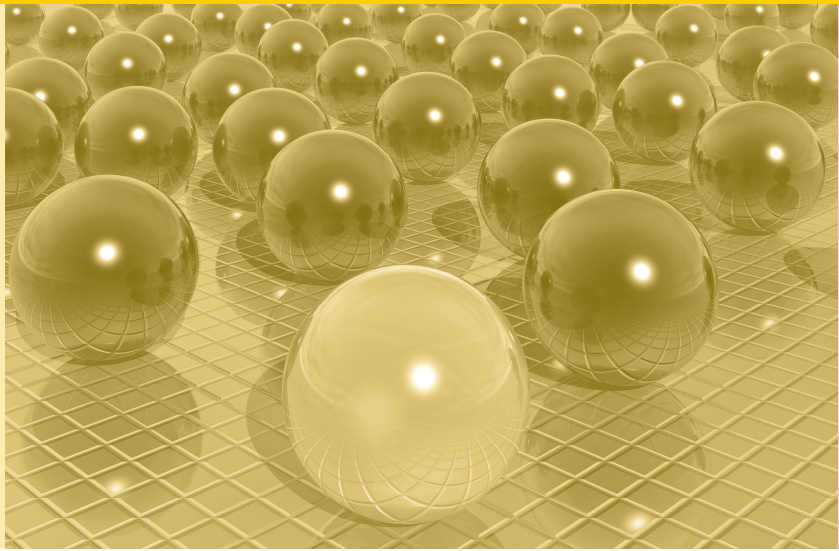


# FDZ-Arbeitspapier Nr. 4\*



Entwicklung des Mikrosimulationsmodells EITDsim

Georg Struch

2013

## Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder  
Herstellung: Information und Technik Nordrhein-Westfalen  
Telefon 0211 9449-01 • Telefax 0211 9449-8000  
Internet: [www.forschungsdatenzentrum.de](http://www.forschungsdatenzentrum.de)  
E-Mail: [forschungsdatenzentrum@it.nrw.de](mailto:forschungsdatenzentrum@it.nrw.de)

### Fachliche Informationen zu dieser Veröffentlichung:

Forschungsdatenzentrum der  
Statistischen Landesämter  
– Geschäftsstelle –  
Tel.: 0211 9449-2873  
Fax: 0211 9449-8087  
[forschungsdatenzentrum@it.nrw.de](mailto:forschungsdatenzentrum@it.nrw.de)

### Informationen zum Datenangebot:

Statistisches Bundesamt  
Forschungsdatenzentrum  
  
Tel.: 0611 75-2420  
Fax: 0611 72-3915  
[forschungsdatenzentrum@destatis.de](mailto:forschungsdatenzentrum@destatis.de)

Forschungsdatenzentrum der  
Statistischen Landesämter  
– Geschäftsstelle –  
Tel.: 0211 9449-2876  
Fax: 0211 9449-8087  
[forschungsdatenzentrum@it.nrw.de](mailto:forschungsdatenzentrum@it.nrw.de)

Erscheinungsfolge: unregelmäßig  
Erschienen im Mai 2013

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter [www.forschungsdatenzentrum.de](http://www.forschungsdatenzentrum.de) angeboten.

© Information und Technik Nordrhein-Westfalen, Düsseldorf, 2013  
(im Auftrag der Herausbergemeinschaft)

Vervielfältigung und Verbreitung, nur auszugsweise, mit Quellenangabe gestattet. Alle übrigen Rechte bleiben vorbehalten.

Fotorechte Umschlag: ©artSILENCEcom – Fotolia.com

**Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse des genannten Autors im Zusammenhang mit der Nutzung der bereitgestellten Daten der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.**

# FDZ-Arbeitspapier Nr. 4\*

Entwicklung des Mikrosimulationsmodells EITDsim

Georg Struch

2013



# Entwicklung des Mikrosimulationsmodells EITDsim

Georg Struch\*

## Zusammenfassung

Dieser Aufsatz dokumentiert den Aufbau des Mikrosimulationsmodells EITDsim. EITDsim kann zur empirischen (ex ante) Evaluation von steuer- und finanzpolitischen Fragestellungen, insbesondere von (potentiellen) Steuerreformen, verwendet werden. Dafür kombiniert EITDsim auf Basis des erweiterten Mikrodatensatzes FAST 2004 ein statisches Steuer-Simulationsmodell zur Schätzung kurzfristiger Erst-Rundeneffekte mit einem diskreten Arbeitsangebotsmodell zur Schätzung mittelfristiger Zweit-Rundeneffekte. Aufgrund unvollständiger Informationen in der FAST 2004 wird als Datengrundlage für das Arbeitsangebotsmodell in EITDsim das Sozioökonomische Panel (SOEP) verwendet. Auf dem Wege einer Datensatzfusion werden die Informationen aus beiden Mikrodatensätzen kombiniert. Hierfür werden drei verschiedene Fusionsalgorithmen hinsichtlich ihrer zu erwartenden Gütequalität evaluiert. Die Analyse der Fusionsergebnisse zeigt, dass ein Kovariaten-Matching inklusive eines geschätzten Propensity Scores am ehesten dazu geeignet ist, aus den Beobachtungseinheiten der FAST und des SOEP statistische Zwillinge zu identifizieren.

*JEL-Klassifikation:* C51, C53, H24, J22

*Schlagworte:* Mikrosimulation, Arbeitsangebot, Datensatzfusion

\*Georg Struch, Lehrstuhl für Öffentlicher Sektor, Finanz- und Sozialpolitik, Wirtschafts- und Sozialwissenschaftliche Fakultät, Universität Potsdam, August-Bebel-Straße 89, 14482 Potsdam. Tel. +49 (0) 331 - 977 33 95, E-Mail: Georg.Struch@Uni-Potsdam.de

## Vorbemerkungen zum Forschungsprojekt EITD

Bevor in den folgenden Abschnitten der Aufbau und die Funktionsweise des Mikrosimulationsmodells EITDsim erläutert wird, soll an dieser Stelle kurz auf die Anfänge des Forschungsprojektes eingegangen werden. Das Akronym EITD steht für *Extended Income Tax Dataset* und ist das Ergebnis eines ersten Forschungsansatzes, eine geeignete Datenbasis für die Simulation steuer- und sozialpolitischer Reformvorschläge und -alternativen zu konstruieren.<sup>1</sup> EITD erweiterte im Zuge eines Fusions- bzw. Erweiterungsalgorithmus den Informationsgehalt der amtlichen Steuerstatistik (FAST 2001) zum einen mit zusätzlichen sozioökonomischen Variablen sowie zusätzlichen Einzeldatensätzen durch das SOEP 2001. Zum anderen wurden, der demographischen und monetären Entwicklung folgend, sowohl die Fallgewichte der Einzeldatensätze als auch die Einkommens- und Ausgabenbestandteile bis an den (damals) aktuellen Rand fortgeschrieben. EITD stellte daher für die Evaluierung von Steuerrechtsänderungen eine umfassendere Datengrundlage dar als die Ausgangsdaten der amtlichen Steuerstatistik.

Die Konstruktion des EITD basierte grundsätzlich auf einer Erweiterung der Steuerstatistik FAST 2001 durch das SOEP. Hierbei wurden nach struktureller Angleichung beider Datensätze mit der Methode des Propensity Score-Matchings zunächst diejenigen Fälle identifiziert, welche sich in beiden Datensätzen nur geringfügig unterscheiden, und anschließend wurden die Informationen aus beiden Datensätzen derart fusioniert, dass die Informationsbreite der amtlichen Steuerstatistik stieg. Um zu erreichen, dass der neue integrierte Mikrodatsatz tatsächlich alle steuerlich veranlagten Fälle in Deutschland abbildet, wurden, dem Ansatz von Bach et al. (2009) folgend, diejenigen Einzeldatensätze, welche nach dem Matching "übrig" waren, dem neuen Datensatz angehängen. Mit diesem Vorgehen wurden die amtlichen Steuerdaten (hochgerechnet) um über 18 Mio. Fälle erweitert. Diese erweiterte Datengrundlage wurde dann mit einem *static aging*-Ansatz an die demographische Struktur des (damals) aktuellen Randes angepasst. Hierbei wurden die Altersstruktur, Familienstände, Erwerbsbeteiligung und Erwerbstätigkeit berücksichtigt. Neben der veränderten demographischen Struktur wurden schließlich die Einkommens- und Ausgabengrößen nominal fortgeschrieben. Hierbei wurden verschiedene Einkommens- und Ausgabenarten, soweit möglich, getrennt mit den entsprechenden Fortschreibungsgrößen angepasst.

---

<sup>1</sup> Struch & Jenderny (2010).

Nach der Veröffentlichung der FAST 2004 im Jahr 2010 wurde der Ansatz des Forschungsprojektes EITD vollständig überarbeitet. Das Ziel war nun nicht mehr (nur) die Konstruktion einer informationell erweiterten Datengrundlage für Mikrosimulationsmodelle, sondern die Konstruktion eines integrierten Steuer-Transfer-Mikrosimulationsmodells zur Evaluation von kurz- und mittelfristigen Effekten von (potentiellen) Steuerreformvorschlägen.

## 1 Einleitung

Gerade in wirtschaftlich schwierigen Zeiten für die öffentlichen Haushalte braucht eine nachhaltig gestaltende Finanz- und Steuerpolitik ein hohes Maß an Information über die Wirkungszusammenhänge von verschiedenen Politikoptionen. Im Vorfeld einer Steuerreform sind die konkreten Auswirkungen nur schwer abzuschätzen. Aufgrund der hohen Komplexität von Steuer- und Abgabensystemen werden zur ex ante Evaluation von Reformvorschlägen deshalb vermehrt Simulationsmodelle eingesetzt. Dabei wird versucht mit dem Modell eine möglichst genaue Abbildung des realen Wirtschafts- und Steuersystems zu erreichen, um davon ausgehend die Wirkungen einzelner finanzpolitischer Instrumente identifizieren und quantifizieren zu können. In den vergangenen Jahren sind zum Zwecke der wissenschaftsorientierten Politikberatung eine Reihe von mikrodatenbasierten Simulationsmodellen entwickelt worden, u.a. von Steiner et al. (2005), Fuest et al. (2005), Wagenhals & Buck (2006) Steiner & Wakolbinger (2009), Flory & Stöwhase (2010) oder Peichl et al. (2010), die sich aber sowohl hinsichtlich der zugrundeliegenden Datenbasis als auch hinsichtlich der methodischen Konzeption zum Teil erheblich unterscheiden. Einig sind sich all diese Modelle aber in der Zielsetzung, dem politischen Entscheidungsträger Informationen über die fiskalischen, allokativen und distributiven Effekte einzelner Reformoptionen bereitzustellen und zwar noch bevor diese in das reale Steuer- und Transfersystem implementiert werden.

In diesem Kontext steht auch das Mikrosimulationsmodell EITDsim, welches zur empirischen (ex ante) Analyse von Reformen des bestehenden Steuer- und Transfersystems auf das verfügbare Einkommen und das Arbeitsangebot von privaten Haushalten in Deutschland entwickelt wurde. EITDsim kombiniert ein Steuer-Transfer-Modell, das die Effekte von Sozialabgaben, Einkommensteuer und Solidaritätszuschlag sowie staatlicher Transferzahlungen auf das Nettohaushaltseinkommen simuliert, mit einem diskreten Arbeitsangebotsmodell,

welches zur ökonomischen Schätzung von Arbeitsangebotsreaktionen der Haushalte verwendet wird. Mit EITDsim ist es daher möglich, Aussagen über die, kurz- und mittelfristigen, fiskalischen, allokativen und distributiven Effekte einer (hypothetischen) Reform des bestehenden Steuer- und Transfersystems zu treffen.<sup>2</sup>

Voraussetzung für die detaillierte Abbildung der komplexen institutionellen Zusammenhänge des Steuer- und Transfersystems und die damit simulierbaren Politikoptionen ist (jedoch) eine aktuell repräsentative und qualitativ hochwertige Datenbasis. Angesichts der hohen Informationsanforderungen kann es durchaus sein, dass eine neue (artifizielle) Datenbasis aus verschiedenen Mikrodatensätzen konstruiert werden muss, um über alle notwendigen Variablen für die Evaluation der zu untersuchenden Politikmaßnahme zu verfügen. Mit der vorliegenden Arbeit wird die Vorgehensweise bei der Konzeption des Mikrosimulationsmodells EITDsim sowie die Fusion zweier repräsentativer Mikrodatensätze, der amtlichen Steuerstatistik und des Sozioökonomischen Panels, dokumentiert.

Die vorliegende Arbeit ist wie folgt gegliedert: In Abschnitt 2 werden die theoretischen Grundlagen der Mikrosimulation aufgezeigt und der Modul-Aufbau des integrierten Mikrosimulationsmodell EITDsim vorgestellt. Die einzelnen Module werden anschließend in den Abschnitten 3, 4 und 5 erörtert. Abschließende Bemerkungen und einen Ausblick auf weiterführende Forschungsprojekte beinhaltet Abschnitt 6.

## **2 Steuer-Transfer-Mikrosimulation mit EITDsim**

### **2.1 Grundlagen**

Seit Orcutt (1957) vor über 50 Jahren den (gedanklichen) Grundstein für einen neuen sozioökonomischen Modelltyp legte, hat sich die Verwendung von Mikrosimulationsmodellen (MSM) in den Sozial- und Wirtschaftswissenschaften aufgrund der hohen Flexibilität immens ausgebreitet.<sup>3</sup> MSM zählen zu den partial-analytischen Modellen und ermöglichen eine detaillierte Abbildung und (empirische) Analyse eines sozio-ökonomischen Systems sowie der darin befindlichen Mikroeinheiten. Insbesondere für die Finanzwissenschaft stellt die Mikro-

<sup>2</sup> Allerdings betrachtet EITDsim nur die Angebotsseite des Arbeitsmarktes. Es ist also kein allgemeines Gleichgewichtsmodell (CGE-Modell), weshalb keine Aussagen über Dritt-Rundeneffekte möglich sind.

<sup>3</sup> Baroni & Richiardi (2007) geben einen Überblick über die Entwicklung der MSM in verschiedenen Wissenschaftsbereichen in den letzten 50 Jahren.



simulation ein ideales Instrument für (ex ante) empirische Untersuchungen der Wirkungen von steuerpolitischen Reformmaßnahmen dar. Als grundlegende Voraussetzungen für die Methode der Mikrosimulation gelten dabei die Verfügbarkeit von repräsentativen Mikrodaten und die Vollständigkeit relevanter Variablen.

MSM können grundsätzlich in statische und dynamische Modellklassen unterschieden werden. Mit dynamischen Mikrosimulations- bzw. Arbeitsangebotsmodellen können die Langzeitwirkungen einer betrachteten Maßnahme auf den gesamten Lebenszyklus und z.B. den Aufbau von Humankapital oder die Ersparnis eines Untersuchungsobjektes analysiert werden. Aufgrund verschiedener Probleme (z.B. fehlende Variablen, Item-Nonresponse, Unit-Nonresponse) in Panel-Datensätzen spielen dynamische MSM in der empirischen Finanzwissenschaft immer noch eine untergeordnete Rolle. Einen Überblick über verschiedene Modelltypen und Anwendungsgebiete zeigen Blundell & MaCurdy (1999).

