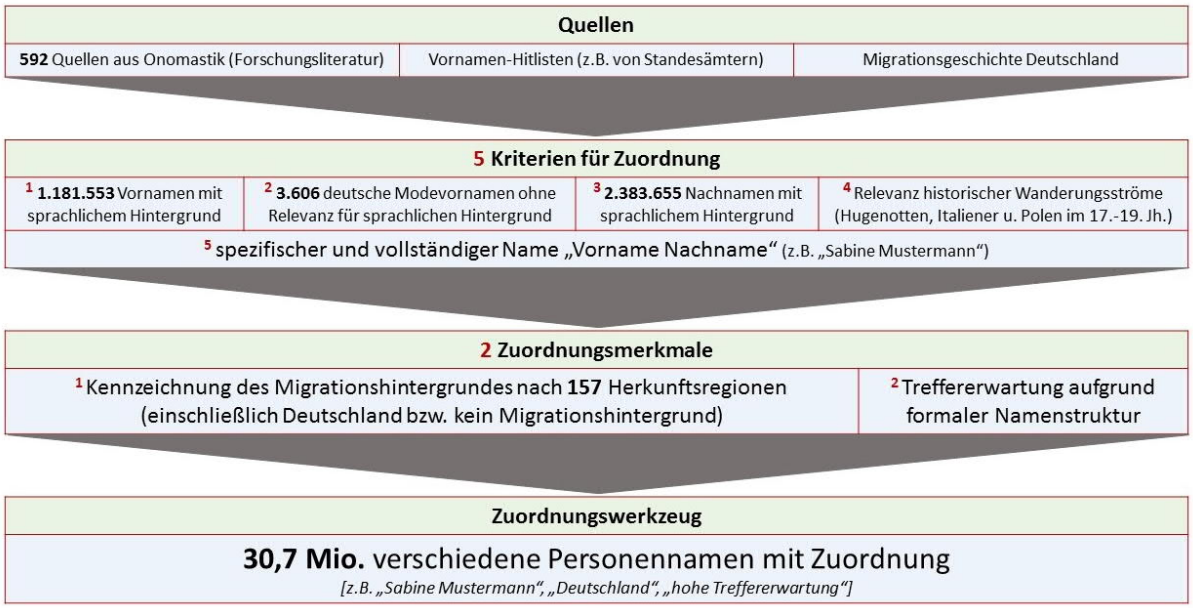
- Anwendungsregion: Die Anwendung ist auf Namen von Personen fokussiert, die in Deutschland leben.
- Sprachliche Analyse mittels wissenschaftlicher Namenforschung (Onomastik): Mit Hilfe methodischer Erkenntnisse und wissenschaftlicher Literatur (zurzeit 592 Quellen) aus der Onomastik wird die Grundlage für die sprachliche Analyse von Vor- und Nachnamen geschaffen. Darüber hinaus werden zusätzliche Quellen mit Namen genutzt, die verlässliche Hinweise auf den sprachlichen Hintergrund geben.
- Berücksichtigung historischer Wanderungsbewegungen: In Deutschland leben durch frühere Wanderungsbewegungen (z.B. Zuzug von Hugenotten im 17. Jahrhundert und von Polen in industrielle Ballungsräume des kaiserlichen Deutschlands) viele Personen mit fremdsprachigem Nachnamen, denen von der Migrationsforschung und der amtlichen Statistik kein Migrationshintergrund zugeordnet wird.
- Erfassung von Fehlschreibungen und Transkriptionsvarianten: In jeder Datenbank mit Personennamen sind Fehler enthalten. Für eine möglichst vollständige Zuordnung ist deshalb die Erfassung von Fehlschreibungen und Transkriptionsvarianten unerlässlich. Durch den Wegfall diakritischer Zeichen, unterschiedliche Transkriptionssysteme oder unterschiedliche Zeichensätze kommen ausländische Namen besonders häufig in mehreren Schreibvarianten vor. Diese Schreibvarianten sind in einer zentralen Kodierungsdatei des Onomastik-Verfahrens ebenfalls enthalten.
- Berücksichtigung möglichst vieler Sprachen: Die Erfassung möglichst vieler Sprachen dient nicht nur dem Zweck einer maximalen Zuordnungsrate, sondern auch der Vermeidung von Fehlern. Die Erfassung zusätzlicher Sprachen vermindert auch Fehlzuordnungen, die auf zufällige Namensgleichheit in zwei oder mehr Sprachen bzw. Herkunftsregionen beruhen.
- Kennzeichnung sprachlicher Mischformen: Ein gewisser Teil von Personennamen enthält Vor- und Nachnamen aus unterschiedlichen Sprachen. Außerdem ist eine Minderheit (ca. 4-5%) von Nachnamen in verschiedenen Staaten gebräuchlich. In diesen Fällen steigt die Wahrscheinlichkeit für falsche Zuordnungen. Für die Erfassung und Kennzeichnung gemischtsprachiger Namen wurde deshalb eine grobe Kategorisierung der sogenannten Treffererwartung („hoch“, „mittel“ oder „niedrig“) eingeführt.
- Gegenstand der endgültigen Zuordnung: Das Verfahren analysiert nur vollständige Namen, die aus Kombinationen von Vor- und Nachnamen bestehen.
- Zentrale Zuordnungsdatei: In einer zentralen Datenbank werden diese vollständigen Personennamen mit ihrem Schlüssel (HSM) und einer Treffererwartung (HS-Typ) gesammelt. Zurzeit (Mai 2018) wird nach 157 verschiedenen Staaten- und Regionalschlüsseln differenziert. In dieser Datenbank sind 30.725.355 verschiedene[10] Personennamen mit Zuordnung zum Herkunftsstaat (einschließlich Deutschland) enthalten. Darin nicht befindliche Kombinationen aus Vor- und Nachnamen werden über ein halbautomatisches Verfahren analysiert.

Der Ablauf des Entscheidungsprozesses im Rahmen des Onomastik-Verfahrens soll mit der folgenden Abbildung veranschaulicht werden:

[10] Verschieden bedeutet in diesem Zusammenhang, dass beispielsweise der Name „Helmut Schmidt“ nur einmal eingetragen ist, obwohl eine Vielzahl von in Deutschland lebenden Personen diesen Namen tragen.

Der Ablauf des Entscheidungsprozesses im Rahmen des Onomastik-Verfahrens soll mit der folgenden Abbildung veranschaulicht werden:

Abbildung 1: Ablauf des Onomastik-Verfahrens



Die Quellen des Verfahrens bestehen aus Ergebnissen der Namenforschung, Mode-Vornamenlisten, migrationsgeschichtlichen Informationen und sonstigen Namenbeständen mit sicherem sprachlichem Bezug. Mit Hilfe dieser Quellen kann jeder einzelne vollständige Personenname anhand von fünf Kriterien einem Herkunftsland bzw. -region zugeordnet werden. Diese fünf Kriterien werden durch den sprachlichen Hintergrund des Vornamens und des Nachnamens gebildet. Außerdem wird die Relevanz bezüglich Vornamenmoden und migrationsgeschichtlicher Besonderheiten (z.B. frühe Zuwanderungswellen) geprüft. Letztlich wird das fünfte Kriterium für eine Zuordnung durch den konkreten Personennamen als Kombination aus Vor- und Nachnamen („Sabine Mustermann“) gebildet.

Auf Grundlage dieser fünf Zuordnungskriterien werden für jeden Personennamen zwei Merkmale gebildet. Dabei handelt es sich um die Herkunftsregion und ein Kennzeichen für die formale Struktur des Namens, der einsprachig sein oder mehrsprachige Teile enthalten kann. Daraus entsteht eine zentrale Zuordnungsdatenbank mit dem Namen (Vor- und Nachname) und der zugewiesenen Herkunftsregion. Zurzeit (Mai 2018) befinden sich darin 30,7 Mio. verschiedene Namen, die von ca. 60 bis 75 Mio. in Deutschland lebenden Personen getragen werden. Namen ohne Eintrag in dieser zentralen Zuordnungsdatenbank werden mit Hilfe der Quellen, die 1,1 Mio. einzelne Vornamen bzw. Vornamenkombinationen und 2,3 Mio. Nachnamen enthalten, mittels halbautomatischer Verfahren und Einzelsichtung analysiert.