Mit statischen MSM wird dagegen anhand einer komparativ-statischen Analyse der kurzfristige Effekt (Erst-Rundeneffekt) von z.B. einer Steuerreform auf das Steueraufkommen oder die Verteilung der verfügbaren Einkommen gemessen. Die zentrale Annahme dieser Modellklasse lautet, dass die Untersuchungssubjekte vor und nach der simulierten Reformmaßnahme die selben Verhaltensmuster aufweisen. Diese eher restriktive Annahme einer vollkommen unelastischen Anpassungsreaktion kann für die meisten steuerlichen Maßnahmen in der kurzen Frist aber durchaus gerechtfertigt werden. Mithin ist es grundsätzlich plausibel, dass die Reaktion der Steuersubjekte auf z.B. die Abschaffung eines steuerlichen Ausnahmetatbestandes erst mit einer gewissen Verzögerung einsetzt, weil u.U. die konkreten Auswirkungen auf das verfügbare Einkommen unsicher sind und ein schnelles Handeln der Steuersubjekte unwahrscheinlich macht. Ebendieser mittelfristige Effekt (Zweit-Rundeneffekt) kann aber mit statischen Arbeitsangebotsmodellen qualitativ und quantitativ empirisch geschätzt werden. Zu diesem Zweck sind in den letzten 50 Jahren verschiedene Modelltypen konzipiert worden.<sup>4</sup> Das wesentliche Unterscheidungsmerkmal dieser statischen Arbeitsangebotsmodelle ist die Konstruktion der Budgetrestriktion. Demnach kann zwischen kontinuierlichen und diskreten Arbeitsangebotsmodellen differenziert werden.

Kontinuierlichen Arbeitsangebotsmodellen liegt die Annahme zugrunde, dass die Mikroeinheiten ihre Arbeitszeit frei wählen können. Für die, zur Schätzung der nutzenmaximalen

---

<sup>4</sup> Einen umfassenden Überblick bieten hierzu Blundell & MaCurdy (1999) oder Creedy & Duncan (2002).

Arbeitszeit notwendige, Budgetrestriktion existieren wiederum verschiedene Konstruktionsansätze: Burtless & Hausman (1978) bilden das Steuersystem möglichst exakt anhand einer stückweise linearisierten Funktion nach. Im Gegensatz dazu schlägt Hall (1973) eine vereinfachte Darstellung des Steuersystems anhand einer linearen Funktion vor, während Flood & MaCurdy (1992) das Steuersystem durch eine kontinuierliche und differenzierbare Funktion approximieren. Trotz dieser unterschiedlichen Ansätze sind die Schätzergebnisse von kontinuierlichen Arbeitsangebotsmodellen laut Creedy & Duncan (2002) ungenau oder verzerrt. Darüber hinaus wird der extensive Arbeitsangebotseffekt laut Brenneisen & Peichl (2007) nur unzureichend abgebildet und Van Soest et al. (1990) zufolge wird auch der Anteil der Teilzeitbeschäftigten tendenziell überschätzt.

Diskrete Arbeitsangebotsmodelle fußen auf der Annahme, dass die Mikroeinheiten ihre Arbeitszeit nicht frei, sondern aus einem endlichen Kontingent diskreter Arbeitszeiten wählen können. Diese grundlegende Annahme vereinfacht zum einen den Rechenaufwand bei der Ermittlung der Budgetrestriktion und zum anderen die Komplexität der Fragestellung, da nicht die gesamte Budgetrestriktion einer betrachteten Mikroeinheit sondern nur einzelne Punkte geschätzt werden müssen. Auch bei der Spezifikation der Nutzenfunktion bieten diskrete Arbeitsangebotsmodelle mehr Flexibilität, da erst ex post überprüft werden muss, ob die grundsätzlichen Anforderungen an die Nutzenfunktion erfüllt sind. Darüber hinaus ermöglichen diskrete Arbeitsangebotsmodelle auch die Analyse von extensiven Arbeitsangebotseffekten. Ein Nachteil diskreter Arbeitsangebotsmodelle ist sicherlich der Verzicht auf Informationen über die tatsächlich gearbeiteten Stunden zugunsten der Modellierung von diskreten Arbeitszeitkategorien. Der damit verbundene Rundungsfehler wiegt die Vorteile der diskreten gegenüber der kontinuierlichen Modellierung des Arbeitsangebots nach herrschender Meinung aber nicht auf.<sup>5</sup>

## 2.2 Datengrundlage von EITDsim

Elementarer Bestandteil eines MSM ist die zugrundeliegende Datenbasis. Die Entscheidung für oder gegen eine bestimmte (verfügbare) Datengrundlage ist zuvorderst durch die mit dem MSM zu beantwortende Forschungsfrage bestimmt. Darüber hinaus kann wiederum,

<sup>5</sup> Zum Beispiel Van Soest (1995, S. 83-84), Van Soest & Das (2000, S. 1-2, 26-28), Fuest et al. (2005, S. 24) oder Peichl et al. (2010, S. 14-15).

wenn z.B. relevante Informationen nicht verfügbar sind, die Datenbasis selbst die potentiellen Möglichkeiten eines MSM limitieren.

Aufgrund der Konzeption von EITDsim als Instrument zur Beantwortung bzw. Evaluation von insbesondere finanz- und sozialpolitisch motivierter Forschungsfragen wird als Primärdatenbasis in EITDsim die Faktisch Anonymisierte Lohn- und Einkommensteuerstatistik (FAST) 2004 verwendet.<sup>6</sup> Dieser Datensatz basiert auf einer geschichteten 10%-Zufallsstichprobe der Lohn- und Einkommensteuerstatistik aus dem Veranlagungsjahr 2004 und wird der Wissenschaft vom FDZ des Bundes und der Länder als Scientific Use File zur Verfügung gestellt. FAST weist ca. 3,5 Millionen Einzeldatensätze auf und stellt eine repräsentative Stichprobe der rund 35 Millionen Einkommensteuerpflichtigen dar. Ehepaare, die eine gemeinsame Veranlagung gewählt haben, werden jedoch als ein Fall ausgewiesen. Unter Berücksichtigung dieses Aspekts repräsentiert die FAST 2004 insgesamt rund 52,78 Millionen Personen.

FAST 2004 liefert detaillierte Informationen über steuerlich relevante Einkommensquellen, Sonderausgaben, außergewöhnliche Belastungen sowie einige (steuerlich relevante) soziodemographische Merkmale, wie Alter, Geschlecht, Religion, Kinderzahl und Familienstand. Diese informationelle Tiefe macht FAST besonders für MSM zur Beantwortung finanz- und steuerpolitischer Forschungsfragen interessant. Ein nicht unerheblicher Nachteil der FAST ist jedoch, dass alle Einkommensbestandteile, die nicht steuerpflichtig sind sowie Informationen über das soziale Umfeld und insbesondere das Arbeitsverhalten der Steuerpflichtigen nicht erfasst werden.<sup>7</sup>

Als Sekundärdatenbasis wird in EITDsim das Sozioökonomische Panel (SOEP) verwendet. Das SOEP ist eine repräsentative Wiederholungsbefragung privater Haushalte in Deutschland, die als anonymisierter Mikrodatsatz vom DIW Berlin für die wissenschaftliche Forschung bereitgestellt wird. Im für EITDsim relevanten Erhebungsjahr 2004 wurden 11.796 Haushalte mit 22.019 Menschen befragt.<sup>8</sup> Diese Stichprobe ist repräsentativ für rund 40

<sup>6</sup> Weil es bei der Veranlagung zur Einkommensteuer zu gewissen zeitlichen Verschiebungen kommen kann, erscheint die amtliche Statistik mit einer etwa fünf- bis sechsjährigen Verzögerung. Aufgrund dessen sind bis zum jetzigen Zeitpunkt keine aktuelleren Steuerdaten verfügbar.

<sup>7</sup> Vgl. Forschungsdatenzentrum (2008).

<sup>8</sup> Die Befragung wird mittels Interviews durchgeführt und erfordert teilweise eine Selbsteinschätzung der Teilnehmer. Insbesondere bei Fragen zur Einkommenshöhe bzw. steuerlich relevanten Themen können hier jedoch Verzerrungen entstehen.

Millionen Menschen in Deutschland, wobei jedoch besonders die unteren und mittleren Einkommensbereiche abgebildet werden. Es werden sowohl kontinuierlich Informationen über u.a. Persönlichkeitsmerkmale, Erwerbs- und Einkommensverläufe, Arbeitsverhalten, Haushaltszusammensetzungen als auch Informationen zu Schwerpunktthemen, wie z.B. Weiterbildung und Qualifikation, erhoben.<sup>9</sup> Das SOEP eignet sich aufgrund der Fülle an soziodemographischen bzw. -ökonomischen Informationen besonders für die Beantwortung von sozialpolitisch motivierter Forschungsfragen und wird in EITDsim insbesondere für die ökonomische Schätzung von Arbeitsangebotsreaktionen in Folge einer Politikmaßnahme verwendet.

Aufgrund des Panelcharakters und der Erhebung von Informationen zu Haushalten sowie Einzelpersonen unterscheidet sich das Datendesign des SOEP deutlich vom Aufbau der FAST. Informationen über gemeinsam veranlagte Ehepaare werden im SOEP, anders als in der FAST, nicht übergreifend als ein (Steuer-)Fall dargestellt. Es ist jedoch grundsätzlich möglich, den Ehepartner anhand einer eindeutigen Zuordnungsnummer zu identifizieren und die Informationen beider Einzeldatensätze zu kombinieren.

## 2.3 Modellaufbau

Eine schematische Darstellung der Grundstruktur von EITDsim zeigt die Abbildung 1. In Anbetracht der komplexen Modellierungsanforderungen, wird der Aufbau des gesamten Modells in drei Module unterteilt. Nach einer Aufbereitung der Daten sowie der Datensatzstruktur werden zunächst sowohl in der FAST als auch im SOEP die Einzeldatensätze mit Hilfe von Dummy-Variablen dahingehend kategorisiert bzw. selektiert, ob dem Haushalt ein (teilweise) flexibles oder unflexibles Arbeitsangebot unterstellt werden kann.<sup>10</sup> Dies geschieht vor dem Hintergrund, dass (1) die ökonomische Schätzung des Arbeitsangebotsverhaltens nur für Einzeldatensätze des SOEP mit (teilweise) flexiblen Arbeitsangebot durchgeführt wird und (2) nur diese ausgewählten SOEP-Einzeldatensätze später auf dem Wege der Datensatzfusion mit (möglichst) ähnlichen Einzeldatensätzen der FAST verknüpft werden sollen. Von der Datensatzfusion werden jedoch diejenigen SOEP-Einzeldatensätze aus-

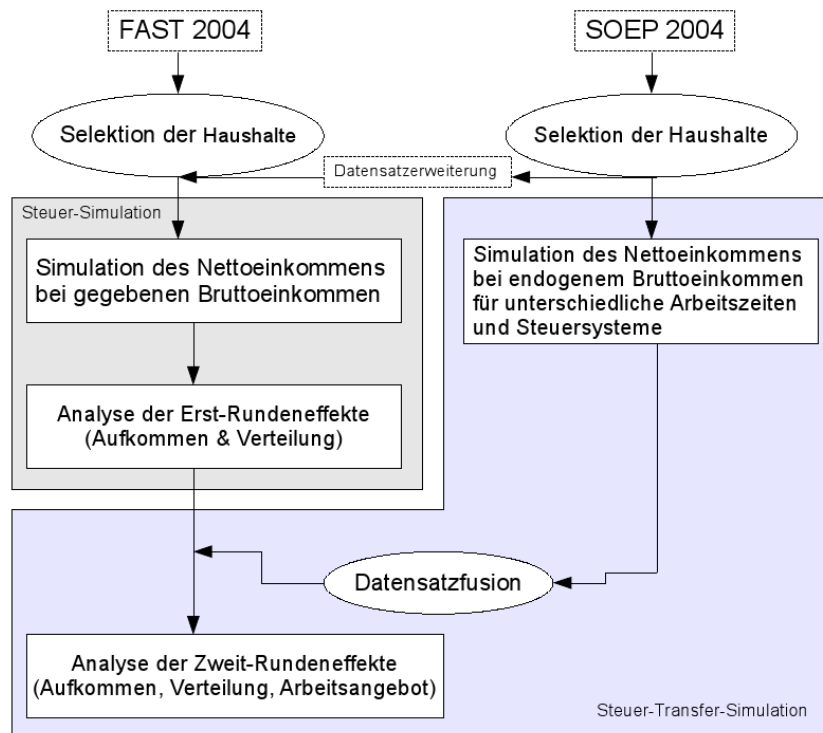
---

<sup>9</sup> vgl. Haisken-DeNew & Frick (2011).

<sup>10</sup> Näheres hierzu folgt in Abschnitt 4.2.

genommen, welche im Status quo erwerbslos sind.<sup>11</sup> Diese SOEP-Einzeldatensätze werden der FAST angehängen und erweitern so die Informationsbreite der Datenbasis.

**Abbildung 1: Modellaufbau EITDsim**



Quelle: Eigene Darstellung.

Das erste Modul enthält ein statisches Steuer-Simulationsmodell auf Basis der (erweiterten) FAST 2004 und dient zur Ermittlung der Aufkommens- und Verteilungseffekte einer potentiellen Steuerreform in der kurzen Frist. Die Simulation des aktuellen Steuersystems dient dabei als Referenzmodell für potentielle Steuerreformvorschläge. Dabei wird implizit angenommen, dass das Arbeitsangebot vollkommen unelastisch ist (Erst-Rundeneffekte). Das zweite Modul enthält ein Steuer-Transfer-Simulationsmodell sowie ein ökonometrisch geschätztes Arbeitsangebotsmodell auf Basis der (selektierten) SOEP-Daten. Das dritte Modul stellt eine Art "interaktive" Schnittstelle, mit der eine Verknüpfung der Module 1 und 2 hergestellt wird. Genauer gesagt werden über einen Fusionsalgorithmus die SOEP-Einzeldatensätze,

<sup>11</sup> Da der aktuelle Beschäftigungsstatus in der FAST nicht beobachtet werden kann, wird angenommen, dass alle in der FAST enthaltenen Mikroeinheiten mit flexiblem Arbeitsangebot auch tatsächlich erwerbstätig sind.

welchen ein (teilweise) flexibles Arbeitsangebot unterstellt werden kann und die im Status quo auch tatsächlich erwerbstätig waren, mit (möglichst) ähnlichen FAST-Einzeldatensätzen verbunden. Mit diesem (artifizialen) Mikrodatensatz ist es dann möglich, die Aufkommens-, Verteilungs- und Arbeitsangebotseffekte einer potentiellen Steuerreform in der mittleren Frist zu ermitteln (Zweit-Rundeneffekte). Es wird in EITDsim außerdem implizit angenommen, dass ein zusätzliches Arbeitsangebot auch tatsächlich nachgefragt wird.<sup>12</sup>

### **3 Das Steuer-Simulationsmodul in EITDsim**

Mit dem Steuer-Simulationsmodul von EITDsim ist es möglich die kurzfristigen Effekte einer Veränderung der steuerlichen Rahmenbedingungen auf das Steueraufkommen oder die empirische Einkommensverteilung zu simulieren. Das Modul kann jedoch keine Verhaltensanpassung der Steuerpflichtigen endogenisieren. D.h. es wird implizit unterstellt, dass die Steuerpflichtigen kurzfristig z.B. einer höheren Steuerbelastung nicht ausweichen bzw. das Arbeitsangebot vollkommen unelastisch ist. Diese starke Vereinfachung beschränkt die Ergebnisse der Analyse einer (potentiellen) Tarifreform dahingehend, dass nur Aussagen über Einkommenseffekte, sogenannte Erst-Rundeneffekte, möglich sind. Der für EITDsim verwendete Ansatz der Steuer-Simulation basiert größtenteils auf Struch (2012).

#### **3.1 Datenbasis**

Die Grundlage für die komparativ-statische Analyse einer (potentiellen) Steuerreform in EITDsim stellt der erweiterte Mikrodatensatz FAST 2004 dar. Diese erweiterte FAST setzt sich aus den Einzeldatensätzen der FAST 2004 und den als erwerbslos identifizierten SOEP-Einzeldatensätzen des Jahres 2004 zusammen.

#### **3.2 Modellierung des aktuellen Einkommensteuerrechts**

Der Aufbau des Steuer-Simulationsmodells orientiert sich am Berechnungsschema der tariflichen Einkommensteuer in § 2 EStG. Nach der Ermittlung des zu versteuernden Einkommens (zvE) erfolgt die Berechnung der Einkommensteuer nach § 32a EStG. Nach der

<sup>12</sup> EITDsim ist also kein allgemeines Gleichgewichtsmodell (CGE-Modell), da hierfür zusätzlich noch eine endogene Arbeitsnachfrage modelliert werden müsste.

Rechtslage 2012 wird die Einkommensteuer durch Anwendung des Formeltarifs in § 32a Abs. 1 EStG berechnet. Bis zu einem zvE von 8.004 Euro (Grundfreibetrag) muss keine Einkommensteuer gezahlt werden. Bei einem zvE von 8.005 Euro beträgt der Eingangsteuersatz 14%. Innerhalb dieser ersten Progressionszone steigt die Grenzbelastung bis auf knapp 24% bei einem zvE von 13.469 Euro stark an. In der anschließenden zweiten Progressionszone steigt die Grenzbelastung langsamer, aber stetig bis auf knapp 42% bei einem zvE von 52.881 Euro an. Innerhalb der folgenden ersten Proportionalzone (ab einem zvE von 52.882 Euro) und bis zu einem zvE von 250.730 Euro beträgt die Grenzbelastung konstant 42% und springt danach auf den Spitzensteuersatz von 45%. Die Durchschnittsbelastung steigt mit dem zvE, wobei die Steigung der Funktion gegen Null strebt, und nähert sich dem Spitzensteuersatz asymptotisch an. Ehepaare können auf Antrag eine gemeinsame Veranlagung wählen. Die Einkommensteuer wird in diesem Fall nach § 32a Abs. 5 EStG (Splittingverfahren) ermittelt: Die Einkommensteuer beträgt das Doppelte des Steuerbetrags, der sich für die Hälfte des gemeinsamen zvE nach Anwendung des Formeltarifs in § 32a Abs. 1 EStG ergibt. Für eine korrekte Erfassung der Einkommensteuer werden außerdem folgende Vorschriften berücksichtigt:

- § 32 EStG Freibeträge für Kinder (Günstigerprüfung)
- § 32b EStG Progressionsvorbehalt
- § 32d EStG Gesonderter Tarif für Einkünfte aus Kapitalvermögen (Abgeltungsteuer)

Bei Inanspruchnahme des Kinderfreibetrags erfolgt eine Hinzurechnung des Kindergeldes zur ermittelten Einkommensteuer. Ferner wird im Modell angenommen, dass jeder Steuerpflichtige die Veranlagungsoption nach § 32d Abs. 6 EStG immer dann wählt, wenn die Abgeltungsteuer höher als die Steuer ist, die sich ergeben würde, wenn die Einkünfte aus Kapitalvermögen gemäß § 32a EStG besteuert werden würden.

Das (ökonomische) Bruttoeinkommen wird als Aggregat aus den einzelnen Bruttoeinkünften sowie den Einkünften, die dem Progressionsvorbehalt des § 32b EStG unterliegen, definiert. Hinzu addiert werden außerdem Transferzahlungen, wie z.B. Kindergeld und Leistungen nach dem SGB II. Das (verfügbare) Nettoeinkommen ergibt sich nach Abzug von Sozialversicherungsabgaben, der tariflichen Einkommensteuer, der Abgeltungsteuer und des Solidaritätszuschlags vom (ökonomischen) Bruttoeinkommen.

## 4 Das Arbeitsangebotsmodul von EITDsim

Mit dem Arbeitsangebotsmodul von EITDsim ist es möglich Verhaltensanpassungen der Haushalte auf eine Veränderung der steuerlichen Rahmenbedingungen zu simulieren. Der für EITDsim verwendete Ansatz basiert auf dem neoklassischen Lehrbuchmodell, wonach ein Individuum bei der Wahl seines Arbeitsangebots vor einem Trade-off zwischen Freizeit und Einkommen bzw. Konsum steht. Das Individuum maximiert seinen Nutzen, welcher positiv vom Einkommen bzw. Konsum und negativ von der Arbeitszeit abhängt, unter den Nebenbedingungen, dass das gesamte Einkommen verkonsumiert wird und außerdem nur eine begrenzte Freizeitausstattung möglich ist. Die Lösung des Problems ergibt für das Individuum ein optimales Bündel aus Freizeit und Einkommen sowie das dafür notwendige Arbeitsangebot.

Mit EITDsim wird zunächst das nutzenmaximierende Arbeitsangebot unter Maßgabe des aktuellen Steuer- und Transfersystems mit einem statisch diskreten Haushaltsarbeitsangebotsmodell nach Van Soest (1995) ökonometrisch geschätzt. Zur Identifizierung von Arbeitsangebotsreaktionen der Haushalte auf eine Veränderung des Steuerregimes werden dann, einem Ansatz von Bonin & Schneider (2006) folgend, exakte Übergangswahrscheinlichkeiten zwischen den Arbeitszeitkategorien berechnet.

### 4.1 Modell

Als theoretische Grundlage dient das diskrete Haushaltsarbeitsangebotsmodell von Van Soest (1995). Van Soest nimmt an, dass die Entscheidung der Haushaltsmitglieder über ihre Arbeitszeit nicht (völlig) beliebig sei. Aus diesem Grund könne man die Wahl der Arbeitszeit statt als kontinuierliche Variable auch als kategoriale Variable modellieren, d.h. die Arbeitsangebotsentscheidung könne als Wahl zwischen einer bestimmten Anzahl von Arbeitszeitalternativen betrachtet werden. Mit diskreten Auswahlmodellen wird die bedingte Wahrscheinlichkeit geschätzt, dass ein Haushalt  $i$  aus  $K$  Alternativen immer dann die Alternative  $k$  wählt, wenn der damit einhergehende Nutzen  $U_{ik}$  größer als der Nutzen aller anderen Alternativen  $l$  ist:

$$P_{ik} = P(U_{ik} > U_{il}) \quad \forall l = 1, \dots, K, l \neq k \quad (1)$$



Dem Prinzip der stochastischen Nutzenmaximierung von McFadden (1974) folgend, kann der Nutzen  $U_{ik}$  in einen deterministischen (direkt beobachtbaren) Teil  $V_{ik}$  und einen zufälligen (unbeobachtbaren) Teil  $\varepsilon_{ik}$  aufgespalten werden.

$$U_{ik} = V_{ik} + \varepsilon_{ik} \quad (2)$$

Der direkte Nutzen  $V_{ik}$  ist eine Funktion von Nettoeinkommen und Freizeit sowie weiteren Merkmalen des Haushaltes. Der stochastische Teil  $\varepsilon_{ik}$  enthält alle unbeobachteten Faktoren, die den Nutzen des Haushaltes beeinflussen. Durch Einsetzen von (2) in (1) und Umformen kann die bedingte Wahrscheinlichkeit, dass Haushalt  $i$  die Alternative  $k$  allen anderen Alternativen vorzieht wie folgt dargestellt werden:

$$P(U_{ik} > U_{il}) = P(V_{ik} - V_{il} > \varepsilon_{il} - \varepsilon_{ik}). \quad (3)$$

Unter der Annahme, dass die Fehlerterme  $\varepsilon_{ik}$  über alle Haushalte  $i$  und Arbeitszeitkategorien  $K$  unabhängig und identisch verteilt sind und einer Extremwertverteilung vom Typ I (Gumbel) folgen, kann die bedingte Auswahlwahrscheinlichkeit dann für jede mögliche Arbeitszeitkategorie über ein Conditional-Logit-Modell nach McFadden (1974) berechnet werden:

$$P(U_{ik} > U_{il}) = \frac{\exp(V_{ik})}{\sum_{l=1}^K \exp(V_{il})}. \quad (4)$$

Das Conditional-Logit-Modell beinhaltet die Annahme, dass die relative Auswahlwahrscheinlichkeit zweier Alternativen unabhängig von irgendeiner anderen Alternative ist.<sup>13</sup> Obwohl diese Annahme (theoretisch) recht problematisch ist, insbesondere bei der Modellierung von diskreten Arbeitszeitkategorien, wird nachfolgend, wie in der Literatur davon ausgegangen, dass diese Annahme erfüllt ist. Ferner impliziert Gleichung (4), dass die Haushalte keiner Beschränkung bei der Arbeitszeitentscheidung unterliegen. D.h. es kann in dieser Modell-Spezifikation keine unfreiwillige Arbeitslosigkeit vorliegen.

Um das nutzenmaximierende Verhalten der Haushalte abbilden zu können, werden die Parameter der (direkten) Nutzenfunktion  $V_{ik}$  empirisch geschätzt. Dabei wird im Folgenden angenommen, dass Ehepartner eine gemeinsame Nutzenfunktion maximieren.<sup>14</sup> In der Li-

<sup>13</sup> Independence of irrelevant alternatives (IIA).

<sup>14</sup> Sog. Household Utility oder Unitary Model.

teratur finden sich verschiedene Ansätze der funktionalen Spezifikation für den zu schätzenden (direkten) Haushaltsnutzen  $V_{ik}$ .<sup>15</sup> Notwendige Bedingung für eine sinnvolle ökonomische Spezifikation ist jedoch, dass der Grenznutzen von Einkommen und Freizeit positiv, aber abnehmend ist.<sup>16</sup> In EITDsim wird der (direkte) Nutzen eines Haushaltes  $i$  bei Wahl einer bestimmten Arbeitszeitkategorie  $k$  mit einer Translog-Funktion modelliert:

$$V_{ik}(x_{ik}) = x'_{ik}Ax_{ik} + \beta'x_{ik}, \quad (5)$$

wobei  $x = (\ln ekn, \ln lm, \ln lf)'$ .<sup>17</sup> Die (beobachtbaren) Elemente von  $x$  sind das Nettohaushaltseinkommen ( $ekn$ ), die Freizeit des Mannes ( $lm$ ) und die Freizeit der Frau ( $lf$ ). In die Nutzenfunktion gehen die Elemente von  $x$  sowohl in linearer und quadratischer Form als auch in Form von Kreuztermen ein. Die Matrix  $A$  mit den Elementen  $\alpha_{ij}$ ,  $ij = (1, 2, 3)$ , beinhaltet die (zu schätzenden) Koeffizienten der quadratischen Terme und der Kreuzterme. Der Vektor  $\beta_j$ ,  $j = (1, 2, 3)$  die (zu schätzenden) Koeffizienten der linearen Terme.

Statt in Matrix-Schreibweise kann (5) umgeformt auch wie folgt dargestellt werden:

$$V_{ik}(x_{ik}) = \beta_1 \ln ekn_{ik} + \beta_2 \ln lm_{ik} + \beta_3 \ln lf_{ik} + \alpha_{11}(\ln ekn_{ik})^2 + \alpha_{22}(\ln lm_{ik})^2 + \alpha_{33}(\ln lf_{ik})^2 + \alpha_{12} \ln ekn_{ik} \ln lm_{ik} + \alpha_{13} \ln ekn_{ik} \ln lf_{ik} + \alpha_{23} \ln lm_{ik} \ln lf_{ik}. \quad (6)$$

Ferner wird die (beobachtbare) Heterogenität der Haushalte wie bei Van Soest (1995) durch folgende Spezifizierung der Parameter  $\beta_m$ ,  $\alpha_{mn}$  berücksichtigt:

$$\beta_m = \beta_{m0} + \sum_{p=1}^P \beta_{mp}z_p \quad (7)$$

$$\alpha_{mn} = \alpha_{mn0} + \sum_{p=1}^P \alpha_{mnp}z_p, \quad (8)$$

wobei  $m, n = 1, 2, 3$  und  $z_p$  ( $p = 1, \dots, P$ ) die Fall-spezifischen Kontrollvariablen darstellen. Dazu zählen z.B. Alter, Nationalität, Schulabschluss, Gesundheitszustand, Anzahl und Alter

<sup>15</sup> Stern (1986) gibt einen Überblick über verschiedene Spezifikationen von direkten Nutzenfunktionen für Arbeitsangebotsfunktionen.

<sup>16</sup> Diese Bedingung muss allerdings erst ex post überprüft werden.

<sup>17</sup> Der entscheidende Vorteil dieser Spezifikation gegenüber einer Cobb-Douglas oder CES-Nutzenfunktion liegt in der Flexibilität. D.h. durch die Modellierung von Kreuztermen ist es möglich, die Abhängigkeit von z.B. dem Freizeitnutzen eines Haushaltsmitglieds auf den Freizeitnutzen seines Partners zu berücksichtigen.

der Kinder im Haushalt sowie die Wohnregion (Ost/West).

Nach Schätzung der Parameter der spezifizierten Nutzenfunktion muss geprüft werden, ob das geschätzte Haushaltsnutzen-Modell auch mit der zugrunde liegenden ökonomischen Theorie übereinstimmt. Wie oben ausgeführt, muss der Grenznutzen der Freizeit sowie des Einkommens positiv, aber abnehmend sein.<sup>18</sup> D.h. unter der Voraussetzung, dass Freizeit für die Haushaltsmitglieder ein normales Gut darstellt, sollte die erste Ableitung der Nutzenfunktion nach der Freizeit eines Haushaltsmitgliedes positiv und die zweite Ableitung negativ sein. Darüber hinaus ist das Vorzeichen des Substitutionseffektes der Freizeit der Haushaltsmitglieder untereinander bzw. von Freizeit und Einkommen der Haushaltsmitglieder jedoch theoretisch unbestimmt.

## 4.2 Datensatz, Selektion und Variable

Als Datenbasis für die ökonometrische Schätzung und Simulation des Arbeitsangebotsverhaltens dient in EITDsim das Sozioökonomische Panel (SOEP) des DIW Berlin.<sup>19</sup> Im für die Simulation des Arbeitsangebotsverhaltens relevanten Erhebungsjahr 2004 wurden 11.796 Haushalte mit 22.019 Menschen befragt.

Wie in der Literatur üblich, wird in einem ersten Schritt die Schätzung des Arbeitsangebotes auf jene Personen beschränkt, denen ein flexibles Arbeitsangebot unterstellt werden kann.<sup>20</sup> Dazu gehören sowohl alle abhängig Beschäftigten als auch alle Arbeitslosen. Ein unflexibles Arbeitsangebot wird Personen unterstellt, die eines der folgenden Merkmale aufweisen:

- Personen, die jünger als 16 oder älter als 65 Jahre sind
- Beamte
- Bezieher von Altersrente, Altersübergangs- oder Vorruhestandsgeld
- Auszubildende und Studenten
- Frauen im Mutterschutz
- Wehr- oder Zivildienstleistende
- Selbstständige

<sup>18</sup> Vgl. hierzu Steiner & Wrohlich (2004) und Brenneisen & Peichl (2007).

<sup>19</sup> Ausführliche Informationen zum SOEP geben Haisken-DeNew & Frick (2011).

<sup>20</sup> Vgl. Van Soest & Das (2000), Steiner & Wrohlich (2004), Fuest et al. (2005) oder Franz et al. (2007).

Der zweite Schritt stellt die Zuordnung der Personen in eine bestimmte Haushaltskategorie dar. Es werden dafür in Abhängigkeit der Arbeitsangebotsflexibilität drei Haushaltstypen definiert.<sup>21</sup> Der erste Haushaltstyp wird als “flexibler Haushalt” bezeichnet, d.h. sowohl dem Haushaltsvorstand als auch seinem Partner kann ein flexibles Arbeitsangebot unterstellt werden. Der zweite Haushaltstyp wird als “gemischter Haushalt” charakterisiert, wenn entweder dem Haushaltsvorstand oder seinem Partner ein flexibles Arbeitsangebot unterstellt werden kann. Den dritten Haushaltstyp, der sogenannte “unflexible Haushalt”, kennzeichnet, dass neben dem Haushaltsvorstand auch sein Partner ein unflexibles Arbeitsangebot im obigen Sinne aufweist. Haushalte, die der Kategorie des “unflexiblen Haushaltes” zugeordnet werden können, werden aus der Schätzstichprobe entfernt. Darüber hinaus wird bei der Modellierung zwischen Single- und Paarhaushalten unterschieden. Zwei Personen zählen nur dann als Paarhaushalt, wenn sie zusammen veranlagt werden. “Gemischte Haushalte” verbleiben in der Schätzstichprobe, jedoch wird der Partner mit dem flexiblen Arbeitsangebot wie ein “technischer” Single-Haushalt behandelt.