3. Empirische Überprüfung des Onomastik-Verfahrens

3.1 Datengrundlage

Für einen möglichst verlässlichen Test des Onomastik-Verfahrens wird eine Datenbasis benötigt, die mindestens vier Bedingungen erfüllt. Erstens muss die Klärung des tatsächlichen herkunftslandspezifischen Migrationshintergrundes einer Person möglich sein. Zweitens sollen darin die relevanten Migrantengruppen in großer Fallzahl enthalten sein. Drittens soll die Datenbasis noch weitere sozioökonomische bzw. integrationsrelevante Erhebungsmerkmale bereitstellen können, um eine mögliche Verzerrung, die durch den Einsatz des Onomastik-Verfahrens hervorgerufen wird, erkennen zu können. Viertens darf die Datenbasis aufgrund fehlender Finanzmittel nicht eigens für den Test erhoben werden, sondern muss bereits vorhanden und kostenfrei nutzbar sein. Alle vier Bedingungen kann das Sozioökonomische Panel (SOEP)[11] erfüllen. Für den Test wurde diesem Datensatz deshalb nur noch die Herkunftslandzuweisung durch das Onomastik-Verfahren hinzugefügt. Das SOEP ist eine seit 1984 in Deutschland laufende Längsschnittbefragung von in Deutschland lebenden Personen, welche mit seinen drei großen Zuwandererstichproben[12] das deutsche Migrationsgeschehen bis zum Jahr 2013 weitgehend abdeckt. Die Teilpopulation jener Personen mit Migrationshintergrund, die mit Hilfe des Onomastik-Verfahrens in der M1 Stichprobe ermittelt wurden (Kroh et al. 2015) bleibt von der Untersuchung ausgenommen.[13] Der SOEP-Datensatz wurde darüber hinaus für die nachfolgenden Analysen auf jene Fälle beschränkt, für die bezüglich Herkunftsland bzw. -region ein spezifischer[14] Migrationshintergrund sicher bestimmt werden konnte. Dies sind aktiv befragte SOEP-Teilnehmer der Erhebungsjahre 1984 bis 2013, die

- um der Herkunftslandgruppe ‚ohne Migrationshintergrund' zugewiesen zu werden, folgende Merkmale aufweisen müssen: Auf allen sieben Variablen - erste Staatsangehörigkeit im SOEP-Beobachtungsfenster (e SA), letzte Staatsangehörigkeit im SOEP-Beobachtungsfenster[15] (l SA), Geburtsland (G), Geburtsland der Mutter (G M),

[11] SOEP.v30

[12] Erstens Teilstichprobe B, die 1984 gezogen wurde und die Zuwanderung bis 1984 in die damalige BRD abdeckt. Zweitens Teilstichprobe D, welche 1994/95 gezogen wurde und die Zuwanderung von 1984 bis 1994 nach Deutschland umfasst und drittens Teilstichprobe M1, die 2013 gezogen wurde und die Zuwanderung seit 1995 nach Deutschland im Stichprobensystem des SOEP ergänzt (Liebau/Tucci 2015).

[13] Auch in den Teilstichproben I und J des SOEP kam für eine überproportionale Ziehung von Personen mit Migrationshintergrund das Onomastik-Verfahren zum Einsatz (für Details siehe Liebau/Tucci 2015). Aus folgenden Gründen wurden diese Fälle nicht ausgeschlossen: Das Onomastik-Verfahren beruhte bei den Teilsamples I und J auf Klingelschildnamen, also in der Regel nur Nachnamen und hat lediglich zwischen zwei Gruppen, Personen ohne und mit Migrationshintergrund unterschieden. Die mit Migrationshintergrund gekennzeichneten Fälle wurden überproportional in die Bruttostichprobe aufgenommen, jene als ohne Migrationshintergrund gekennzeichneten jedoch nicht ausgeschlossen.

[14] Zu geringe Fallzahlen sowie abweichende Gruppenabgrenzungen zwischen SOEP und der HS-Klassifizierung erforderten die Zusammenfassung von mehreren Staaten zu Ländergruppen bzw. Regionen (Details siehe Tabelle 11 im Appendix).

[15] Wenn die Person nur ein Jahr lang am SOEP teilgenommen hat, dann sind die erste und letzte beobachtete Staatsangehörigkeit automatisch identisch.

Geburtsland des Vaters (G V), Staatsangehörigkeit der Mutter (SA M), Staatsangehörigkeit des Vaters (SA V) - muss der Wert Deutsch/Deutschland vorliegen

- um einem/einer spezifischen ausländischen Herkunftsland bzw. -region zugewiesen zu werden, folgende Merkmale aufweisen müssen: Die Kombination der Werte folgender fünf Variablen - erste und letzte Staatsangehörigkeit im SOEP-Beobachtungsfenster, Geburtsland, Geburtsland der Mutter, Geburtsland des Vaters - verweist eindeutig auf einen spezifischen Migrationshintergrund.

Tabelle 1: Übersicht zur Zuweisung des spezifischen Migrationshintergrundes bzw. Ausschluss eines Falles aus der SOEP-Grundgesamtheit mit Beispielen und Fallzahlen

	Zuordnungsbeispiele							
	e SA	l SA	G	G M	G V	SA M	SA V	N
1. Fälle ohne Migrationshintergrund mit fehlenden Angaben	d	d	d	-	-	d	d	37.849
2. Fälle mit unspezifischen Angaben zum Migrationshintergrund (z.B. „Ausland“)	d	d	d	d	-	a	-	402
3. Fälle mit Angaben zum Migrationshintergrund, die auf kein spezifisches Herkunftsland/-region hinweisen	t	d	d	t	g			281
4. Fälle ohne Migrationshintergrund	d	d	d	d	d	d	d	10.842
5. Fälle mit eindeutig bestimmbaren Migrationshintergrund	t	d	d	-	t			12.335
Aktiv befragte SOEP-Teilnehmer								61.709

t=Türkei, g=Griechenland, a=Ausland;
Personen, die vor 1949 zugewandert sind, werden der Gruppe ohne Migrationshintergrund zugeordnet

Nur die Fälle der 4. und 5. Gruppe aus Tabelle 1 bilden die Grundgesamtheit der folgenden Analysen. Obwohl alle vorhandenen Angaben der 1. Gruppe auf keinen Migrationshintergrund deuten, könnte dieser aus den fehlenden Angaben hervorgehen, weshalb - um ganz sicher zu gehen - solche Fälle ausgeschlossen sind. Dies gilt auch für die 2. Gruppe, weil bei diesen SOEP-Befragten nur allgemein ein Migrationshintergrund ohne konkretes Herkunftsland eingetragen ist. Für die SOEP-Befragten der 3. Gruppe ist ebenfalls keine eindeutige Zuordnung zu einem Herkunftsland möglich, weil sie bezüglich der sieben Indikatoren einen Migrationshintergrund von zwei unterschiedlichen ausländischen Herkunftsstaaten/-regionen aufweist. Wir wollen hier keine willkürliche Zuweisung des Herkunftslandes vornehmen, weshalb auch diese Fälle ausgeschlossen sind.[16] Dadurch bleiben viele Befragte ausgeschlossen, die vermutlich ohne Migrationshintergrund sind. Deshalb ist hinzunehmen, dass in der verbleibenden Analysepopulation der Anteil von Personen mit Migrationshintergrund im Vergleich z.B. mit dem Mikrozensus überhöht ausfällt (12.335 von 23.177 = 53,2% zu Bevölkerung in Deutschland nach Mikrozensus 2015 17,1 Mio. von 81,4 Mio. = 21,0%). Eine Übertragung der Analyseergebnisse auf die Bevölkerung in Deutschland, insbesondere die Messwerte zum Grad der Leistungsfähigkeit des Verfahrens (siehe Tabelle 3), erfordert also mindestens eine entsprechende Gewichtung nach vorhandenem und nicht vorhandenem Migrationshintergrund.

[16] Diese Gruppe stellt einen Spezialfall dar, der gesondert untersucht werden sollte.

3.2 Gütemaße

Beim nachfolgenden Test soll die Güte des Onomastik-Verfahrens bestimmt werden, welche sich aus verschiedenen Einzelmaßen zusammensetzt. Eines dieser Gütemaße ist der Grad der Leistungsfähigkeit, welcher darauf beruht, dass jede Zuordnung des Onomastik-Verfahrens als falsch, ungenau oder richtig klassifiziert wird. Fehlende Zuordnungen werden ebenfalls erfasst. Die Ungenauigkeit wird nach drei verschiedenen Stärken beurteilt. Dies ergibt sich zwangsläufig aus sprachtechnischen Gründen, die statt einer Zuweisung zu einzelnen Staaten nur die Kennzeichnung einer Staatengruppe (z.B. „Skandinavien“) zulässt. Das Verfahren führt hier also zwangsläufig zu ungenauen Zuordnungen, weil der jeweilige SOEP-Befragte von Seiten des DIW immer einem konkreten Staat (z.B. „Schweden“) zugeordnet ist. Der Grad der Ungenauigkeit hängt nun von der Ausdehnung des regionalen Raumes ab, der durch die Onomastik-Kategorie definiert ist. Je größer die regionale Ausdehnung ist, desto größer ist auch die Ungenauigkeit der Onomastik-Zuordnung.
Demnach vermittelt dieses Gütemaß eine Qualitätseinschätzung, die unabhängig von der Richtung bzw. Perspektive auf die Identifikation einer bestimmten Zielgruppe (z.B. Personen mit [türkischem] Migrationshintergrund) ist.