Die Schätzstichprobe umfasst (ohne Berücksichtigung der Hochrechnungsfaktoren) insgesamt 2.806 Paar-Haushalte und 5.470 (technische) Single-Haushalte, welche wiederum in 2.386 Single-Männer und 3.083 Single-Frauen unterschieden werden können.

#### 4.2.1 Budgetrestriktion des Haushaltes

Um den geschlechtsspezifischen Unterschieden bei der tatsächlichen Wochenarbeitszeit Rechnung zu tragen, wird für Männer und Frauen eine unterschiedliche Anzahl möglicher Arbeitszeitkategorien ( $k$ ) modelliert. Dies bedeutet, dass Männer zwischen drei Alternativen wählen können. Die Modellierung der Arbeitszeitkategorien für Männer ist in Tabelle 1 dargestellt.

**Tabelle 1:** Arbeitszeitkategorien - Männer

$k$	Wochenarbeitszeit (in Stunden)	
0	0	Arbeitslos
1	1 - 40	Voll erwerbstätig
2	> 40	Überstunden

<sup>21</sup> Ähnlich gehen Fuest et al. (2005) oder Steiner & Wakolbinger (2009) vor.

Bei Frauen wird angenommen, dass sie sich zwischen fünf Alternativen entscheiden. Die Modellierung der Arbeitszeitkategorien für Frauen ist in Tabelle 2 dargestellt.

**Tabelle 2:** Arbeitszeitkategorien - Frauen

$k$	Wochenarbeitszeit (in Stunden)	
0	0	Arbeitslos
1	1 - 15	Teilzeit erwerbstätig I
2	16 - 34	Teilzeit erwerbstätig II
3	35 - 40	Vollzeit erwerbstätig
4	> 40	Überstunden

Um die Arbeitsangebotsentscheidung von Paar-Haushalten modellieren zu können, wird angenommen, dass diese Haushalte eine gemeinsame Nutzenfunktion maximieren. Zu diesem Zweck werden die geschlechtsspezifischen Arbeitszeitkategorien für Männer und Frauen in Arbeitszeitkombinationen für Paare überführt. Dieses Vorgehen führt zu insgesamt 15 verschiedenen Arbeitszeitkombinationen, aus denen ein Paar-Haushalt das gemeinsame Arbeitsangebot wählen kann.

Van Soest (1995) folgend, wird die Freizeitausstattung eines Haushalts  $i$  für jede korrespondierende Arbeitszeitkategorie  $k$  gemäß folgender Regel bestimmt:

$$lm_{ik} = TE - hm_{ik} \text{ bzw. } lf_{ik} = TE - hf_{ik},$$

wobei  $lm$  ( $lf$ ) für die Freizeitausstattung des Mannes (der Frau),  $TE$  für das Zeitbudget pro Woche und  $hm$  ( $hf$ ) für die Arbeitszeit des Mannes (der Frau) stehen. Es wird angenommen, dass die gesamte Zeitausstattung  $TE$  eines Haushaltsmitgliedes pro Woche 80 Stunden beträgt.<sup>22</sup> Die Arbeitszeiten  $hm$  bzw.  $hf$  werden definiert als Produkt aus der Arbeitszeitkategorie  $k$  und einem bestimmten Intervall  $IL_g$ ,  $g \in \{m, f\}$ :

$$hm_{ik} = k \cdot IL_m \text{ bzw. } hf_{ik} = k \cdot IL_f.$$

Aufgrund der unterschiedlichen Anzahl an Arbeitszeitkategorien für Männer und Frauen, wird für Männer eine Intervalllänge von  $IL_m = 24$  und für Frauen eine Intervalllänge von  $IL_f = 12$  definiert.

<sup>22</sup> Van Soest (1995), Van Soest & Euwals (1999) und Steiner & Wrohlich (2004) zeigen, dass alternative Zeitbudgetwerte, z.B.  $TE = 60$ , kaum Einfluss auf die Schätzergebnisse haben.

Für die Simulation des Nettohaushaltseinkommens wird in EITDsim ein Steuer-Transfer-Modell, das alle relevanten Komponenten des EStG beinhaltet, verwendet. Der vorgenommenen Selektion des Datensatzes entsprechend, stellen die Einkünfte aus abhängiger Beschäftigung für die Mehrzahl der betrachteten Haushalte die Haupteinkommenskomponente dar. Die (beobachtbare) Information über das Bruttomonatseinkommen wird, zusammen mit der (beobachtbaren) Information über die Wochenarbeitszeit des Haushaltsmitglieds, benutzt, um einen Bruttostundenlohn zu ermitteln. Bei der Simulation des (hypothetischen) Bruttomonatseinkommens eines Haushaltsmitgliedes für jede Arbeitszeitkategorie  $k$  wird der ermittelte Bruttostundenlohn mit der durchschnittlichen Wochenarbeitszeit jeder Arbeitszeitkategorie  $k$  multipliziert. Bei Ehepaaren wird das für jeden Ehepartner ermittelte (hypothetische) Bruttomonatseinkommen einer Arbeitszeitkategorie zu einem gemeinsamen Bruttohaushaltseinkommen aufsummiert. Wie in der Literatur üblich wird dabei implizit angenommen, dass der Bruttostundenlohn unabhängig von der tatsächlichen Arbeitszeit ist.<sup>23</sup> Für Haushaltsmitglieder, die entweder erwerbslos sind oder bei denen keine Informationen über den Bruttostundenlohn vorliegen, wird mit einem zweistufigen Selektionsmodell ein (hypothetischer) Bruttostundenlohn geschätzt. Zu diesem Zweck wird eine Lohn- bzw. Partizipationsgleichung durch eine Regression empirisch geschätzt.<sup>24</sup> Die Schätzergebnisse sind in der Tabelle 14 im Appendix abgetragen.

Für die Simulation des Nettoeinkommens werden ausgehend vom (hypothetischen) Bruttomonatseinkommen eines Haushaltes in der Kategorie  $k$  die Sozialversicherungsbeiträge des Arbeitnehmers, die simulierte Einkommensteuer sowie der Solidaritätszuschlag abgezogen und (potentielle) staatliche Transferzahlungen hinzu addiert. Die Simulation der Einkommensteuer orientiert sich am Berechnungsschema der tariflichen Einkommensteuer in § 2 EStG. Nach Ermittlung des zu versteuernden Einkommens (zvE) erfolgt die Berechnung der Einkommensteuer entsprechend § 32a EStG.

Für eine korrekte Erfassung der Einkommensteuer werden außerdem folgende Vorschriften berücksichtigt:

- § 32 EStG Freibeträge für Kinder (Günstigerprüfung)
- § 32d EStG Besonderer Tarif für Einkünfte aus Kapitalvermögen (Abgeltungsteuer)

<sup>23</sup> Vgl. Van Soest (1995), Van Soest & Das (2000) oder Fuest et al. (2005).

<sup>24</sup> Vgl. hierzu Heckman (1976) und Heckman (1979).

Bei Inanspruchnahme des Kinderfreibetrags erfolgt eine Hinzurechnung des Kindergeldes zur ermittelten Einkommensteuer. Ferner wird im Modell angenommen, dass jeder Steuerpflichtige die Veranlagungsoption nach § 32d Abs. 6 EStG immer dann wählt, wenn die Abgeltungsteuer höher als die Steuer ist, die sich ergeben würde, wenn die Einkünfte aus Kapitalvermögen gemäß § 32a EStG besteuert würden.

Die Modellierung des staatlichen Transferanspruchs bei Erwerbslosigkeit umfasst in EITDsim neben dem o.g. Kindergeld auch das Arbeitslosengeld II, das Kindersozialgeld sowie die Zulagen für Unterkunft und Heizung. Der Transferanspruch verringert sich durch einen Hinzuverdienst wie folgt: Nach aktueller Rechtslage bleiben die ersten 100 Euro anrechnungsfrei. Zwischen 100 und 1.000 Euro werden 80 Prozent und zwischen 1.000 und 1.200 Euro 90 Prozent vom Nettoeinkommen auf den Transferanspruch angerechnet.

#### **4.2.2 Deskriptive Analysen der Variablen**

Tabelle 3 zeigt die Erwerbsquoten für die verschiedenen Gruppen der Schätzstichprobe bzw. der Grundgesamtheit. Die Erwerbsquote in der Grundgesamtheit (GG) beträgt 56,78%.<sup>25</sup> Die Erwerbsquote für Männer (Frauen) innerhalb der Schätzstichprobe liegt bei 78,41% (62,80%) und damit nur marginal unter den Werten, die das Statistische Bundesamt für 2004 ermittelte.<sup>26</sup> Hinsichtlich der Veranlagungsart können in der GG keine gravierenden Unterschiede in der Erwerbsquote festgestellt werden. In der Schätzstichprobe unterscheidet sich der Anteil der erwerbstätigen Single-Männer nur geringfügig vom Anteil der erwerbstätigen verheirateten Männer, wohingegen in der GG eine deutlich höhere Erwerbsquote für verheiratete Männer festgestellt werden kann. Auffällig ist weiterhin, dass vor allem in der Schätzstichprobe aber auch in der GG der Anteil der Erwerbstätigen bei den Single-Frauen höher ist als bei verheirateten Frauen. Eine naheliegende Erklärung hierfür wäre, dass der potentielle Arbeitslohn von verheirateten Frauen durch das Splitting-Verfahren (bei Erwerbstätigkeit des Mannes) mit einem hohen Grenzsteuersatz belastet wird und diese Frauen dann eher von einer Arbeitsaufnahme absehen.

<sup>25</sup> Die Erwerbsquote gibt das Verhältnis der erwerbstätigen Personen zur Gesamtbevölkerung im erwerbsfähigen Alter (16-65 Jahre) an.

<sup>26</sup> Die vom Statistischen Bundesamt (2006) ermittelte Erwerbsquote bezogen auf die männlichen (weiblichen) Erwerbspersonen beträgt 79,3% (65,2%).

**Tabelle 3: Erwerbsquoten**

	Paare		(Techn.) Singles		GG
	Männer	Frauen	Männer	Frauen	
ZV	81,68%	57,05%	73,83%	50,10%	56,51%
EV	-	-	75,43%	75,20%	57,09%
GG	74,81%	50,70%	58,30%	55,35%	56,78%

Anmerkungen: ZV = Zusammenveranlagung, EV = Einzelveranlagung.

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Tabelle 4 gibt einen Überblick über die Verteilung der Arbeitszeitkategorien der (technischen) Single-Haushalte in der Schätzstichprobe. Fast ein Viertel aller Single-Männer ist nicht erwerbstätig. Der überwiegende Teil (rund 70%) der Single-Männer arbeitet zwischen 1 und 40 Stunden pro Woche, wobei die durchschnittliche Wochenarbeitszeit rund 38 Stunden beträgt. Nur etwas mehr als 5% haben eine reale Wochenarbeitszeit von mehr als 40 Stunden. Im Mittel arbeiten diese Single-Männer rund 49 Stunden pro Woche.

**Tabelle 4: Durchschnittliche Wochenarbeitszeit und relative Verteilung der Single-Haushalte auf die Arbeitszeitkategorien**

(a) Männer					
Arbeitszeitkategorie $k$	0	1	2		
Wochenarbeitszeit	0	1 ≤ 40	> 40		
∅ Wochenarbeitszeit	0	37,65	48,92		
Relativer Anteil	24,57%	69,95%	5,48%		

(b) Frauen					
Arbeitszeitkategorie $k$	0	1	2	3	4
Wochenarbeitszeit	0	1 ≤ 15	16 ≤ 34	35 ≤ 40	> 40
∅ Wochenarbeitszeit	0	10,19	25,03	38,41	48,46
Relativer Anteil	24,79%	2,68%	15,36%	54,97%	2,20%

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Ähnlich den Single-Männern, ist fast jede vierte Single-Frau in der Schätzstichprobe nicht erwerbstätig. Ungefähr 18% der Single-Frauen arbeiten in Teilzeit, d.h. nicht mehr als 34 Stunden pro Woche. Für rund 55% aller Single-Frauen beträgt die Wochenarbeitszeit wenigstens 35 und höchsten 40 Stunden. Nur etwas mehr als 2% der Single-Frauen arbeiten (tatsächlich) mehr als 40 Stunden pro Woche.



**Tabelle 5:** Relative Verteilung der Paar-Haushalte auf die Arbeitszeitkategorien

		Frauen					Summe		
		Arbeitszeitkategorie $k$		0	1	2		3	4
		Wochenarbeitszeit		0	$1 \leq 15$	$16 \leq 34$		$35 \leq 40$	$> 40$
Männer	0	0	21,53%	3,37%	8,84%	7,28%	0,56%	41,58%	
	1	$1 \leq 40$	24,65%	6,24%	11,63%	9,48%	0,34%	52,34%	
	2	$> 40$	3,02%	0,63%	1,33%	1,01%	0,07%	6,06%	
Summe			49,20%	10,24%	21,80%	17,77%	0,97%	100%	

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Tabelle 5 zeigt die Verteilung der Arbeitszeitkategorien für verheiratete Paare in der Schätzstichprobe. In fast 22% aller Paar-Haushalte der Schätzstichprobe sind beide Partner nicht erwerbstätig. Wie schon in Tabelle 3 gezeigt wurde, gibt es in Paar-Haushalten die Tendenz zum Allein-Verdienermodell, wobei der Anteil der erwerbstätigen Ehemänner (58,42%) im Vergleich deutlich größer ist als der Anteil der erwerbstätigen Ehefrauen (50,80%). Im Vergleich zu den Single-Frauen arbeiten auch relativ mehr Ehefrauen in Teilzeit, d.h. zwischen 1 und 15 bzw. 16 und 34 Stunden pro Woche, statt in Vollzeit-Arbeitsverhältnissen.

### 4.3 Spezifizierung des Schätzmodells

Zur ökonometrischen Schätzung der Parameter der (direkten) Nutzenfunktion des Haushaltes wird ein Conditional-Logit-Modell nach McFadden (1974) spezifiziert. Auf das Vorliegen von beobachtbarer Heterogenität wird mittels folgender Fall-spezifischer (Dummy-)Variablen kontrolliert: Alter, Gesundheitszustand, Nationalität, Schulabschluss, Anzahl und Alter der Kinder im Haushalt und Wohnregion (Ost/West).

Die Spezifikation des Schätzmodells wird für jede Schätzstichprobe (Paar-Haushalte, Single-Männer/-Frauen) getrennt vorgenommen. Dabei wird implizit angenommen, dass alle Beobachtungseinheiten innerhalb einer Schätzgruppe ein vergleichbares Einkommens- bzw. Freizeitnutzenkalkül aufweisen. Der Nutzen eines Haushaltes in jeder möglichen Arbeitszeitkategorie wird anschließend durch Einsetzen der geschätzten Parameter und der Variablenwerte in Gleichung (6) bestimmt. Für jeden Haushalt können die bedingten Auswahlwahrscheinlichkeiten der jeweiligen Arbeitszeitkategorien gemäß Gleichung (4) ermittelt werden. Zur Identifizierung von Arbeitsangebotsreaktionen der Haushalte auf eine Veränderung des Steuer- oder Transfersystems werden, einem Ansatz von Bonin & Schneider (2006) folgend,

exakte Übergangswahrscheinlichkeiten zwischen den Arbeitszeitkategorien berechnet.

#### 4.3.1 Schätzergebnisse

Die oben spezifizierte Nutzenfunktion wird in allen drei Schätzgruppen mit der Maximum-Likelihood Methode geschätzt. Die Tabellen 6 und 7 zeigen die geschätzten Parameter der direkten Nutzenfunktion. Aufgrund der Logarithmierung und der Modellierung von Kreuz- sowie Interaktionstermen von Einkommen und Freizeit können die geschätzten Koeffizienten jedoch nicht als marginale Effekte interpretiert werden. Die Interaktionen von Einkommen  $ekn$  mit den sozioökonomischen Merkmalen  $z$  sind in den beiden Single-Schätzgruppen mit nur einer Ausnahme signifikant: Für ledige Männer und Frauen mit Abitur hat das Nettoeinkommen bei der Arbeitszeitwahl *ceteris paribus* einen niedrigeren Stellenwert als bei denen ohne Abitur. Ferner hat für ledige Männer und Frauen in den neuen Bundesländern das Nettoeinkommen *ceteris paribus* ein höheres Gewicht bei der Arbeitszeitsentscheidung als bei denen in den alten Bundesländern.