Tabelle 2: Grad der Leistungsfähigkeit

Grad	Bedeutung
1	Treffer bzw. richtige Zuordnung
2	gering ungenaue Zuordnung (nur Region (z.B. „Skandinavien“)
3	ungenaue Zuordnung (nur Obergruppe (z.B. „Muslimische Staaten“)
4	sehr ungenaue Zuordnung (nur „Migrationshintergrund“)
5	keine Zuordnung
6	falsche Zuordnung

Demgegenüber ist eine Fehleranalyse, die in der Logik „falsch positiver“ und „falsch negativer“ Fälle durchgeführt wird, immer auf eine bestimmte Zielgruppe mit spezifischen Migrationshintergrund fokussiert. Die „falsch positiven“ Fälle sind dann solche Personen, denen das Onomastik-Verfahren einen spezifischen Migrationshintergrund zuweist, obwohl sie diesen tatsächlich gar nicht aufweisen. Deshalb sprechen wir bei dieser Fehlerart auch von der Effektivität des Verfahrens. Bei „falsch negativen“ Fällen wird hingegen der spezifisch interessierende Migrationshintergrund vom Verfahren nicht erkannt. Diese Fehlerart stellt ein Potential für Verzerrungen durch das Verfahren dar und kennzeichnet deshalb das Selektivitätspotential des Verfahrens.
Ziel ist es nun zunächst, den Umfang des Grads der Leistungsfähigkeit des Onomastik-Verfahrens sowie der zwei Fehlerformen zu ermitteln. Für die Bestimmung der Güte des Verfahrens ist die Fehlerart „falsch negativ“ am bedeutsamsten, weil im Falle größerer Abweichungen zwischen Grundgesamtheit und den Identifizierten hinsichtlich relevanter Merkmale die Gefahr verzerrter Stichproben bestehen könnte. In einem weiteren Schritt soll deshalb auch der möglichen Selektion bestimmter Sozialgruppen durch das Onomastik-Verfahren anhand zentraler Integrationsindikatoren nachgegangen werden.

3.3 Testergebnisse

Hinsichtlich der zentralen Frage der Leistungsfähigkeit des Onomastik-Verfahrens beinhaltet die folgende Tabelle 3 die Verteilungen für die Gesamtheit der Befragten und differenziert nach Migrationshintergrund, Generationenstatus und Geschlecht. Insgesamt ist zunächst festzustellen, dass nur in 0,2 Prozent aller Fälle keine Zuordnung durch das Verfahren erfolgt ist. Etwa 14 Prozent werden falsch und knapp über 4 Prozent werden vom Verfahren ungenau zugeordnet.

Tabelle 3: Häufigkeiten über den Grad der Leistungsfähigkeit

Grad der Leistungsfähigkeit des Verfahrens	Maß*	alle	Migrationshintergrund nach Angaben im SOEP		Generationen-status		Geschlecht	
			ja	nein	1.	2.+	m	w
1 Richtig	Prozent	82	66	98	67	61	83	79
	Absolut	18.801	8.175	10.626	6.573	1.601	9.291	9.510
2 Gering ungenau	Prozent	2	3	0	4	1	2	2
	Absolut	394	394	0	355	38	199	195
3 Ungenau	Prozent	1	3	0	3	1	2	1
	Absolut	313	313	0	261	37	181	132
4 Sehr ungenau	Prozent	1	3	0	3	2	1	2
	Absolut	313	313	0	261	52	120	193
5 Keine Zuordnung	Prozent	0	0	0	0	0	0	0
	Absolut	41	24	17	23	1	14	27
6 Falsch	Prozent	14	25	2	23	35	12	16
	Absolut	3.315	3.116	199	2.192	920	1.400	1.915
Total	Prozent	100	100	100	100	100	100	100
	Absolut	23.177	12.335	10.842	9.680	2.649	11.205	11.972

* Die Prozentwerte sind auf ganze Zahlen gerundet.

Die differenzierte Betrachtung nach Migrationshintergrund zeigt, dass die Leistungsfähigkeit des Verfahrens bei Personen ohne Migrationshintergrund mit fast 98 Prozent richtigen Zuordnungen sehr hoch ist. Die Identifikation innerhalb der Personen mit Migrationshintergrund führt hingegen in jedem vierten Fall zu einer falschen Zuordnung. Von allen falschen Zuordnung (3.315) sind 2.617 (79%) darauf zurückführen, dass eine Person mit Migrationshintergrund der Gruppe ohne Migrationshintergrund zugeordnet wird. Die Verfahrensleistung variiert ebenso hinsichtlich der Generationenzugehörigkeit und des Geschlechts der Namensträger. Bei Personen ohne eigener Migrationserfahrung jedoch mit Migrationshintergrund funktioniert das Onomastik-Verfahren deutlich schlechter als bei Personen, die selbst zugewandert sind (35% zu 23% falsche Zuordnungen). Dies steht im Einklang zu den Befunden von Kruse/Dollmann 2017. Bei weiblichen SOEP-Befragten funktioniert das Onomastik-Verfahren etwas schlechter, der Anteil der Falschzuordnungen liegt mit 16 Prozent 4 Prozentpunkte über jenem der Männer.

Wird nach der groben Zuordnung mit und ohne Migrationshintergrund unterschieden, siehe Tabelle 4, so schreibt das Onomastik-Verfahren 1 Prozent aller zugeordneten Fälle einen Migrationshintergrund zu, obwohl dieser nach den Angaben im SOEP nicht besteht. Diese machen 2 Prozent aller Fälle ohne Migrationshintergrund nach dem SOEP aus. Bei 11 Prozent aller zugeordneten Fälle wird hingegen der

bestehende Migrationshintergrund vom Onomastik-Verfahren nicht erkannt, dies ist ein gutes Fünftel aller Personen mit Migrationshintergrund im SOEP. Das Verfahren übersieht demzufolge eher einen bestehenden Migrationshintergrund, als dass es diesen fälschlicherweise zuschreibt. Insgesamt rund 88 Prozent aller Befragten werden bei der groben Differenzierung nach mit bzw. ohne Migrationshintergrund vom Verfahren richtig zugeordnet.

Tabelle 4: Güte nach grober Einteilung mit und ohne Migrationshintergrund

			Migrationshintergrund nach Onomastik-Verfahren		
			nein	ja	Total
Migrationshintergrund nach Angaben im SOEP	nein	N	10.626	199	10.825
		Zellenprozente	46	1	47
		Zeilenprozente	98	2	100
	ja	N	2.617	9.694	12.311
		Zellenprozente	11	42	53
		Zeilenprozente	21	79	100
		Total	13.243	9.893	23.136

Tabelle 5 gibt nun Auskunft darüber, wie viel Prozent jener Personen, die seitens des Onomastik-Verfahrens ein spezifisches Herkunftsland zugeschrieben bekommen haben, tatsächlich aus diesem Land stammen. Es besteht die Erwartung, dass die Effektivität des Onomastik-Verfahrens insbesondere für solche Herkunftsländer niedrig ausfällt, die entweder zum gleichen Sprachraum gehören (wie die USA und Großbritannien), bei denen starke Wanderungsströme zu einer staatenübergreifenden Verbreitung von Namen einer Sprache geführt haben (wie Ungarn[17]) oder deren Zuwanderer häufig deutschsprachige Namen tragen (Schweiz, Österreich, Ex-Sowjetunion).

Für 22 der 36 untersuchten Herkunftsstaaten bzw. -regionen kann eine hohe Effektivität des Onomastik-Verfahrens von mindestens 70 Prozent beobachtet werden. Darunter sind die wichtigsten Zuwanderungsländer und die Gruppe ohne Migrationshintergrund. Das Verfahren produziert besonders hohe Fehleranteile bei Namen aus Herkunftsstaaten, die sich nicht trennscharf von Sprachräumen unterscheiden lassen (z.B. USA, Großbritannien, Frankreich), die historisch eine starke Wanderung in Nachbarstaaten aufweisen (z.B. Ungarn, Frankreich) und zum deutschsprachigen Ausland zählen (z.B. Schweiz, Österreich).