Von den Interaktionstermen von Freizeit  $l$  und den sozioökonomischen Merkmalen  $z$  sind in der Schätzgruppe der Single-Männer v.a. das Alter relevant. Mit zunehmenden Alter steigt der Wert der Freizeit. Bei Single-Frauen spielt das Alter bei der Freizeitabwägung eine ähnliche Rolle wie bei Single-Männern. Ferner erhöht das Vorhandensein von Kindern im Haushalt den Wert der Freizeit für Single-Frauen. Je jünger die Kinder, desto stärker ist dieser Effekt. Ferner legen Single-Frauen in den neuen Bundesländern mehr Wert auf Freizeit als Single-Frauen aus den alten Bundesländern. Und bei Single-Frauen mit Abitur ist der Freizeitnutzen *ceteris paribus* geringer als bei denen ohne Abitur.

Tabelle 7 zeigt, dass in der Schätzgruppe der zusammenveranlagten Paare die Interaktionen von Einkommen  $ekn$  mit den sozioökonomischen Merkmalen  $z$  nur in zwei Fällen signifikant sind. Für Ehepaare in den neuen Bundesländern spielt bei der Arbeitsangebotsentscheidung das gemeinsame Nettoeinkommen im Vergleich zu den Ehepaaren in den alten Bundesländern *ceteris paribus* eine stärkere Rolle. Darüber hinaus hat für verheiratete Männer mit Abitur das Nettoeinkommen bei der Arbeitszeitwahl *ceteris paribus* einen niedrigeren Stellenwert als bei denen ohne Abitur. Die Interaktionen von Freizeit  $lm$  bzw.  $lf$  mit den sozioökonomischen Merkmalen  $z$  sind alle signifikant und haben die gleichen Vorzeichen wie in den Single-Schätzgruppen.

**Tabelle 6:** Parameter der Arbeitsangebotsschätzung - Singles

Variable	Männer		Frauen	
	Koeffizient	Standardfehler	Koeffizient	Standardfehler
<i>Einkommen</i>				
ln ekn	-8,851**	4,068	-27,320***	4,504
(ln ekn) <sup>2</sup>	0,553**	0,222	2,076***	0,260
ln ekn × Ost	0,488*	0,259	1,165***	0,375
ln ekn × Abitur	-1,164***	0,280	-2,007***	0,346
ln ekn × Deutsch	-0,149	0,472	0,425	0,679
<i>Freizeit</i>				
ln l	74,870***	5,125	33,542***	4,788
(ln l) <sup>2</sup>	-9,828***	0,421	-5,015***	0,367
ln l × Ost	1,156	0,304	0,798**	0,323
ln l × Abitur	-2,141	0,440	-1,502***	0,370
ln l × Deutsch	-0,665	0,494	0,490	0,521
ln l × Erwerbsbehinderung	0,305	0,300	0,303	0,384
ln l × 20 ≤ Alter < 25	-2,231***	0,452	-3,676***	0,547
ln l × 25 ≤ Alter < 30	-1,100***	0,353	-3,713***	0,423
ln l × 30 ≤ Alter < 35	-1,341***	0,351	-3,429***	0,421
ln l × 35 ≤ Alter < 40	-1,244***	0,349	-3,300***	0,435
ln l × 40 ≤ Alter < 45	-1,131***	0,348	-3,032***	0,435
ln l × 45 ≤ Alter < 50	-0,635*	0,350	-2,230***	0,428
ln l × 50 ≤ Alter < 55	-0,718*	0,390	-2,005***	0,436
ln l × 55 ≤ Alter < 60	-0,163	0,404	-1,421***	0,475
ln l × Kinder 0-6 Jahre	-3,761	5,136	3,831***	0,723
ln l × Kinder 7-16 Jahre	-1,633	5,292	1,862***	0,372
ln l × Kinder ab 17 Jahre	-2,600**	1,075	0,667**	0,285
<i>Interaktion von Einkommen und Freizeit</i>				
ln ekn × ln l	0,916**	0,441	1,630***	0,460
Loglikelihood-Wert	-1431		-2669	
Pseudo-R <sup>2</sup>	0,3438		0,1227	
Beobachtungszahl N	5955		9455	

Anmerkungen: \*  $p < 0,1$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$ .

Quelle: Eigene Berechnungen auf Basis von EITDsim.

**Tabelle 7: Parameter der Arbeitsangebotschätzung - Paare**

Variable	Koeffizient	Standardfehler
<i>Einkommen</i>		
ln ekn	-8,241**	3,214
(ln ekn) <sup>2</sup>	0,315**	0,144
ln ekn × Ost	0,801***	0,280
ln ekn × Abitur <sub>m</sub>	-0,495**	0,216
ln ekn × Abitur <sub>f</sub>	0,175	0,162
ln ekn × Deutsch <sub>m</sub>	-0,696	0,424
ln ekn × Deutsch <sub>f</sub>	0,090	0,273
<i>Freizeit</i>		
ln lm	62,585***	4,292
(ln lm) <sup>2</sup>	-8,877***	0,311
ln lf	24,741***	3,667
(ln lf) <sup>2</sup>	-3,386***	0,269
ln lm × Ost	-5,201**	2,024
ln lf × Ost	-6,934***	1,937
ln lm × Abitur <sub>m</sub>	-0,803**	0,319
ln lf × Abitur <sub>f</sub>	-0,860***	0,213
ln lm × Deutsch <sub>m</sub>	-1,164***	0,443
ln lf × Deutsch <sub>f</sub>	-1,170***	0,371
ln lm × Erwerbsbehinderung <sub>m</sub>	0,950***	0,226
ln lf × Erwerbsbehinderung <sub>f</sub>	0,574*	0,315
ln lm × 25<Alter<30	-0,723***	0,856
ln lm × 30<Alter<35	-1,647***	0,468
ln lm × 35<Alter<40	-1,921***	0,324
ln lm × 40<Alter<45	-1,811***	0,272
ln lm × 45<Alter<50	-1,757***	0,250
ln lm × 50<Alter<55	-1,477***	0,231
ln lm × 55<Alter<60	-1,184***	0,221
ln lf × 25<Alter<30	-1,708**	0,764
ln lf × 30<Alter<35	-2,907***	0,434
ln lf × 35<Alter<40	-2,597***	0,355
ln lf × 40<Alter<45	-2,653***	0,312
ln lf × 45<Alter<50	-2,368***	0,295
ln lf × 50<Alter<55	-1,865***	0,276
ln lf × 55<Alter<60	-1,300***	0,279
ln lm × Kinder 0-6 Jahre	-0,512*	0,312
ln lm × Kinder 7-16 Jahre	-0,503**	0,227
ln lm × Kinder ab 17 Jahre	-0,374**	0,179
ln lf × Kinder 0-6 Jahre	6,148***	0,497
ln lf × Kinder 7-16 Jahre	2,312***	0,243
ln lf × Kinder ab 17 Jahre	1,405***	0,176
ln lm × ln lf	0,939***	0,365
ln lm × ln lf × Ost	1,535***	0,509
<i>Interaktion von Konsum und Freizeit</i>		
ln ekn × ln lm	1,096***	0,288
ln ekn × ln lf	0,402*	0,216
Loglikelihood-Wert	-5794	
Pseudo-R <sup>2</sup>	0,2366	
Beobachtungszahl N	42045	

Anmerkungen: \*  $p < 0,1$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$ .

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Tabelle 8 zeigt, dass die in Abschnitt 4.1 genannten notwendigen theoretischen Bedingungen an die geschätzte Nutzenfunktion in den drei Schätzgruppen nicht für jede Beobachtungseinheit erfüllt werden.<sup>27</sup> Die Ableitung nach dem Einkommen ist für Paare immer positiv, für ledige Frauen und Männer ist die Bedingung in wenigen Fällen nicht erfüllt. Der Grenznutzen der Freizeit ist für ledige Männer immer positiv. Für ledige Frauen ist diese Bedingung in wenigen Fällen nicht erfüllt. Für verheiratete Männer und insbesondere Frauen wird in deutlich geringerem Umfang ein positiver Grenznutzen der Freizeit ermittelt. Darüber hinaus zeigt Tabelle 8, dass die zweite Koherenzbedingung (Konkavität) für Paarhaushalte nahezu immer erfüllt ist. Für ledige Männer und Frauen ist die zweite Ableitung nach dem Einkommen dagegen in deutlich weniger Fällen tatsächlich negativ.

**Tabelle 8:** Koherenzkriterien

	Paare	Singles	
		Männer	Frauen
$V'_{ekn}$	100,00%	92,74%	93,49%
$V'_{lf}^m$	89,06%	100,00%	-
$V'_{lf}^f$	76,76%	-	95,55%
$V''_{ekn}$	99,77%	40,62%	15,00%
$V''_{lf}^m$	100,00%	100,00%	-
$V''_{lf}^f$	100,00%	-	100,00%

Anmerkung: Angegeben ist jeweils der Anteil der Beobachtungseinheiten in Prozent, für den die Bedingung erfüllt ist.

Quelle: Eigene Berechnungen auf Basis von EITDSim.

### 4.3.2 Arbeitsangebotselastizitäten

Nachfolgend werden exemplarisch die im Modell simulierten Arbeitsangebotsreaktionen, für eine (allgemeine) Bruttolohnerhöhung um 1%, zur Schätzung von Arbeitsangebotselastizitäten verwendet.<sup>28</sup> In der Literatur finden sich diesbezüglich verschiedene Unterscheidungen, die häufigste Darstellung ist jedoch die Elastizität des Arbeitsangebots bezüglich des gesamten Arbeitsvolumens und der Partizipation.<sup>29</sup> Letztere misst den Arbeitsangebotseffekt auf extensiver Ebene, erstere den Arbeitsangebotseffekt auf intensiver Ebene. Ferner können

<sup>27</sup> Um konsistente Präferenzrelationen abbilden zu können, muss die Nutzenfunktion monoton zunehmend im Einkommen und in der Freizeit sein (1. Kriterium) und der marginale Nutzenzuwachs umso geringer ausfallen, je höher das Einkommen bzw. die Freizeit ist (2. Kriterium).

<sup>28</sup> Das zur Schätzung der Arbeitsangebotselastizitäten verwendete Verfahren basiert auf einem Ansatz von Steiner (2000).

<sup>29</sup> Vgl. Peichl et al. (2010) für eine Darstellung verschiedener Arbeitsangebotselastizitäten.

diese Elastizitäten kompensiert oder unkompensiert sein. Kompensierte Elastizitäten messen nur den Substitutionseffekt, während unkompensierte Elastizitäten den Gesamteffekt (bestehend aus Einkommens- und Substitutionseffekt) darstellen. Darüber hinaus können die Elastizitäten bedingt oder unbedingt hinsichtlich der Partizipation vor der untersuchten Reform sein.

**Tabelle 9: Arbeitsangebotselastizitäten**

	Singles	Technische Singles	Paare nur Mann <sup>a</sup>	nur Frau <sup>b</sup>
<i>Partizipation (in Prozentpunkten)</i>				
Männer	0,09	0,03	0,11	0,01
Frauen	0,11	0,09	0,03	0,08
<i>Gesamte Arbeitsstunden (in Prozent)</i>				
Männer	0,14	0,05	0,15	0,03
Frauen	0,32	0,28	0,11	0,23

Anmerkung: Unkompensierte und unbedingte Arbeitsangebotselastizitäten für eine (allgemeine) Bruttolohnerhöhung um 1%. <sup>a</sup> Bruttolohnerhöhung nur für Männer. <sup>b</sup> Bruttolohnerhöhung nur für Frauen.

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Tabelle 9 zeigt die unkompensierten und unbedingten Arbeitsangebotselastizitäten hinsichtlich der Partizipation und des (gesamten) Arbeitsvolumens für eine (allgemeine) Bruttolohnerhöhung um 1% für verschiedene Haushaltskategorien. Alle ermittelten Elastizitäten hinsichtlich des Arbeitszeitvolumens sind positiv, d.h. der positive Substitutionseffekt der Bruttolohnerhöhung auf die Arbeitszeit überkompensiert den negativen Einkommenseffekt. Ferner ist die Elastizität des Arbeitszeitvolumens immer größer als die Partizipationselastizität, da es sich um unbedingte Elastizitäten handelt. Überdies zeigt sich, dass Frauen bei einer (eigenen) Bruttolohnerhöhung wesentlich elastischer hinsichtlich ihres Arbeitsangebots reagieren als Männer. So würde z.B. eine Bruttolohnerhöhung um 1% die Partizipationsquote von Single-Frauen um 0,11 Prozentpunkte erhöhen und das gesamte Arbeitszeitvolumen würde um 0,32% steigen.

Die hier geschätzten Arbeitsangebotselastizitäten unterscheiden sich qualitativ und quantitativ nicht wesentlich von den Ergebnissen anderer Studien bzw. Modelle.<sup>30</sup>

<sup>30</sup> Vgl. Steiner (2000), Steiner & Wrohlich (2004), Haan & Steiner (2005), Jacobebbinghaus (2006), Peichl et al. (2010).

## 5 Das Datensatzfusionsmodul in EITDsim

Die Verbindung des Arbeitsangebotsmoduls mit dem Steuer-Simulationsmodul wird in EITDsim mit der Methode des *Statistical Matchings* vollzogen. Das Ziel dieser Datensatzfusion ist es, aus den zwei zugrundeliegenden Datenbeständen diejenigen Beobachtungseinheiten zu identifizieren, welche sich in bestimmten sozioökonomischen Merkmalen möglichst ähnlich sind. Wird ein sogenannter statistischer Zwilling gefunden, können die Informationen kombiniert werden. Dieser neue (artifizielle) Mikrodatensatz dient in EITDsim dann zur Evaluation von mittelfristigen Effekten (Zweit-Rundeneffekten) potentieller Politikmaßnahmen.

### 5.1 Grundlagen

Den Auftakt für die theoretische und praktische Forschung zur Verknüpfung zweier (oder mehrerer) unabhängiger Datensätze setzte Okner (1972), indem er aus dem 1966er US Tax File und dem 1967er Survey of Economic Opportunities eine neue artifizielle Mikrodatenbasis konstruierte. Seitdem hat sich ein außerordentlich breites Literaturspektrum zur Problematik der Datensatzfusion entwickelt.<sup>31</sup> Bei einem Teil der Wissenschaft stößt jedoch die Idee, zwei (oder mehrere) unabhängige Datenbestände zu verknüpfen, auf Skepsis oder Ablehnung.<sup>32</sup> Andererseits hat in den letzten Jahren die theoretische Forschung (im Zusammenspiel mit der immens gestiegenen Rechenkapazität aktueller PCs) zu den Möglichkeiten (und Grenzen) der Datensatzfusion erhebliche Fortschritte gemacht.<sup>33</sup> Darüber hinaus werden auch in der Praxis Datensatzverknüpfungen sehr häufig angewendet. Zum Beispiel entwickelte die GfK (Gesellschaft für Konsum-, Markt- und Absatzforschung) vier Analyseinstrumente auf dem Weg der Datensatzfusion. Ferner wurden integrierte Mikrodatensätze als Basis für Steuer-Simulationsmodelle in den letzten Jahren u.a. von Bork (2000), Steiner et al. (2005), Fuest et al. (2005), Wagenhals & Buck (2006), Steiner & Wakolbinger (2009) und Peichl et al. (2010) entwickelt.

Die in diesem Beitrag durchgeführte Datensatzfusion ist zuallererst von einer sogenannten exakten Datenzusammenführung abzugrenzen. Von einer exakten Zusammenführung

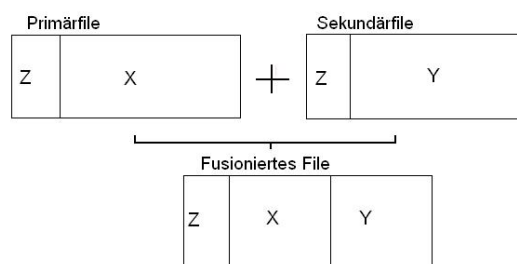
<sup>31</sup> Einen guten Überblick bieten z.B. Rodgers (1984), Rässler (2002) oder D’Orazio et al. (2006).