[17] Beispielsweise verteilen sich die 69 vom Onomastik-Verfahren mit dem ungarischen Migrationshintergrund gekennzeichneten Befragten hinsichtlich ihrer tatsächlichen Herkunftsstaaten auf die Region Österreich, Ex-Tschechoslowakei, Rumänien und Ex-Jugoslawien. Dies kann als Folge der historischen Wanderungsbewegungen im ehemaligen Vielvölkerstaat Österreich-Ungarn gesehen werden.

Tabelle 5: Effektivität des Onomastik-Verfahrens geordnet nach Anteil der Richtigen in Prozent*

Spezifische(s) Herkunftsland/-region nach dem Onomastik-Verfahren	Übereinstimmung mit dem/der spezifischen Herkunftsland/-region im SOEP			
	richtig	ungenau	falsch	N
Korea	100	0	0	12
China	100	0	0	24
Vietnam	100	0	0	33
Türkei	98	0	2	2.761
Griechenland	98	0	2	843
Iran	97	0	3	35
Ex-Jugoslawien**	97	0	3	1.031
Thailand	96	0	4	25
Sub-Sahara**	95	0	5	65
Italien	95	0	5	1.242
Rumänien	94	0	6	174
Ex-Sowjetunion**	93	0	7	911
Spanien	93	0	7	450
Polen	91	0	9	459
Japan	89	0	11	9
Sri Lanka	87	0	13	15
Deutschland	80	0	20	13.243
Finnland	80	0	20	10
Maghreb-Staaten**	78	6	16	50
Bulgarien	77	0	23	22
Benelux	76	0	24	37
Ex-Tschechoslowakei**	72	0	28	29
Arabische Staaten**	67	27	6	51
Indien	67	0	33	27
Portugal	65	0	35	46
Frankreich	54	0	46	74
Mittlerer Osten**	51	11	38	37
Südostasien**	44	11	45	18
USA	42	0	58	24
Ungarn	39	0	61	69
Großbritannien	38	0	62	92
Österreich	31	0	69	16
Dänemark	22	0	78	9
Schweiz	17	0	83	6
Lateinamerika**	16	69	15	202
Albanien	6	92	2	359

* Die Prozentwerte sind auf ganze Zahlen gerundet, ** Vgl. Übersicht zu den einzelnen Ländern einer Herkunftsregion Tabelle 11 im Appendix. Länder mit N<=5 sind ausgeschlossen; richtig=Kategorie 1 beim Grad der Leistungsfähigkeit, ungenau=Kategorien 2 bis 4 beim Grad der Leistungsfähigkeit, falsch=Kategorie 6 beim Grad der Leistungsfähigkeit

In der folgenden Tabelle 6 wird die Treffgenauigkeit des Onomastik-Verfahrens aus anderer Perspektive beleuchtet. Vom tatsächlichen Migrationshintergrund ausgehend (SOEP-Zuordnung) soll der Anteil der durch das Onomastik-Verfahren identifizierten Grundgesamtheit offengelegt werden. Mit sinkendem Anteil der identifizierten Grundgesamtheit steigt das Selektivitätspotential, also die Wahrscheinlichkeit für eine verzerrte Abbildung. Parallel zur Betrachtung der Effektivität ist auch hinsichtlich der Selektivität zu erwarten, dass diese insbesondere bei Herkunftsländern hoch ausfällt, die sich mit anderen Staaten den Sprachraum teilen (z.B. Australien, Kanada, USA, Irland, Benelux und Großbritannien, Lateinamerika, Frankreich). Auch bei deutschsprachigen Herkunftsländern oder Zuwanderergruppen, die als deutsche Minderheit nach Deutschland gekommen sind, müsste das Onomastik-Verfahren nur geringe Anteile mit Identifizierten einer Grundgesamtheit produzieren.

Für 11 der 40 untersuchten Herkunftsländer kann ein niedriges Selektivitätspotential des Onomastik-Verfahrens beobachtet werden. Hierbei wurden mindestens 70 Prozent aus der Gesamtpopulation eines/r spezifischen Herkunftslandes bzw. -region seitens des Onomastik-Verfahrens richtig identifiziert. Darunter sind erneut die Türkei, Griechenland, Italien und die Gruppe ohne Migrationshintergrund.

Tabelle 6: Selektivitätspotential des Onomastik-Verfahrens geordnet nach Anteil der Richtigen in Prozent*

Spezifische(s) Herkunftsland/-region im SOEP	Übereinstimmung mit dem/der spezifischen Herkunftsland/-region nach dem Onomastik-Verfahren			
	richtig	ungenau	falsch	N
Japan	100	0	0	8
Deutschland	98	0	2	10.825
Türkei	97	2	1	2.786
Griechenland	92	3	5	893
Vietnam	92	5	3	36
Italien[2]	90	1	9	1.311
Korea	86	0	14	14
Albanien	81	4	15	26
Thailand	80	10	10	30
China	78	3	19	31
Sri Lanka	72	11	17	18
Spanien	69	23	8	606
Ex-Jugoslawien**	62	25	13	1.613
Indien	60	7	33	30
Finnland	57	14	29	14
Bulgarien	55	13	32	31
Iran	55	38	7	53
Großbritannien[1]	52	8	40	67
Ex-Sowjetunion[2] **	51	3	46	1.648
Sub-Sahara**	48	14	38	128
Maghreb-Staaten**	44	52	4	89
Polen[2]	40	1	59	1.036
Frankreich[1 2]	38	4	58	106
Rumänien[2]	37	6	57	443
Mittlerer Osten**	37	45	18	65
Ungarn[1]	36	5	59	76
Lateinamerika[1] **	34	7	59	97
Irland[1]	29	0	71	7
Arabische Staaten**	23	66	11	149
Benelux[1] **	21	4	75	136
Südostasien**	16	18	66	50
Norwegen	14	14	72	7
Schweden	11	6	83	18
Ex-Tschechoslowakei[2] **	11	4	85	195
USA[1]	9	8	83	109
Dänemark	7	3	90	30
Schweiz[2]	3	5	92	37
Österreich[2]	2	1	97	224
Kanada[1]	0	0	100	16
Australien[1]	0	17	83	6

* Die Prozentwerte sind auf ganze Zahlen gerundet. ** Vgl. Übersicht zu den einzelnen Ländern einer Herkunftsregion Tabelle 11 im Appendix. Länder mit N<=5 sind ausgeschlossen; richtig=Kategorie 1 beim Grad der Leistungsfähigkeit, ungenau=Kategorien 2 bis 4 beim Grad der Leistungsfähigkeit, falsch=Kategorie 6 beim Grad der Leistungsfähigkeit.

Tabelle 6 bestätigt die oben erwähnte Erwartung falscher Zuordnungen von Zuwanderern aus Staaten mit deutschsprachigen Bevölkerungsgruppen (in der Tabelle 6 mit [2] markiert). Von den relevanten acht

Staaten sind mit Ausnahme von Italien hohe Fehlerraten festzustellen. Der Anteil von Zuwanderern aus Südtirol mit deutschsprachigen Namen scheint gering auszufallen. Auch für Herkunftsstaaten mit Landessprachen überregionaler Bedeutung bzw. starker Wanderungsströme in Nachbarstaaten (siehe Staaten mit einer [1]) können in der nachfolgenden Tabelle des Selektivitätspotentials ausnahmslos niedrige Anteile mit richtiger Onomastik-Zuordnung beobachtet werden.

Die Erwartung einer strukturellen Schwäche des Onomastik-Verfahrens bei Namen aus Sprachen mit hohem Verbreitungsgrad oder deutschsprachigen Namen von Personen mit Migrationshintergrund wird also bestätigt.

In den Tabellen 7a/b wurde für sechs Herkunftsstaaten mit geringerem Anteil richtig identifizierter Personen und einer hinreichenden Fallzahl ein Vergleich bezüglich verschiedener integrationsrelevanter Merkmale (angelehnt an die Integrationsdimensionen von Gordon 1964) zwischen der Gruppe sprachlich richtig Identifizierter und ihrer Grundgesamtheit durchgeführt.[18]

Werden die Befunde aus Tabellen 7a/b systematisch betrachtet, kann festgehalten werden, dass für Staaten der ehemaligen Sowjetunion, Polen und Rumänien, also Herkunftsländer deren Zuwanderung nach Deutschland sowohl von der autochthonen Bevölkerung sowie deutschen Minderheiten geprägt ist, die häufigsten Abweichungen (Anzahl signifikant abweichender Indikatoren) feststellbar sind. Für Polen, Rumänien und die arabischen Staaten, deren Identifizierungsanteil besonders gering ausfällt, sind hingegen die stärksten Abweichungen (Betrag beim Mittelwertvergleich) feststellbar.

Inhaltlich betrachtet kommt es für die Herkunftsländer Ex-Sowjetunion, Polen und Rumänien durch das Onomastik-Verfahren zu einer Unterschätzung des Integrationsfortschrittes. Das Onomastik-Verfahren identifiziert zugewanderte (Spät-)Aussiedler nur teilweise, was dazu führt, dass die Teilpopulation der richtig Identifizierten gegenüber der Gesamtpopulation der Zuwanderer aus diesen Ländern geringere Deutschkenntnisse und höhere Arbeitslosenraten aufweist, sich weniger über gegenseitige Besuche mit der deutschen Bevölkerung austauscht und mit Deutschland identifiziert. Auch sind die durch das Onomastik-Verfahren richtig identifizierten im Schnitt jünger und kürzer in Deutschland, verfügen zu höheren Anteilen über Kinder im Haushalt und weisen eine höhere Erwerbstätigenrate auf. Diese Merkmale bedingen sich natürlich untereinander.

Bei den drei verbleibenden Herkunftsländern/-regionen namentlich Staaten des ehemaligen Jugoslawien, Spanien und arabische Staaten kommt es hingegen zu einer Überschätzung des Integrationsfortschrittes bei den wenigen signifikant abweichenden Indikatoren. So leben die richtig identifizierten Zuwanderer aus Staaten des ehemaligen Jugoslawiens und den arabischen Staaten schon länger in Deutschland und sind stärker sozial integriert. Richtig identifizierte Personen aus arabischen Staaten und Spanien sind darüber hinaus seltener von Arbeitslosigkeit betroffen. Gemeinsam ist den richtig Identifizierten dieser Herkunftsländer/-regionen eine durchschnittlich längere Aufenthaltsdauer, was möglicherweise die stärkere soziale sowie strukturelle Integration mit bedingt.

[18] Für die Teilgruppe der Ausländer (Personen ohne deutsche Staatsangehörigkeit) wurde bereits ein ähnlicher Test für ein anderes namenbasiertes Verfahren (Trigramme) durchgeführt. Siehe dazu Schnell et al. 2014.

Tabelle 7a: Mittelwertvergleiche über soziodemografische und Integrationsmerkmale aller Personen mit Migrationshintergrund mit jenen, die seitens des Onomastik-Verfahrens richtig identifiziert wurden, für spezifische Herkunftslandgruppen mit hohem Selektivitätspotential

Merkmale	Zuwanderergruppen											
	Staaten der ehemaligen Sowjetunion				Polen				Rumänien			
	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$
	MW	KI	MW		MW	KI	MW		MW	KI	MW	
Soziodemografische Merkmale												
Durchschnittsalter	42,86	42,06 – 43,67	41,61	1,25	41,95	41,05 – 42,85	39,97	1,98	43,72	42,21 – 45,23	38,57	5,15
Ant. Frauen	0,57	0,54 – 0,59	0,58		0,57	0,54 – 0,60	0,57		0,57	0,52 – 0,61	0,54	
Ant. Verheiratet	0,65	0,62 – 0,67	0,67		0,58	0,55 – 0,61	0,58		0,63	0,59 – 0,68	0,66	
Ant. mit Kindern <16 Jahren im H	0,46	0,44 – 0,49	0,49		0,42	0,39 – 0,45	0,48	0,06	0,35	0,30 – 0,39	0,46	0,11
Integrationsdimensionen												
Durchschnittliche Aufenthaltsdauer	14,02	13,71 – 14,33	13,05	0,97	16,26	15,52 – 17,00	13,06	3,20	13,24	12,35 – 14,13	8,18	5,06
Kulturell: Ant. mit guten Sprachfertigkeiten*	0,53	0,49 – 0,56	0,50		0,62	0,58 – 0,66	0,52	0,10	0,60	0,54 – 0,66	0,51	0,09
Strukturell: Ant. mit hoher Bildung**	0,42	0,40 – 0,45	0,46	0,04	0,52	0,49 – 0,55	0,50		0,51	0,46 – 0,56	0,53	
Erwerbsstatus												
-Ant. arbeitslos	0,11	0,09 – 0,12	0,15	0,04	0,07	0,06 – 0,09	0,10	0,03	0,07	0,05 – 0,10	0,09	
-Ant. erwerbstätig	0,59	0,57 – 0,62	0,61		0,68	0,65 – 0,71	0,68		0,65	0,61 – 0,70	0,77	0,12
Sozial: Ant. mit Besuch von/ bei Deutschen***	0,79	0,76 – 0,81	0,74	0,05	0,88	0,86 – 0,90	0,86		0,83	0,79 – 0,87	0,79	
Emotional: Ant. mit hoher Identifikation mit Deutschland****	0,76	0,73 – 0,78	0,63	0,13	0,75	0,71 – 0,79	0,60	0,15	0,88	0,83 – 0,93	0,65	0,23

Merkmale stammen jeweils aus dem letzten Erhebungsjahr einer Person; * mindestens gute Kenntnisse beim Sprechen und Schreiben der deutschen Sprache; **mindestens Mittlere Reife mit Berufsausbildung; ***sowohl in den letzten 12 Monaten zu Besuch bei Deutschen als auch diese zu Besuch gehabt; **** sich mindestens überwiegend als Deutscher fühlen; MW = Mittelwert; KI = Konfidenzintervall; H = Haushalt; Ant. = Anteil; $\|\Delta_{\mu}\|$ = Differenzbetrag der Mittelwerte (MW).

Fehler! Verweisquelle konnte nicht gefunden werden.7b: Mittelwertvergleiche über soziodemografische und Integrationsmerkmale aller Personen mit Migrationshintergrund mit jenen, die seitens des Onomastik-Verfahrens richtig identifiziert wurden, für spezifische Herkunftslandgruppen mit hohem Selektivitätspotential

Merkmale	Zuwanderergruppen											
	Staaten des ehemaligen Jugoslawiens				Spanien				Arabische Staaten			
	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$	alle Personen mit Migrationshintergrund		richtig identifiziert	$\|\Delta_{MW}\|$
	MW	KI	MW		MW	KI	MW		MW	KI	MW	
Soziodemografische Merkmale												
Durchschnittsalter	41,14	40,40 – 41,88	42,25	1,11	41,72	40,54 – 42,91	41,75		36,60	34,52 – 38,69	36,29	
Ant. Frauen	0,51	0,48 – 0,53	0,49		0,46	0,42 – 0,50	0,43		0,48	0,40 – 0,56	0,53	
Ant. Verheiratet	0,64	0,61 – 0,66	0,63		0,63	0,59 – 0,67	0,64		0,63	0,55 – 0,71	0,55	
Ant. mit Kindern <16 Jahren im H	0,41	0,39 – 0,44	0,38	0,03	0,40	0,36 – 0,44	0,37		0,75	0,68 – 0,82	0,71	
Integrationsdimensionen												
Durchschnittliche Aufenthaltsdauer	20,63	19,99 – 21,27	21,74	1,11	22,01	21,09 – 22,93	22,34		12,97	11,78 – 14,16	16,60	3,63
Kulturell: Ant. mit guten Sprachfertigkeiten*	0,50	0,47 – 0,53	0,48		0,39	0,35 – 0,43	0,41		0,43	0,34 – 0,51	0,49	
Strukturell: Ant. mit hoher Bildung**	0,25	0,23 – 0,28	0,27		0,21	0,18 – 0,24	0,22		0,32	0,24 – 0,40	0,38	
Erwerbsstatus												
-Ant. arbeitslos	0,10	0,08 – 0,11	0,09		0,07	0,05 – 0,09	0,04	0,03	0,23	0,16 – 0,30	0,15	0,08
-Ant. erwerbstätig	0,57	0,55 – 0,60	0,60		0,64	0,60 – 0,68	0,67		0,47	0,39 – 0,55	0,53	
Sozial: Ant. mit Besuch von/ bei Deutschen***	0,80	0,78 – 0,82	0,84	0,04	0,87	0,84 – 0,90	0,88		0,62	0,53 – 0,71	0,73	0,11
Emotional: Ant. mit hoher Identifikation mit Deutschland****	0,32	0,29 – 0,34	0,29		0,13	0,10 – 0,16	0,14		0,59	0,43 – 0,75	0,67	

Merkmale stammen jeweils aus dem letzten Erhebungsjahr einer Person; * mindestens gute Kenntnisse beim Sprechen und Schreiben der deutschen Sprache; **mindestens Mittlere Reife mit Berufsausbildung; ***sowohl in den letzten 12 Monaten zu Besuch bei Deutschen als auch diese zu Besuch gehabt; **** sich mindestens überwiegend als Deutscher fühlen; MW = Mittelwert; KI = Konfidenzintervall; H = Haushalt; Ant. = Anteil; $\|\Delta_{\mu}\|$ = Differenzbetrag der Mittelwerte (MW).