<sup>32</sup> Zum Beispiel bei Sims (1972a), Sims (1972b) oder Rodgers (1984). Hauptkritikpunkt ist die implizite Annahme der bedingten Unabhängigkeit zwischen den niemals gemeinsam beobachtbaren Variablen in den zu fusionierenden Datensätzen für gegebene gemeinsame Variablen.

<sup>33</sup> Vgl. Gu & Rosenbaum (1993).

spricht man, wenn Informationen von identischen Beobachtungseinheiten aus getrennten Datensätzen in einem gemeinsamen Datenbestand verschmolzen werden. Die Voraussetzung für eine exakte Zusammenführung ist grundsätzlich das Vorliegen eines eindeutigen Primärschlüssels, z.B. einer Personen-ID. Liegt kein derartiger Schlüssel vor, dann ist die Zusammenführung von Informationen in der Regel nur zwischen (möglichst) ähnlichen Merkmalsträgern möglich.<sup>34</sup> Ähnlichkeit heißt hier, dass auf Basis eines gemeinsamen Variablenvektors  $z$ , diejenigen Mikroeinheiten in beiden Datenbeständen identifiziert werden können, welche sich nur marginal unterscheiden. Das Prinzip ist in Abbildung 2 dargestellt.

**Abbildung 2:** Prinzip der Datensatzfusion



Quelle: Eigene Darstellung nach Buck (2006).

Für die Durchführung einer Datensatzfusion stehen in der Literatur eine Reihe von Fusionsalgorithmen zur Auswahl, die spezifische Vor- und Nachteile haben.<sup>35</sup> Einige Fusionsverfahren, sog. Covariate-Matching-Methods (CVM), basieren auf Distanzfunktionen zur Messung der Ähnlichkeit zwischen einer Beobachtung  $i$  aus dem Primär- und einer Beobachtung  $j$  aus dem Sekundärdatensatz. Anhand der Kovariaten  $Z$  werden dabei zunächst Abstandsmaße berechnet. Die Ermittlung des Fusionspartners erfolgt dann durch einen Vergleich der (skalaren) Maße: Mit, beispielsweise, einem Nearest-Neighbor-Verfahren können die Beobachtungseinheiten aus Primär- und Sekundärdatensatz fusioniert werden, welche einen minimalen Abstand aufweisen. Gängige Distanzfunktionen zur Bestimmung des Fusionspartners stellen z.B. die absolute Distanz (auch City-Block-Metric genannt), die euklidische Distanz, die quadratische Distanz oder die Mahalanobis-Distanz dar. Die Verwendung von Distanzfunktionen erfordert teilweise eine Normierung und eine Gewichtung der Kovariaten  $Z$ , weil es bei der Berechnung der Abstandsmaße ansonsten zu einer Übergewichtung der

<sup>34</sup> Vgl. Rosenbaum & Rubin (1983).

<sup>35</sup> Eine ausführliche Diskussion verschiedener Matching-Methoden liefert z.B. Zhao (2004).



Variablen mit dem größten Wertebereich kommen kann.<sup>36</sup> Ein weiterer Nachteil dieser Fusionsalgorithmen ist, dass bei sehr großen Datenbeständen und mit zunehmender Dimension des Kovariatenvektors die Rechenintensität stark zunimmt, weil für jede Beobachtungseinheit des Primärdatensatzes der Abstand zu allen Beobachtungen im Sekundärdatensatz berechnet werden muss. Darüber hinaus sinkt die Wahrscheinlichkeit statistische Zwillinge zu identifizieren mit zunehmender Anzahl der Kovariaten in  $Z$ .<sup>37</sup>

Eine Alternative zur Verwendung von Distanzfunktionen stellt das Propensity Score-Matching (PSM) dar.<sup>38</sup> Dieser Ansatz wurde zuerst von Rosenbaum & Rubin (1983) vorgeschlagen. Zur Bestimmung des Fusionspartners wird im PSM zunächst mittels einer Probit- oder Logitschätzung auf der Basis der Kovariaten  $Z$  ein Ähnlichkeitsindex bzw. Propensity Score ermittelt. Anhand dieses (skalaren) Maßes kann dann der Fusionspartner, z.B. mit einem Nearest-Neighbor-Verfahren, bestimmt werden. Bei großen Datensätzen weist dieser Ansatz einige Vorteile gegenüber der Verwendung von Distanzfunktionen auf. Zwar sind auch Propensity Score-Verfahren rechenintensiv, da aber die eigentliche Fusion wiederum nur auf Basis des Propensity Scores vorgenommen wird, relativiert sich bei großen Datensätzen dieser Nachteil. Ein weiterer Vorteil des PSM besteht (ferner) darin, dass mit dem Schätzverfahren eine implizite Gewichtung der Kovariaten einhergeht.

Mithin ist es aber auch möglich, die zwei oben beschriebenen Matching-Ansätze zu kombinieren. So schlagen beispielsweise Rosenbaum & Rubin (1985) ein Verfahren vor, bei dem zunächst mit einem Logit-Modell auf Basis aller (vorhandenen) unabhängigen Variablen für jede Beobachtungseinheit ein Propensity Score geschätzt wird. Anschließend wird mit einer Teilmenge der Kovariaten ein Mahalanobis-Metric-Matching mit Caliper, auf Basis des geschätzten Propensity Scores, durchgeführt. Rosenbaum & Rubin (1985), Rubin & Thomas (2000) und Baser (2006) zeigen, dass dieser Fusionsansatz den (vergleichsweise) besten Ausgleich zwischen Primär- und Sekundärdatensatz bezüglich der Kovariaten  $Z$  herstellt, d.h. die Selektionsverzerrung minimiert.

<sup>36</sup> Dies gilt jedoch nicht für die Mahalanobis-Distanz.

<sup>37</sup> Vgl. Rosenbaum & Rubin (1985), Gu & Rosenbaum (1993).

<sup>38</sup> Eine anwenderorientierte Einführung in das PSM geben Calliendo & Kopeinig (2005).

## 5.2 Umsetzung im Modell

Da ex ante nicht bestimmbar ist, welches Verfahren optimal für die Fusion der vorliegenden Datensätze wäre, werden zunächst drei verschiedene Fusionsalgorithmen durchgeführt und die Fusionsergebnisse in einer Qualitätsanalyse miteinander verglichen.<sup>39</sup> Folgende Fusionsverfahren werden nachfolgend evaluiert:

- (M1) - PSM: Nearest-Neighbor-Matching auf Propensity Score-Basis
- (M2) - CVM: Mahalanobis-Metric-Matching inkl. Propensity Score
- (M3) - PSM/CVM: Mahalanobis-Metric-Matching mit Caliper auf Propensity Score-Basis

Letztendlich wird die Zusammenführung von FAST und SOEP in EITDsim mit dem Verfahren durchgeführt, welches die (vergleichsweise) geringste Selektionsverzerrung aufweist.

Die Datenfusion wird mit der Statistiksoftware STATA und dem Modul PSMATCH2 durchgeführt.<sup>40</sup> Dafür ist gemäß Gleichung (9) zunächst in den Datensatz eine künstliche Treatment-Variable  $S$  einzuführen. Diese nimmt den Wert 1 an, wenn die Beobachtungseinheit im Primärdatensatz FAST enthalten ist (Treatment-Gruppe) und den Wert 0, falls der Merkmalsträger Teil des Sekundärdatensatzes SOEP (Control-Gruppe) ist. Anhand dieser Treatment-Variablen  $S$  kann PSMATCH2 unterscheiden, zu welchem Ursprungsdatensatz der Merkmalsträger gehört.

$$S_i = \begin{cases} 1, & \text{falls die Beobachtung Teil der FAST ist und} \\ 0, & \text{falls die Beobachtung Teil des SOEP ist.} \end{cases} \quad (9)$$

In der Literatur finden sich verschiedene Ansichten zur Frage, ob ein vergleichsweise großer Datensatz wie die FAST eher als Primär- oder Sekundärdatensatz im Matchingprozess verwendet werden sollte. So ist z.B. Rässler (2002) der Ansicht, ein großer Datensatz sollte eher als Primärdatensatz verwendet werden, damit keine Informationen außen vor bleiben.<sup>41</sup> Demgegenüber meinen D’Orazio et al. (2006), dass ein vergleichsweise großer Datensatz

<sup>39</sup> In der Literatur finden sich keine Anhaltspunkte für die Existenz eines überlegenen Fusionsalgorithmus. Vgl. S. 380 in Baser (2006).

<sup>40</sup> Für eine ausführliche Dokumentation des Moduls PSMATCH2 siehe Leuven & Sianesi (2003).

<sup>41</sup> Vgl. S. 18 in Rässler (2002).

eher als Sekundärdatensatz verwendet werden sollte, weil damit die Wahrscheinlichkeit für gute Fusionsergebnisse steigen würde.

Da die FAST sich v.a. aufgrund ihrer informationellen Tiefe hinsichtlich verschiedener Einkommensquellen, Sonderausgaben, außergewöhnliche Belastungen etc. als Datenbasis für steuerorientierte MSM empfiehlt, wird in EITDsim die FAST als Primär- und das SOEP als Sekundärdatenbasis verwendet.

### **(M1) - PSM: Nearest-Neighbor-Matching auf Propensity Score-Basis**

Auf Basis der Kovariaten  $Z$  wird für jede der drei Subfiles eine Schätzung der Propensity Scores durchgeführt. Der Propensity Score  $e(z_i)$  für Beobachtungseinheit  $i$  gibt dabei die bedingte Wahrscheinlichkeit dafür an, dass  $i$ , gegeben die Kovariate  $z_i$ , dem Primärdatensatz ( $S = 1$ ) angehört.

$$e(z_i) = P(S_i = 1 | Z_i = z_i)$$

Für die Schätzung der Propensity Scores wird ein Logit-Modell verwendet. Die geschätzten Koeffizienten und deren Signifikanzniveaus finden sich in den Tabellen 17, 15 und 16 im Appendix. Zur Bestimmung des Fusionspartners werden dann die geschätzten Propensity Scores  $\hat{e}(z)$  verglichen. Dabei wird für jede Beobachtungseinheit des Primärdatensatzes ein Partner aus dem Sekundärdatensatz gesucht, dessen Propensity Score sich nur minimal unterscheidet (Nearest-Neighbor-Verfahren).

$$d(i, j) = |\hat{e}(z_j) - \hat{e}(z_i)|$$

Die Beobachtungseinheit  $j$  des Sekundärdatensatzes, welcher eine minimale Distanz  $d(i, j)$  aufweist, wird als Fusionspartner für die Beobachtungseinheit  $i$  der des Primärdatensatzes ausgewählt und  $i$  wird aus dem Matching-Pool entfernt. Dieser Prozess wird solange wiederholt, bis für jeden Merkmalsträger  $i$  des Primärdatensatzes ein hinreichend ähnlicher Fusionspartner  $j$  aus dem Sekundärdatensatz gefunden wurde. Durch die Berücksichtigung der Common Support Bedingung wird ferner sichergestellt, dass nur jene Beobachtungseinheiten des Primärdatensatzes FAST bei der Fusion berücksichtigt werden, für die auch ein passender Partner im Sekundärdatensatz gefunden werden kann. D.h. es bleiben alle Beobachtungseinheiten des Primärdatensatzes FAST bei der Fusion außen vor, bei denen der geschätzte Propensity Score höher als das Maximum bzw. niedriger als das Minimum

der geschätzten Propensity Scores der SOEP-Beobachtungseinheiten ist.

### **(M2) - CVM: Mahalanobis-Metric-Matching inkl. Propensity Score**

Für das Mahalanobis-Metric-Matching werden zunächst alle Beobachtungseinheiten in den Subfiles in eine zufällige Ordnung gebracht und anschließend wird die Distanz von der ersten Beobachtungseinheit der Treatment-Gruppe (FAST) zu allen Beobachtungseinheiten der Control-Gruppe (SOEP) bestimmt. Die Mahalanobis-Distanz einer Beobachtungseinheit  $i$  der Treatment-Gruppe zu einer Beobachtungseinheit  $j$  der Control-Gruppe entspricht dabei

$$d(i, j) = (u - v)^T C^{-1} (u - v),$$

wobei  $u$  und  $v$  die Werte von der verwendeten Matchingvariablen enthält und  $C$  die Kovarianz-Matrix der Matchingvariablen aller Beobachtungseinheiten des Sekundärdatensatzes darstellt. Die Beobachtungseinheit  $j$  des Sekundärdatensatzes SOEP, welche eine minimale Distanz  $d(i, j)$  aufweist, wird als Fusionspartner für die Beobachtungseinheit  $i$  der des Primärdatensatzes FAST ausgewählt und  $i$  wird aus dem Matching-Pool entfernt. Dieser Prozess wird solange wiederholt, bis für jede Beobachtungseinheit  $i$  des Primärdatensatzes ein hinreichend ähnlicher Fusionspartner  $j$  aus dem Sekundärdatensatz gefunden wurde. Das Mahalanobis-Metric-Matching wird für jede der drei Subfiles auf Basis der Kovariaten Bundesland, Alter, Anzahl der Kinder und Einkommen aus nichtselbstständiger Beschäftigung sowie dem Logarithmus des zuvor geschätzten Propensity Score durchgeführt. Darüber hinaus wird zur Verbesserung der Fusionsergebnisse ein Caliper verwendet. Der Caliper definiert für jede Beobachtungseinheit einen Common Support Bereich, d.h. potentielle Fusionspartner mit einer Distanz außerhalb dieses Bereichs werden als Fusionspartner ausgeschlossen.<sup>42</sup>

### **(M3) - PSM/CVM: Mahalanobis-Metric-Matching mit Caliper auf Propensity Score-Basis**

Bei diesem Verfahren werden die beiden oben beschriebenen Algorithmen kombiniert. D.h. es wird Mahalanobis-Metric-Matching auf Basis der Kovariaten Bundesland, Alter, Anzahl der Kinder und Einkommen aus nichtselbstständiger Beschäftigung durchgeführt. Dem Ansatz von Rosenbaum & Rubin (1985) folgend, wird ferner zur Verbesserung der Fusionser-

<sup>42</sup> Vgl. S. 10 in Calliendo & Kopeinig (2005).

gebnisse ein Caliper, diesmal auf Basis des zuvor geschätzten Propensity Scores, verwendet.

### 5.3 Datenmanagement und Tests

Vor der Anwendung der Fusionsalgorithmen wird der Datensatz wieder in drei Subfiles geteilt, d.h. es wird eine Selektion entsprechend Kapitel 4.2 vorgenommen. Nach Buck (2006) ist dieses Vorgehen dann sinnvoll, wenn a priori Informationen vorliegen, dass hinreichende Unterschiede in den Datensätzen existieren, die eine Fusion dieser heterogenen Datensätze nicht ratsam erscheinen lassen.<sup>43</sup> Durch dieses Vorgehen wird somit der potentielle Fehler ausgeschlossen, dass es bei der Datensatzfusion z.B. zu einem Matching von Single-Frauen mit Single-Männern kommt.

Tabelle 10 zeigt, dass im Subfile der Single-Männer (Subsample I) vor der Datensatzfusion deutliche statistische Unterschiede zwischen Merkmalsträgern aus der FAST (Treatment-Gruppe) und dem SOEP (Control-Gruppe) vorliegen.