3.4 Erklärungsansätze für falsche Zuordnungen

Aufgrund datenschutzrechtlicher Bedingungen war eine Zusammenführung von Namen und SOEP-Befragungsdaten natürlich nicht möglich. Dies erschwert eine verlässliche Klärung der Gründe für die 3.315 falschen Zuordnungen des Onomastik-Verfahrens. Trotzdem kann man auf Grundlage der empirischen Daten Hinweise auf plausible Erklärungsansätze für falsche Zuordnungen erhalten.
Deutliche Hinweise und Indizien auf zentrale Fehlerursachen konnten bereits anhand der bisher aufgeführten Tabellen gewonnen werden. Mit Hilfe der folgenden Analysen, die sich auf die 3.315 falsch zugeordneten Fälle beschränken, soll möglichen Erklärungen für die Fehlerursachen noch eingehender nachgegangen werden.

Tabelle 8: Verteilung falscher Zuordnungen nach Migrationshintergrund

			Migrationshintergrund nach SOEP-Angaben		gesamt
			nein	ja	
Migrationshintergrund nach Onomastik	nein	N	0	2.617	2.617
		Zellenprozente	0	79	79
	ja	N	199	499	698
		Zellenprozente	6	15	21
		Gesamt	199	3.116	3.315
			6	94	100

Immerhin knapp 79 Prozent aller falschen Zuordnungen wurden dadurch verursacht, dass einem SOEP-Befragten mit Migrationshintergrund fälschlicherweise vom Onomastik-Verfahren kein Migrationshintergrund (Deutschland) zugeordnet wurde. Eine plausible Erklärung für diese 1. Fehlerart wären Personen mit Migrationshintergrund, die einen komplett deutschsprachigen Namen tragen.
Die 2. Fehlerart beruht auf der Zuordnung des falschen (ausländischen) Herkunftsgebietes bei Personen mit Migrationshintergrund. Diese machen rund 15% aller falschen Zuordnungen aus. Hier könnten die staatsübergreifende Ausdehnung von Sprachräumen oder starke Wanderungsströme die Fehlerursache sein. Die verbleibende 3. Fehlerart (die fälschliche Zuordnung eines Migrationshintergrundes bei SOEP-Befragten ohne Migrationshintergrund) ist aufgrund seines Anteils von sechs Prozent an allen Fehlern des Onomastik-Verfahrens nur von geringer Bedeutung. Die folgende Aufstellung ist der Versuch einer Systematisierung von möglichen Gründen für die beiden erstgenannten relevanten Fehlerarten.

Tabelle 9: Mögliche Gründe für Fehler

1. Fehlerart: Zuordnung von kein Migrationshintergrund bei Personen mit Migrationshintergrund […durch komplett deutschsprachige Namen von Migranten]	
1.1	binationale Ehepartner [der Nachname des deutschen Partners wird übernommen und der Vorname war bereits in Deutschland üblich]
1.2	Nachkommen binationaler Eltern [Kind trägt den Nachnamen des deutschen Elternteils und erhält deutschen Vornamen]
1.3	Zuwanderung aus deutschsprachigem Ausland [Österreich, Schweiz, Italien [Südtirol], Frankreich [Elsass u. Lothringen]]
1.4	Zuwanderung als deutsche Minderheit [Aussiedler aus Polen, Ex-Sowjetunion, Siebenbürgen/Banat, Karpaten, Böhmen/Mähren]
2. Fehlerart: Zuordnung falscher ausländischer Herkunftsregion bei Personen mit Migrationshintergrund […durch Spracheräume mit mehreren Staaten und historische Wanderungsströme in Nachbarstaaten]	
2.1	Ein Sprachraum umfasst mehrere Staaten bzw. Herkunftsregionen [z.B. Großbritannien-USA-Australien, Kanada-Frankreich, Brasilien-Portugal, Spanien-Lateinamerika]
2.2	Starke Wanderungsströmungen zwischen benachbarten Staaten [z.B. Ungarn-Österreich-Vielvölkerstaat]

Die Zuweisung dieser sechs möglichen Gründe auf die 3.315 Fehler wurde auf Grundlage der SOEP-Befragungsdaten realisiert. Dazu wurden Angaben zur Nationalität und Herkunftsstaaten der Eltern, der Partner und der Befragten genutzt. Es ergibt sich daraus folgende Verteilung auf die möglichen Fehlerursachen.

Tabelle 10: Häufigkeit möglicher Gründe für falsche Zuordnungen

Mögliche Gründe für falsche Zuordnungen	n	%	%
1.1 binationale Ehepartner	95	3	
1.2 binationale Eltern	122	4	
1.3 Zuwanderung aus deutschsprachigem Ausland	265	8	
1.4 Zuwanderung als deutsche Minderheit	1.368	41	77
2.1 Sprachraum umfasst mehrere Staaten	53	2	
2.2 Starke Wanderungen zwischen Nachbarstaaten	150	5	
mehrere Gründe (1.1 - 2.2)	495	14	
Grund nicht benennbar	767	23	23
Total	3.315	100	100

Immerhin rund 77 Prozent aller falschen Zuordnungen des Onomastik-Verfahrens lassen sich einem oder mehreren der sechs aufgeführten Ursachen zuordnen. Darunter ist die Zuwanderung als deutsche Minderheit (1.4) mit rund 41 Prozent aller Fehler der potentielle Erklärungsansatz mit dem größten Anteil. Addiert man die Zuwanderung aus dem deutschsprachigen Ausland (1.3) hinzu, ist jeder zweite Fehler des Onomastik-Verfahrens bereits zum Zeitpunkt der Zuwanderung durch Personen mit vermutlich vollständig deutschsprachigen Namen unvermeidbar.

Eine fortschreitende gesellschaftliche Entwicklung und Integration in Deutschland, die sich z.B. auch durch Ehen zwischen Personen mit und ohne Migrationshintergrund ausdrückt, spielt mit ihren Konsequenzen für eine deutschsprachige Namensgebung offenbar eine untergeordnete Rolle. Beide Gruppen (1.1 und 1.2) vereinen nur insgesamt 7 Prozent der Fehler.[19]

[19] Bei separater Betrachtung jeder potentiellen Fehlerursache und Auflösung der Kategorie „mehrere Gründe“ weisen 15 Prozent aller Fehlzuweisungen ein binationales Elternhaus und 10 Prozent eine binationale Partnerschaft auf. D.h. diese beiden Gründe für Fehlzuweisungen treten mehrheitlich gleichzeitig mit anderen Gründen für Fehlzuweisungen auf. Der Ausschluss von Personen, die zwei ausländische

4. Zusammenfassung und Ausblick

Der Gütetest hat gezeigt, dass das Onomastik-Verfahren nahezu alle vollständig vorliegenden Namensangaben zu einer Sprachherkunftslandgruppe zuordnen kann. Nur in 0,2 Prozent der Fälle war dies nicht möglich. Für weitere 4 Prozent konnte nur eine ungenaue Zuordnung erfolgen und 14 Prozent wurden vom Verfahren tatsächlich falsch zugeordnet. Da die ungenauen und falschen Zuordnungen überwiegend bei Personen mit Migrationshintergrund vorkommen und diese Teilpopulation in den Testdaten überproportional vorzufinden ist, würde bei einer Übertragung der Testergebnisse dieser Gesamtleistungsfähigkeitsmerkmale des Onomastik-Verfahrens auf die Gesamtbevölkerung der Anteil ungenauer und falscher Zuordnungen deutlich niedriger ausfallen.
Die differenzierte Betrachtung nach Migrationshintergrund hat darüber hinaus gezeigt, dass die Leistungsfähigkeit des Verfahrens bei Personen ohne Migrationshintergrund mit 98 Prozent richtigen Zuordnungen sehr hoch ausfällt. Die Identifikation innerhalb der Personen mit Migrationshintergrund führt hingegen in jedem vierten Fall zu einer falschen Zuordnung. Das Verfahren übersieht eher einen bestehenden Migrationshintergrund, als dass es diesen fälschlicherweise zuschreibt. Für die zwei Gütemaße Effektivität und Selektivitätspotential ergibt sich ein unterschiedliches Bild. Für 22 der 36 untersuchten Herkunftsländer/-regionen kann eine hohe Effektivität des Onomastik-Verfahrens von mindestens 70 Prozent beobachtet werden. Darunter sind die wichtigsten Zuwanderungsländer und die Gruppe ohne Migrationshintergrund. Die im Gütetest ermittelten länder/-regionenspezifischen Fehlerquoten bei der Effizienzanalyse könnten bei der Planung von Bruttostichprobenumfängen eingesetzt werden, weil sie einen Anhaltspunkt für den zu erwartenden screenout-Anteil[20] darstellen. Hinsichtlich des Selektivitätspotentials kann so eine hohe Identifizierungsrate nur für 11 der 40 untersuchten Herkunftsländer/-regionen festgestellt werden. Darunter sind erneut die Türkei, Griechenland, Italien und die Gruppe ohne Migrationshintergrund. Die beiden Herkunftsländer Türkei und Griechenland weisen auf beiden Gütemaßen eine Trefferquote von über 90 Prozent auf. Ob das Selektivitätspotential tatsächlich zu einer selektiven Gruppe führt, konnte anhand des Vergleichs zwischen Grundgesamtheit und vom Verfahren richtig Identifizierten näher betrachtet werden. Für die hier betrachteten Vergleichsmerkmale und Herkunftsländer/-regionen konnte erwartungsgemäß festgestellt werden, dass zwischen Identifizierungsanteil und Abweichungsstärke (Abweichungsbetrag beim Mittelwertvergleich) ein negativer Zusammenhang besteht. Für die Herkunftsländer, deren Zuwanderung nach Deutschland sowohl von der autochthonen Bevölkerung sowie deutschen Minderheiten geprägt ist (Ex-Sowjetunion, Polen und Rumänien), kommt es durch das Onomastik-Verfahren zu einer Unterschätzung des Integrationsfortschrittes über alle Integrationsdimensionen. Bei den anderen drei betrachteten Herkunftsländern wird hingegen durch den Einsatz des Onomastik-

Herkunftsländer in ihrer Migrationsgeschichte aufweisen (siehe die 3. Gruppe in Tabelle 1) aus der Analysegrundgesamtheit bedingt den geringen Umfang der Fehlerursache „binationales Elternhaus“ mit.

[20] Der screenout-Anteil ist jener Anteil an der Bruttostichprobe, der nicht die Zielgruppenkriterien erfüllt. So weisen z.B. 9 Prozent der HS-Gruppe ‚Polen‘ gar keinen polnischen Migrationshintergrund auf und reduzieren entsprechend die Bruttostichprobe neben dem Anteil jener, die nicht kontaktiert werden können und nicht teilnehmen möchten weiter.

Verfahrens die soziale und strukturelle Integration überschätzt. Diese hier ermittelten gruppenspezifisch ausfallenden Befunde zur Verzerrung geben erneut Anhaltspunkte, die von anderen Forschern zur Beurteilung ihrer eigenen inhaltlichen Befunde herangezogen werden könnten, wenn bei der Stichprobenziehung der verwendeten Datengrundlage das Onomastik-Verfahren zum Einsatz kam. Und zu guter Letzt konnte gezeigt werden, dass die Zuwanderung von Personen mit vermutlich vollständig deutschsprachigen Namen für mindestens die Hälfte aller Fehler einen plausiblen Erklärungsansatz darstellt. Dieser Fehleranteil kann nicht durch eine Verbesserung des Verfahrens in der Zukunft reduziert werden und stellt somit eine grundlegende Schwäche des Verfahrens dar. Personen mit Migrationshintergrund und vollständig deutschsprachigen Namen (z.B. Spätaussiedler, Nachkommen binationaler Eltern, Personen mit deutschem Ehepartner) sind jedoch meist nicht durch alternative Methoden (z.B. Staatsangehörigkeit oder Geburtsort bei Melderegister-Stichproben) zu identifizieren. In der Regel besitzen sie die deutsche Staatsangehörigkeit oder das ergänzend benötigte Ziehungsmerkmal Geburtsort ist aufgrund der Gesetzeslage nicht einsetzbar (siehe Bundesmeldegesetz (BMG)). Somit schließt das Onomastik-Verfahren eine bestehende Lücke, auch wenn es dies nicht perfekt, sondern nur mit Einschränkungen kann.

Die Verallgemeinerbarkeit der ermittelten Befunde ist aufgrund der verwendeten Datenbasis, den darin enthaltenen Fallzahlen und der bislang erfolgten Güteanalysen natürlich mehrfach beschränkt. So haben nur Personen, die in einen der vielfältigen Stichprobenziehungsrahmen des SOEPs gefallen sind, die darüber hinaus an mindestens einer Befragung des SOEPs teilgenommen und Angaben zu jenen Merkmalen gemacht haben, die zur sicheren Bestimmung eines spezifischen Migrationshintergrundes herangezogen worden sind, Eingang in die Datengrundlage gefunden. Letztere Beschränkung dürfte auch dazu geführt haben, dass die Testdaten hinsichtlich des Migrationshintergrundes deutlich von der Verteilung in der Bevölkerung in Deutschland abweichen. Da die Teilnahme am SOEP freiwillig ist, lassen sich die Befunde nur auf befragungsbereite Personen verallgemeinern. Aus diesem Pool speist sich jedoch auch der Großteil der sozialwissenschaftlichen Befragungen. Auch beruhen die herkunftslandspezifischen Analysen nur in Teilen auf einer ausreichend großen Fallzahl, so dass für diese mit den vorliegenden Befunden nur ein erster Eindruck und kein robustes Ergebnis vermittelt werden kann.[21]

Die bisherigen Analysen sind darüber hinaus nicht umfassend, sondern es stellen sich anschließende Forschungsfragen. Dies wäre beispielsweise die Analyse einer Zuwanderergruppe unabhängig von der Kategorisierung nach Herkunftsstaaten bzw. -regionen (z.B. die Zuwanderergruppe der (Spät-)Aussiedler oder Muslime). Die Gruppe der (Spät-)Aussiedler stellt für die Stichprobenziehung eine besondere Herausforderung dar. Diese spezifische Zuwanderergruppe lässt sich über das Merkmal der Staatsangehörigkeit überhaupt nicht identifizieren, da (Spät-)Aussiedler kurz nach Zuzug nach Deutschland die deutsche Staatsangehörigkeit erhalten. Darüber hinaus lässt sich diese spezifische Gruppe auch mit Hilfe des Onomastik-Verfahrens, wie bereits dieser Test sowie Liebau 2011 angedeutet haben, nur teilweise aufspüren. (Spät-)Aussiedler finden sich sowohl in anderen Herkunftslandgruppen aus der ehemaligen Sowjetunion und der Gruppe jener Personen ohne

[21] Diese Einschränkung könnte nur mit Hilfe eines disproportionalen Ziehungsansatzes (vom Gruppenumfang kleine Zuwanderergruppen werden überproportional bei der Stichprobenziehung berücksichtigt) oder einem sehr großen Stichprobenumfang, wie er nur im Mikrozensus vorliegt, überwunden werden.

Migrationshintergrund (Liebau 2011). Eine genauere Betrachtung der (Spät-)Aussiedler scheint somit besonders kenntnisfördernd zu sein. Auch das große Interesse an der spezifischen Betrachtung von Muslimen (Brettfeld/Wetzels 2007, Haug et al. 2009) und die Möglichkeit der Identifizierung dieser Gruppe über das Onomastik-Verfahren scheint ein weiteres gutes Anwendungsbeispiel für den Test der Leistungsfähigkeit dieses Verfahrens darzustellen. Genauso spannend wäre ein eingehender Blick in die bislang bewusst ausgeschlossene Gruppe jener Personen mit Migrationshintergrund, die in ihrer Migrationsgeschichte zwei oder mehr ausländische Wurzeln aufweist. Auch herkunftslandspezifische Analysen zur zweiten Zuwanderergeneration versprechen Einblick in das Potential sowie die Grenzen des Onomastik-Verfahrens zu liefern (erste Befunde dazu siehe Kruse/Dollmann 2017). Auch die Simulation einer idealen Stichprobenbildung für eine spezifisch interessierende Zuwanderergruppe, die auf das Onomastik-Verfahren zurückgreift, aber bei der Zusammensetzung der HS-Sprachgruppen gleich das Verzerrungspotential dieser mit berücksichtigt, stellt ein denkbares Anwendungsfeld dar.
Ebenso beruhen die bisherigen Befunde auf bivariaten Analysen. Ob die Kontrolle von relevanten Merkmalen (wie der Aufenthaltsdauer etc.) in multivariaten Analysen die aufgezeigten Abweichungen nivellieren oder sich erst neue herauskristallisieren ist noch eine empirisch offene Frage.
Trotz aller benannter Einschränkungen und weiterführender Fragestellungen bietet dieses Papier den ersten umfassenden Gütetest des Onomastik-Verfahrens von Humpert und Schneiderheinze, welches sich in Deutschland als Standardverfahren für sozialwissenschaftliche Stichprobenziehungen bei Migranten etabliert hat (Schnell 2011: 290). Der Test darf als umfassend bezeichnet werden, weil die Datengrundlage verlässlich den Migrationshintergrund einer Person klärt, große Fallzahlen und ein großes Spektrum von Herkunftslandgruppen bietet.