**Tabelle 10:** Mittelwertvergleich, Unmatched, Single-Männer (Subsample I)

Variable	Mean		%bias*	t-test	
	Treated	Control		t	p> t
Bundesland	8,441	8,341	2,5	0,95	0,344
Ost	0,315	0,285	6,6	2,57	0,010
Alter	32,396	33,884	-13,4	-5,32	0,000
Anzahl Kinder	0,214	0,015	48,3	13,88	0,000
EK aus nS	16720	27819	-67,9	-32,16	0,000
Gesamtes Einkommen	16873	28210	-68,2	-32,63	0,000

\* Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung:  $100 * (Z_i - Z_j) / [0,5 * (S_{Z_i} + S_{Z_j})]^{1/2}$ , wobei  $Z_i$  bzw.  $Z_j$  die Mittelwerte und  $S_{Z_i}$  bzw.  $S_{Z_j}$  die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

Tabelle 11 zeigt, dass auch im Subfile der Single-Frauen (Subsample II) signifikante Differenzen in den Mittelwerten der Kovariaten hervortreten. Gleiches zeigt sich im Subfile der zusammenveranlagten Paare (Subsample III). Auch in diesem Sample weisen alle Kovariate der Treatment- und Control-Gruppe in den Mittelwerten signifikante Differenzen auf (Tabelle 12).

<sup>43</sup> Vgl. Buck (2006), Kapitel 2.5.

**Tabelle 11: Mittelwertvergleich, Unmatched, Single-Frauen (Subsample II)**

Variable	Mean		%bias*	t-test	
	Treated	Control		t	p> t
Bundesland	8,231	8,116	2,9	1,07	0,285
Ost	0,297	0,255	9,4	3,49	0,000
Alter	34,063	34,826	-6,7	-2,50	0,013
Anzahl Kinder	0,342	0,267	11,6	4,22	0,000
EK aus nS	15674	21978	-50,5	-20,62	0,000
Gesamtes Einkommen	15907	22085	-49,4	-20,08	0,000

\* Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung:  $100 * (Z_i - Z_j) / [0,5 * (S_{Z_i} + S_{Z_j})]^{1/2}$ , wobei  $Z_i$  bzw.  $Z_j$  die Mittelwerte und  $S_{Z_i}$  bzw.  $S_{Z_j}$  die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

Wie oben bereits ausgeführt wurde, ist es das Ziel der Datensatzfusion (möglichst) ähnliche Merkmalsträger aus beiden Datenbeständen zu identifizieren und danach zu kombinieren. Mithin sollten nach Durchführung der Datensatzfusion die signifikanten Differenzen in den Mittelwerten der Kovariaten möglichst nicht mehr vorliegen.

**Tabelle 12: Mittelwertvergleich, Unmatched, Paare (Subsample III)**

Variable	Mean		%bias*	t-test	
	Treated	Control		t	p> t
Bundesland	8,621	7,921	17,0	9,55	0,000
Ost	0,348	0,239	24,2	13,54	0,000
Alter Mann	46,815	46,517	2,9	1,60	0,109
Alter Frau	43,85	43,75	1,0	0,56	0,572
Anzahl Kinder	1	0,8494	14,7	8,50	0,000
EK aus nS	24667	42070	-76,5	-58,75	0,000
Gesamtes Einkommen	27293	48117	-83,7	-70,06	0,000

\* Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung:  $100 * (Z_i - Z_j) / [0,5 * (S_{Z_i} + S_{Z_j})]^{1/2}$ , wobei  $Z_i$  bzw.  $Z_j$  die Mittelwerte und  $S_{Z_i}$  bzw.  $S_{Z_j}$  die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

## 5.4 Qualitätsanalyse der Matchingergebnisse

Um zu entscheiden, welches Fusionsverfahren für die Zusammenführung von FAST und SOEP letztlich am geeignetsten ist, werden, dem Ansatz von Rosenbaum & Rubin (1985) bzw. Baser (2006) folgend, nach dem Matching drei Kriterien geprüft:

- (1) *Mittelwertvergleich*: Mit zweiseitigen  $t$ -Tests wird überprüft, ob sich die Mittelwerte der einzelnen Kovariaten in Primär- und Sekundärdatensatz statistisch unterscheiden.
- (2) *Standardisierte mittlere Differenz*: Berechnung der durchschnittlichen Differenz der Ko-

varianten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung:

$$\frac{100 * (Z_i - Z_j)}{[0.5 * (S_{Z_i} + S_{Z_j})]^{1/2}} ,$$

wobei  $Z_i$  bzw.  $Z_j$  die Mittelwerte und  $S_{Z_i}$  bzw.  $S_{Z_j}$  die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.<sup>44</sup>

(3) *Prozentualer Mittelwertangleich*: Ermittlung der prozentualen Angleichung (oder Verzerrungsreduzierung) in den Mittelwerten der Kovariaten vor und nach der Fusion:

$$\frac{(Z_{Ni} - Z_{Nj}) - (Z_{Vi} - Z_{Vj})}{Z_{Vi} - Z_{Vj}} * 100 ,$$

wobei  $Z_{Vi}$  bzw.  $Z_{Vj}$  die Mittelwerte der Kovariaten in Primär- respektive Sekundärdatensatz vor der Fusion und  $Z_{Ni}$  bzw.  $Z_{Nj}$  die Mittelwerte der Kovariaten in Primär- bzw. Sekundärdatensatz nach der Fusion sind.

Diese drei Kriterien können einen Aufschluss darüber geben, welche qualitative Güte die verwendeten Fusionsalgorithmen haben. D.h. um darauf zu schließen, dass mit einem bestimmten Fusionsverfahren (möglichst) ähnliche Merkmalsträger aus beiden Datenbeständen identifiziert werden konnten, sollten nach der Datensatzfusion die Mittelwerte der Kovariaten keine signifikanten Differenzen mehr aufweisen (hohe p-Werte), die standardisierte mittlere Differenz der Kovariaten sollte gleich Null sein und der Mittelwertangleich 100% betragen.

Tabelle 13 zeigt die Ergebnisse der Qualitätsanalyse für alle drei Subsamples (Subsample I-III) und alle drei Fusionsalgorithmen (M1-M3). Der Fusionsalgorithmus M1, also das Propensity Score-Matching, ist in allen drei Subsamples die qualitativ schlechteste Matchingmethode. Nach Kriterium (1) weisen alle Kovariate von Treatment- und Control-Gruppen nach dem Matching (weiterhin) signifikante Unterschiede auf. Die Ergebnisse legen die Vermutung nahe, dass mit dem Propensity Score-Matching auch weniger exakte Matches vollzogen werden. D.h. es werden auf Basis des geschätzten Propensity Scores mithin auch Merkmalsträger aus Treatment- und Control-Gruppe als statistische Zwillinge identifiziert, die sich hinsichtlich ihrer sozioökonomischen Eigenschaften deutlich unterscheiden.

<sup>44</sup> Vgl. S. 34 in Rosenbaum & Rubin (1985).

**Tabelle 13:** Qualitätsanalyse der Fusionsergebnisse

Variable	Subsample I			Subsample II			Subsample III		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
<i>(1) p-Werte der Mittelwerttests</i>									
Bundesland	0,000	1,000	1,000	0,000	1,000	1,000	0,000	1,000	1,000
Ost	0,000	1,000	1,000	0,000	1,000	1,000	0,000	1,000	1,000
Alter <sub>m</sub>	0,000	0,772	0,558	-	-	-	0,000	0,867	0,828
Alter <sub>f</sub>	-	-	-	0,000	0,824	0,934	0,000	0,863	0,821
Anzahl Kinder	0,000	1,000	1,000	0,000	1,000	1,000	0,000	1,000	1,000
EK aus nS	0,000	0,758	0,747	0,000	0,624	0,610	0,000	0,950	0,820
Gesamtes Einkommen	0,000	0,709	0,000	0,000	0,594	0,482	0,000	0,968	0,000
<i>(2) Standardisierte mittlere Differenz in %</i>									
Bundesland	9,8	0,0	0,0	1,7	0,0	0,0	1,0	0,0	0,0
Ost	4,4	0,0	0,0	-1,4	0,0	0,0	4,7	0,0	0,0
Alter <sub>m</sub>	3,4	0,1	0,2	-	-	-	-4,8	0,2	0,2
Alter <sub>f</sub>	-	-	-	6,8	-0,1	0,0	-5,6	0,2	0,2
Anzahl Kinder	-3,2	0,0	0,0	-6,8	0,0	0,0	4,3	0,0	0,0
EK aus nS	-9,2	-0,1	-0,1	-1,8	-0,2	-0,2	-2,3	-0,1	-0,1
Gesamtes Einkommen	-8,6	-0,1	-1,7	-4,0	-0,2	-0,2	-2,1	0,0	-21,4
<i>(3) Prozentualer Mittelwertvergleich</i>									
Bundesland	-301,7	100,0	100,0	42,1	100,0	100,0	94,2	100,0	100,0
Ost	33,2	100,0	100,0	85,5	100,0	100,0	80,7	100,0	100,0
Alter <sub>m</sub>	74,6	99,4	98,8	-	-	-	-65,3	93,5	93,3
Alter <sub>f</sub>	-	-	-	-2,3	98,8	99,6	-450,6	80,6	80,1
Anzahl Kinder	93,4	100,0	100,0	41,1	100,0	100,0	71,1	100,0	100,0
EK aus nS	86,4	99,9	99,9	96,4	99,7	99,7	97,0	99,9	99,8
Gesamtes Einkommen	87,4	99,9	97,5	91,9	99,7	99,6	97,5	100,0	74,4

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Die Fusionsergebnisse der Methoden M2 und M3 unterscheiden sich qualitativ in den drei Subsamples nur wenig. Dem Fusionsalgorithmus M2 kann jedoch eine, v.a. im Hinblick auf Kriterium (1), leicht bessere Performance attestiert werden. Mithin weisen deutlich mehr Kovariate der Treatment- und Control-Gruppen nach dem Matching mit M2 insignifikante Unterschiede hinsichtlich der Mittelwerte auf als nach dem Matching mit dem Fusionsalgorithmus M3. Die standardisierten mittleren Differenzen der Kovariaten (Kriterium (2)) sind für M2 und M3 nahezu identisch, nur in Subsample III (Zusammenveranlagte Paare) fällt die standardisierte mittlere Differenz des gesamten Einkommens nach dem Matching mit M3 deutlich höher aus als mit dem Fusionsalgorithmus M2. Ähnliches gilt für die Bewertung der Fusionsergebnisse nach Kriterium (3). Sowohl M2 als auch M3 reduzieren die ex ante vorliegenden mittleren Unterschiede in den Kovariaten von Treatment- und Control-Gruppen nahezu gleich stark, wobei aber M2 den etwas stärkeren Effekt hat.

Die Ergebnisse der Qualitätsanalyse in Tabelle 13 zeigen, dass sich das Verfahren M2, ein (klassisches) Kovariaten-Matching inklusive eines geschätzten Propensity Scores, deutlich vom (populären) Propensity Score-Matching und leicht vom kombinierten Verfahren M3 abhebt. D.h. die Wahrscheinlichkeit einen statistischen Zwilling aus der FAST und dem SO-



EP zu identifizieren ist mit dem Fusionsalgorithmus M2 größer als bei Anwendung eines der anderen beiden Verfahren.

## **6 Zusammenfassung und Ausblick**

Mit diesem Beitrag wurde der Aufbau und die Funktionsweise des integrierten Mikrosimulationsmodells EITDsim dokumentiert. EITDsim stellt ein flexibel einsetzbares Instrument zur empirischen ex ante Evaluation und Analyse von (potentiellen) finanz- und sozialpolitischen Politikmaßnahmen dar. Dafür kombiniert EITDsim auf Basis des erweiterten Mikrodatensatzes FAST 2004 ein statisches Steuer-Simulationsmodell zur Schätzung kurzfristiger Erst-Rundeneffekte mit einem diskreten Arbeitsangebotsmodell zur Schätzung mittelfristiger Zweit-Rundeneffekte. Aufgrund unvollständiger Informationen in der FAST wird als Datengrundlage für das Arbeitsangebotsmodell in EITDsim das Sozioökonomische Panel (SOEP) verwendet. Auf dem Wege einer Datensatzfusion werden die Informationen aus beiden Mikrodatensätzen kombiniert. Hierfür werden drei verschiedene Fusionsalgorithmen hinsichtlich ihrer zu erwartenden Gütequalität evaluiert. Die Analyse der Fusionsergebnisse zeigt, dass ein Kovariaten-Matching inklusive eines geschätzten Propensity Scores am ehesten dazu geeignet ist, aus den Beobachtungseinheiten der FAST und des SOEP statistische Zwillinge zu identifizieren.

Eine zukünftig denkbare und auch sinnvolle Modul-Erweiterung für EITDsim würde zum einen die Modellierung einer endogenen Arbeitsnachfrage und zum anderen die Konzeption eines Wohlfahrtsmoduls zur Evaluation von Wohlfahrtswirkungen (potentieller) Politikmaßnahmen beinhalten. Darüber hinaus stellt sich grundsätzlich die Frage, ob, in Bezug auf die Bevölkerungsentwicklung und die Einkommen, eine Anpassung der Datengrundlagen bis an den aktuellen Rand möglich und zielführend wäre.

## Appendix

**Tabelle 14:** Lohnregression mit Selektionskorrektur

	Singles		Ehepaare	
	Männer	Frauen	Männer	Frauen
<b>Lohnleichung</b>				
Alter	0,797***	0,091***	0,547***	0,068**
VZ-Beschäftigung ohne Beschäftigung	-0,386***	0,160***	-0,349***	0,134***
Abitur	-1,714***	-0,526***	-1,834***	-0,716***
Fachabitur	7,903***	8,877***	12,438***	9,717***
Realschule	6,128***	6,358***	10,178***	4,857***
Hauptschule	3,366***	3,355***	4,591***	2,845***
Hauptschule	0,944	-0,010	0,778	-0,330
BL: SH	4,033**	3,744***	7,442***	2,916*
BL: HH	2,035	3,956**	10,819***	5,016**
BL: N	6,745***	3,284***	7,893***	2,890**
BL: HB	3,965	3,640*	7,477**	3,775
BL: NRW	5,796***	3,489***	9,334***	3,994***
BL: H	6,508***	4,410***	10,084***	4,916***
BL: RP, S	5,777***	3,686***	9,051***	2,774*
BL: BW	7,118***	4,494***	10,388***	4,175***
BL: BY	7,697***	3,803***	8,645***	3,260**
BL: B	4,509***	1,204	4,729**	7,255***
BL: BB	1,591	-0,664	0,933	2,106
BL: MV	2,124	0,279	-1,250	2,846
BL: SN	0,275	-0,691	-1,040	1,213
BL: ST	-0,801	-1,158	-0,604	1,361
Konstante	-14,200***	0,332	-6,079***	3,375*
<b>Partizipationsgleichung</b>				
Ost	-0,385***	-0,022	-0,162**	0,253***
Abitur	0,611***	0,397***	0,373***	0,351***
Fachabitur	0,436***	0,473***	0,538***	0,326**
Realschule	0,396***	0,344***	0,346***	0,358***
Hauptschule	0,145	0,058	0,206**	0,072
Alter	-0,005**	-0,024***	-0,024***	-0,020***
Erwerbsbehinderung	-0,415***	0,065	-0,415***	-0,209*
Kinder 0-6 Jahre		-0,927***		-1,366***
Kinder 7-16 Jahre		-0,501***		-0,529***
Kinder ab 17 Jahre		-0,135**		-0,172***
Konstante	0,620***	1,164***	1,623***	1,008***
Number of obs	2385	3082	2804	2804
Censored obs	670	1184	685	1254
Uncensored obs	1715	1899	2119	1550
Wald $\chi^2$	1149	537	1128	369

Anmerkungen: \*  $p < 0,1$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$ .

Quelle: Eigene Berechnungen auf Basis von EITDsim.

**Tabelle 15: Propensity Score Schätzung - Single-Männer**

Logistic regression	Number of obs	=	441121
	LR chi <sup>2</sup> (7)	=	1299
	Prob > chi <sup>2</sup>	=	0,0000
Log pseudolikelihood = -9690	Pseudo R <sup>2</sup>	=	0,0628

	S	Coef.	Std. Err.	z	P> z
Ost		-0,211	0,057	-3,67	0,000
Alter		0,030	0,016	1,87	0,061
Alter <sup>2</sup>		-0,000	0,000	-0,75	0,453
Anzahl Kinder		2,426	0,199	12,15	0,000
EK aus nS		-0,000	0,000	-4,34	0,000
EK aus nS/Gesamteinkommen		1,115	0,548	2,03	0,042
Gesamteinkommen × Alter		0,000	0,000	-4,23	0,000
Konstante		4,598	0,617	7,45	0,000

Quelle: Eigene Berechnungen.