Danksagung:

Die Autorengruppe insgesamt dankt den mitwirkenden MitarbeiterInnen von Kantar Public (vormals TNS-Infratest), die alle Datenschutzanforderungen beachtend die Onomastik-Zuweisung aller Namen der SOEP-Befragten umgesetzt haben.

Literaturverzeichnis

Akademie der Wissenschaften und der Literatur Mainz: Digitales Familiennamenwörterbuch Deutschlands (DFD). Abgerufen unter: http://www.namenforschung.net/dfd/projektvorstellung, Zugriff am 09.05.2018.

Angenendt, Steffen, David Kipp und Amrei Meier (2017): Gemischte Wanderungen. Herausforderungen und Optionen einer Dauerbaustelle der deutschen und europäischen Asyl- und Migrationspolitik. Gütersloh: Bertelsmann Stiftung.

Brettfeld, K. und P. Wetzels, 2007: Muslime in Deutschland. Integration, Integrationsbarrieren, Religion sowie Einstellungen zu Demokratie, Rechtsstaat und politisch-religiös motivierter Gewalt. Berlin: Bundesministerium des Innern.

Die Bundesregierung (2016): 11. Bericht der Beauftragten der Bundesregierung für Migration, Flüchtlinge und Integration – Teilhabe Chancengleichheit und Rechtsentwicklung in der Einwanderungsgesellschaft Deutschland. Berlin, Dezember 2016.

Familiennamenbuch der Schweiz. Répertoire des noms de famille suisses. Repertorio dei nomi di famìglis svizzeri (1968-1971). 2., erw. Aufl. 6 Bde. Zürich: Polygraph Verlag.

Gordon, Milton (1964): Assimilation in American Life. New York: Oxford University Press.

Groves, Robert M. (2006): Nonresponse Rates and Nonresponse Bias in Household Surveys. In: Public Opinion Quarterly, Volume 70 (5), S. 646–675.

Haug, Sonja, Stephanie Müssig und Anja Stichs (2009): Muslimisches Leben in Deutschland. Nürnberg: Bundesamt für Migration und Flüchtlinge.

Humpert, Andreas und Klaus Schneiderheinze (2000): Stichprobenziehung für telefonische Zuwandererumfragen - Einsatzmöglichkeiten der Namenforschung (Onomastik). In: ZUMA Nachrichten 24 (47). S. 36-63.

Humpert, Andreas und Klaus Schneiderheinze (2002): Stichprobenziehung für telefonische Zuwandererumfragen - Praktische Erfahrungen und Erweiterung der Auswahlgrundlage. In: Gabler, Siegfried und Sabine Häder (Hrsg.): Telefonstichproben - Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann. S. 187-208.

Kroh, Martin, Simon Kühne, Jan Goebel und Friedrike Preu (2015): The 2013 IAB-SOEP Migration Sample (M1): Sampling Design and Weighting Adjustment. SOEP Survey Papers 271: Series C. Berlin: DIW/SOEP.

Kruse, Hanno und Jörg Dollmann (2017): Name-based measures of neighborhood composition: how telling are Neighbors' names? In: Survey Research Methods, Vol. 11 (4), S. 435-450.

Leicht, René, Andreas Humpert, Markus Leiss, Michael Zimmer-Müller und Maria Lauxen-Ulbrich (2005): Existenzgründungen und berufliche Selbständigkeit unter Aussiedlern (Russlanddeutsche). Mannheim: Institut für Mittelstandsforschung der Universität Mannheim.

Liebau, Elisabeth (2011): Arbeitsmarktintegration von hochqualifizierten Zuwanderern. Erklärung des spezifischen Integrationsmusters in den deutschen Arbeitsmarkt von Aussiedlern und jüdischen Kontingentflüchtlingen aus der ehemaligen Sowjetunion. Dissertation. Mannheim: Universität Mannheim.

Liebau, Elisabeth und Ingrid Tucci (2015): Migrations- und Integrationsforschung mit dem SOEP von 1984 bis 2012: Erhebung, Indikatoren und Potenziale. SOEP Survey Papers 270: Series C. Berlin: DIW/SOEP.

Nederlands Repertorium van Familienamen (1963-88). Hrsg. Meertens, Pieter. 14 Bde. Assen/Amsterdam.

Rymut, Kazimierz (1993): Słownik nazwisk współcześnie w Polsce używanych [Wörterbuch der in Polen gebräuchlichen Familiennamen]. 10 Bde. Kraków: Instytut Języka Polskiego.

Salentin, Kurt (2014): Sampling the Ethnic Minority Population in Germany – The Background to „Migration Background“. In: Methoden, Daten, Analysen 8 (1), S. 25-52.

Schneiderheinze, Klaus (2004): Verzeichnis ausgewählter Literatur aus der Namenforschung (Onomastik). Abgerufen unter: http://www.stichproben.de/herunterladen/onomastik_literatur_HS_GbR.pdf, Zugriff am 09.05.2018.

Schnell, Rainer, Mark Trappmann und Tobias Gramlich (2014): A study of Assimilation bias in name-based sampling of migrants. In: Journal of Official Statistics 30 (2). S. 231-249.

Schnell, Rainer, Tobias Gramlich, Tobias Bachteler, Jörg Reiher, Mark Trappmann, Menno Smid und Inna Becher (2013): Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. In: Methoden, Daten, Analysen 7 (1), S. 5-33.

Schnell, Rainer, Paul B. Hill und Elke Esser (2011): Methoden der empirischen Sozialforschung. 9. Auflage. München: Oldenbourg.

Schupp, J., Wagner G.G. (1995): Die Zuwanderer-Stichprobe des Sozio-ökonomischen Panels (SOEP). In: Vierteljahrshefte zur Wirtschaftsforschung, 64 (1), S. 16-25.

Sozio-oekonomisches Panel (SOEP), Daten für die Jahre 1984-2013, Version 30, SOEP, 2015, doi:10.5684/soep.v30.

Statistisches Bundesamt (2016): Bevölkerung mit Migrationshintergrund auf Rekordniveau. (Pressemitteilung vom 16. September 2016). Online unter: https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2016/09/PD16_327_122pdf.pdf?__blob=publicationFile, Zugriff am 06.06.2017.

Thygesen, Lau Caspar, Camilla Daasnes, Ivan Thaulow und Henrik Brønnum-Hansen (2011): Introduction to Danish (nationwide) registers on health and social issues: Structure, access, legislation, and archiving. In: Scandinavian Journal of Public Health, 39 (7), S. 12-16.

Appendix

Tabelle 11: Übersicht zu den einzelnen Ländern einer Herkunftsregion

Ex-Jugoslawien	Ex-Jugoslawien, Montenegro, Kroatien, Bosnien und Herzegowina, Mazedonien, Slowenien
Sub-Sahara	Benin, Äthiopien, Ghana, Nigeria, Tansania, Mosambik, Somalia, Südafrika, Eritrea, Burkina Faso, Sambia, Angola, Namibia, Kenia, Botswana, Guinea, Elfenbeinküste, Uganda, Mali, Kamerun, Sudan, Kongo, Togo, Tschad
Ex-Sowjetunion	Ex-Sowjetunion, Russland, Moldawien, Kasachstan, Ukraine, Weißrussland, Usbekistan, Georgien, Aserbaidschan, Litauen, Lettland, Kirgisistan, Tadschikistan, Armenien, Turkmenistan, Estland
Maghreb-Staaten	Tunesien, Marokko, Algerien
Benelux	Luxemburg, Belgien, Niederlande
Arabische Staaten	Syrien, Saudi Arabien, Irak, Libanon, Ägypten, Vereinigte Arabische Emirate, Jordanien, Libyen, Kuwait, Oman, Jemen, Palästina
Mittlerer Osten	Afghanistan, Bangladesch, Pakistan
Südostasien	Indonesien, Philippinen, Singapur, Laos, Malaysia
Lateinamerika	Chile, Bolivien, Mexiko, Argentinien, Kolumbien, Venezuela, Kuba, Brasilien, Peru, El Salvador, Costa Rica, Ecuador, Puerto Rico, Dominikanische Republik, Nicaragua, Haiti, Paraguay, Uruguay
Ex-Tschechoslowakei	Ex-Tschechoslowakei, Tschechien, Slowakei