**Tabelle 16: Propensity Score Schätzung - Single-Frauen**

Logistic regression	Number of obs	=	409205
	LR chi <sup>2</sup> (7)	=	401
	Prob > chi <sup>2</sup>	=	0,0000
Log pseudolikelihood = -9485	Pseudo R <sup>2</sup>	=	0,0207

	S	Coef.	Std. Err.	z	P> z
Ost		0,112	0,061	1,84	0,066
Alter		-0,016	0,017	-0,96	0,338
Alter <sup>2</sup>		0,000	0,000	0,53	0,594
Anzahl Kinder		0,156	0,048	3,24	0,001
EK aus nS		-0,000	0,000	-6,97	0,000
EK aus nS/Gesamteinkommen		-2,988	0,788	-3,79	0,000
Gesamteinkommen × Alter		0,000	0,000	2,60	0,009
Konstante		9,546	0,830	11,50	0,000

Quelle: Eigene Berechnungen.

**Tabelle 17: Propensity Score Schätzung - Paare**

Logistic regression	Number of obs	=	427152
	LR chi <sup>2</sup> (8)	=	3857
	Prob > chi <sup>2</sup>	=	0,0000
Log pseudolikelihood = -18400	Pseudo R <sup>2</sup>	=	0,0949

	S	Coef.	Std. Err.	z	P> z
Ost		0,352	0,040	8,62	0,000
Alter <sub>m</sub>		-0,024	0,021	-1,13	0,259
Alter <sub>m</sub> <sup>2</sup>		0,000	0,000	2,27	0,023
Alter <sub>f</sub>		-0,146	0,022	-6,42	0,000
Alter <sub>f</sub> <sup>2</sup>		0,002	0,000	7,89	0,000
Anzahl Kinder		0,354	0,020	16,95	0,000
EK aus nS		-0,000	0,000	-9,94	0,000
(EK aus nS) <sup>2</sup>		0,000	0,000	2,86	0,004
EK aus nS/Gesamteinkommen		0,973	0,118	8,24	0,000
Gesamteinkommen × Alter <sub>m</sub> × Alter <sub>f</sub>		-0,000	0,000	-17,77	0,000
Konstante		7,424	0,370	20,06	0,000

Quelle: Eigene Berechnungen.

## Literatur

- Bach, S., Corneo, G., & Steiner, V. (2009). *From Bottom to Top: The Entire Income Distribution in Germany, 1992-2003*. Review of Income and Wealth. Nr. 55, S. 303-330.
- Baroni, E. & Richiardi, M. (2007). *Orcutt's Vision, 50 years on*. LABORatorio R. Revelli. Working Paper, Nr. 65.
- Baser, O. (2006). *Too much ado about propensity score models? Comparing methods of propensity score matching*. Value in health the journal of the International Society for Pharmacoeconomics and Outcomes Research. Vol. 9 (6), S. 377-385.
- Blundell, R. & MaCurdy, T. (1999). Labor Supply: A Review of Alternative Approaches. In: Ashenfelter, O. & Card, D., (Hrsgb.), *Handbook of Labor Economics*. Elsevier. Vol. 3A, S. 1559-1695.
- Bonin, H. & Schneider, H. (2006). *Analytical prediction of transition probabilities in the conditional logit model*. Economics Letters. Vol. 90 (1), S. 102-107.
- Bork, C. (2000). *Steuern, Transfers und private Haushalte: Eine mikroanalytische Simulationsstudie der Aufkommens- und Verteilungswirkungen*. Peter Lang Verlag, Frankfurt/Main.
- Brenneisen, F. & Peichl, A. (2007). *Dokumentation des Wohlfahrtsmoduls von FiFoSiM*. Universität zu Köln.
- Buck, J. (2006). *Datenfusion und Steuersimulation - Theorie und Empirie im Rahmen des Mikrosimulationsmodells GMOD*. Shaker Verlag, Aachen.
- Burtless, G. & Hausman, J. (1978). *The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment*. Journal of Political Economy. Vol. 86 (6), S. 1103-1130.
- Calliendo, M. & Kopeinig, S. (2005). *Some Practical Guidance for the Implementation of Propensity Score Matching*. IZA Bonn. Discussion Paper Series, Nr. 1588.
- Creedy, J. & Duncan, A. (2002). *Behavioural Microsimulation with Labour Supply Responses*. Journal of Economic Surveys. Vol. 16 (1), S. 1-39.

- D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice (Wiley Series in Survey Methodology)*. John Wiley & Sons.
- Flood, L. & MaCurdy, T. (1992). *Work disincentive effects of taxes: An empirical analysis of Swedish men*. Carnegie-Rochester Conference Series on Public Policy. Nr. 37, S. 239-277.
- Flory, J. & Stöwhase, S. (2010). *MIKMOD-EST- A Static Microsimulation Model for the Evaluation of Personal Income Taxation in Germany*. Preliminary Version.
- Forschungsdatenzentrum (2008). *Beschreibung Lohn- und Einkommensteuerstatistik*. Forschungsdatenzentrum-Website, <http://www.forschungsdatenzentrum.de/bestand/1est/index.asp>, zuletzt besucht am 30.07.2008.
- Franz, W., Guertzgen, N., Schubert, S., & Clauss, M. (2007). *Reformen im Niedriglohnsektor: Eine integrierte CGE-Mikrosimulationsstudie der Arbeitsangebots- und Beschäftigungseffekte*. ZEW Discussion Papers. Nr. 07-085.
- Fuest, C., Peichl, A., & Schaefer, T. (2005). *Dokumentation FiFoSiM: Integriertes Steuer-Transfer-Mikrosimulations- und CGE-Modell*. Universität zu Köln. Finanzwissenschaftliche Diskussionsbeiträge Nr. 05-3.
- Gu, X. S. & Rosenbaum, P. R. (1993). *Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms*. Journal of Computational and Graphical Statistics. Vol. 2 (4), S. 405-420.
- Haan, P. & Steiner, V. (2005). *Labor Market Effects of the German Tax Reform 2000*. DIW Berlin. Discussion Papers Nr. 472.
- Haisken-DeNew, J. & Frick, J. (2011). *DTC - Desktop Companion to the German Socio-Economic Panel (SOEP)*. DIW-Website: [http://www.diw.de/documents/dokumentenarchiv/17/diw\\_01.c.38951.de/dtc.409713.pdf](http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.38951.de/dtc.409713.pdf), Zuletzt aufgerufen am 30.05.2012.
- Hall, R. E. (1973). *Wages, Incomes and Hours of Work in the US Labor Force*. In: Cain, C. & Watts, H., (Hrsgb.), *Income Maintenance and Labour Supply*. Institute for Research on Poverty Monography Series, Academic Press, London. S. 102-162.

- Heckman, J. (1976). *The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models*. Annals of Economic and Social Measurement. Nr. 5, S. 475-492.
- Heckman, J. (1979). *Sample Selection Bias as a Specification Error*. Econometrica. Nr. 47, S. 153-161.
- Jacobebbinghaus, P. (2006). *Steuer-Transfer-Mikrosimulation als Instrument zur Bestimmung des Einflusses von Steuern und Transfers auf Einkommen und Arbeitsangebot einzelner Haushalte*. Dissertation, Universität Bielefeld.
- Leuven, E. & Sianesi, B. (2003). Psmatch2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components, Boston College Department of Economics.
- McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behaviour*. Frontiers in Econometrics. Vol. 1 (2), S. 105-142.
- Okner, B. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. In: *Annals of Economic and Social Measurement*, NBER Chapters. National Bureau of Economic Research, Inc. Vol. 3 (1), S. 325-341.
- Orcutt, G. (1957). *A new type of socio-economic system*. Review of Economics and Statistics. Vol. 39 (2), S. 116-123.
- Peichl, A., Schneider, H., & Siegloch, S. (2010). *Documentation IZAΨMOD: The IZA Policy Simulation MODel*. IZA Bonn. IZA Discussion Paper Nr. 4865.
- Rodgers, W. L. (1984). *An Evaluation of Statistical Matching*. Journal of Business & Economic Statistics. Vol. 2 (1), S. 91-102.
- Rosenbaum, P. R. & Rubin, D. B. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects*. Biometrika. Vol. 70, S. 41-55.
- Rosenbaum, P. R. & Rubin, D. B. (1985). *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score*. The American Statistician. Vol. 39, S. 33-38.

- Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics. Springer.
- Rubin, D. B. & Thomas, N. (2000). *Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates*. *Journal of the American Statistical Association*. Vol. 95 (450), S. 573-585.
- Sims, C. A. (1972a). Comments. In: *Annals of Economic and Social Measurement*, NBER Chapters. National Bureau of Economic Research, Inc. Vol. 1 (3), S. 343-345.
- Sims, C. A. (1972b). Rejoinder. In: *Annals of Economic and Social Measurement*, NBER Chapters. National Bureau of Economic Research, Inc. Vol. 1 (3), S. 355-357.
- Statistisches Bundesamt (2006). *Datenreport 2006 - Zahlen und Fakten über die Bundesrepublik Deutschland*. Statistisches Bundesamt, Bonn.
- Steiner, V. (2000). *Können durch einkommensabhängige Transfers an Arbeitnehmer die Arbeitsanreize gestärkt werden? Eine ökonometrische Analyse für Deutschland*. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, IAB Nürnberg. Vol. 33 (3), S. 385-395.
- Steiner, V., Haan, P., & Wrohlich, K. (2005). *Dokumentation des Steuer-Transfer-Mikrosimulationsmodells STSM 1992-2002*. DIW Berlin. Data Documentation 9.
- Steiner, V. & Wakolbinger, F. (2009). *The Austrian Tax Transfer Modell ATTM, Version 1.1*. ATTM Research.
- Steiner, V. & Wrohlich, K. (2004). *Household Taxation, Income Splitting and Labor Supply Incentives - A Microsimulation Study for Germany*. DIW Berlin. Discussion Papers Nr. 421.
- Stern, N. (1986). On the specification of labour supply functions. In: Blundell, R. & Walker, I., (Hrsgb.), *Unemployment, search and labour supply*. Cambridge University Press, Cambridge, UK. S. 143-189.
- Struch, G. (2012). *Eine verteilungspolitische Beurteilung aktueller Reformvorschläge zur deutschen Einkommensbesteuerung*. *Journal of Economics and Statistics*. Band 232 (5), S. 567-588.

- Struch, G. & Jenderny, K. (2010). *Dokumentation des Extended Income Tax Dataset EITD*. Working Paper.
- Van Soest, A. (1995). *Structural Models of Family Labor Supply - A Discret Choice Approach*. The Journal of Human Resources. Nr. 30, S. 63-88.
- Van Soest, A. & Das, M. (2000). *Family Labor Supply and Proposed Tax Reforms in the Netherlands*. Tilburg University, Center for Economic Research. Nr. 2000-20.
- Van Soest, A. & Euwals, R. (1999). *Desired and Actual Labour Supply of Unmarried Men and Women in the Netherlands*. Labour Economics. Nr. 6, S. 95-118.
- Van Soest, A., Woittiez, I., & Kapteyn, A. (1990). *Labor Supply, Income Taxes and Hours Restrictions in The Netherlands*. The Journal of Human Resources. Nr. 25, S. 517-558.
- Wagenhals, G. & Buck, J. (2006). GMOD+: An Innovative Tax-Benefit Microsimulation Modeling Tool. In: Borutzky, W., Orsoni, A., & Zobel, R., (Hrsgb.), *Proceedings 20th European Conference on Modelling and Simulation*., ECMS. S. 354-359.
- Zhao, Z. (2004). *Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence*. The Review of Economics and Statistics. Vol. 86 (1), S. 91-107.



## **Bisher sind in der Reihe folgende FDZ-Arbeitspapiere erschienen:**

*Arbeitspapier Nr. 45:* Ein Mikrosimulationsmodell zur Berechnung der einkommensteuerlichen Bemessungsgrundlage und der Steuerzahlung auf Basis des Taxpayer-Panels, T. P., Schmidt, April 2013

*Arbeitspapier Nr. 44:* Das SAS-Makro newvar. Entwicklung und Anwendung eines Hilfsinstruments zur effizienten Erstellung neuer Variablen in der DRG-Statistik, T. Hochgürtel, T. Lösch, März 2012

*Arbeitspapier Nr. 43:* Average wage, qualification of the workforce and export performance in German enterprises: Evidence from KombiFiD data, J. Wagner, Januar 2012

*Arbeitspapier Nr. 42:* The Quality of the KombiFiD-Sample of Enterprises from Manufacturing Industries: Evidence from a Replication Study, J. Wagner, Dezember 2011

*Arbeitspapier Nr. 41:* How to define an enterprise and assign trade declarations to the right one: Exploration of German traders' micro transaction data, C. Stirböck, August 2011

*Arbeitspapier Nr. 40:* Definition von nutzerseitigen Kriterien für Datenstrukturfiles, J. Höninger/M. Rosemann/R. Voshage, Juli 2011

*Arbeitspapier Nr. 39:* Improvement of data access – The long way to remote data access in Germany, M. Brandt/M. Zwick, Juni 2011

*Arbeitspapier Nr. 38:* Decentralised Access to Confidential Microdata in Europe, M. Brandt/P. Eilsberger/M. Zwick, Juni 2011

*Arbeitspapier Nr. 37:* Masking Micro Data with Stochastic Noise, J. Höhne/J. Höninger, Mai 2011

*Arbeitspapier Nr. 36:* Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, A. Vogel, April 2011

*Arbeitspapier Nr. 35:* Temporary agency work and firm performance, S. Nielen/A. Schiersch, April 2011

*Arbeitspapier Nr. 34:* Harmonisation of statistical confidentiality in the Federal Republic of Germany, M. Brandt/A. Crößmann/C. Gürke, März 2011

*Arbeitspapier Nr. 33:* Remote Access. Eine Welt ohne Mikrodaten ??, G. Ronning/P. Bleninger/ J. Drechsler/C. Gürke, Februar 2011

*Arbeitspapier Nr. 32:* Compiling a Harmonized Database from Germany's 1978 to 2003 Sample Surveys of Income and Expenditure. T. Bönke/C. Schröder/C. Werdt, Mai 2010

*Arbeitspapier Nr. 31:* The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009

*Arbeitspapier Nr. 30:* Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten - Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/M. Zwick, Januar 2008

*Arbeitspapier Nr. 29:* Reichtum in Niedersachsen und anderen Bundesländern -Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008

*Arbeitspapier Nr. 28:* Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009

*Arbeitspapier Nr. 27:* Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008

*Arbeitspapier Nr. 26:* Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/K. Kohn/A. C. Lembcke, August 2008

*Arbeitspapier Nr. 25:* German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008

*Arbeitspapier Nr. 24:* Zum Einkommensreichtum Älterer in Deutschland – Neue Reichtumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/J. Merz, Februar 2008

*Arbeitspapier Nr. 23:* Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/D. Oberschachtsiek/R. Pohl, November 2007

*Arbeitspapier Nr. 22:* Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung - Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007

*Arbeitspapier Nr. 21:* KombiFid - Kombinierte Firmendaten für Deutschland, S. Bender/ J. Wagner/ M. Zwick, November 2007

*Arbeitspapier Nr. 20:* Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007

*Arbeitspapier Nr. 18:* Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007

*Arbeitspapier Nr. 17:* Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/ R. Lenz, Dezember 2006

*Arbeitspapier Nr. 16:* Die europäische Union - Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006

*Arbeitspapier Nr. 15:* Keeping in Touch - A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/L. Osberg, November 2006

*Arbeitspapier Nr. 14:* Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/ C. Gräß/S. Kriete-Dodds, November 2006

*Arbeitspapier Nr. 13:* Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006

*Arbeitspapier Nr. 12:* Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006

*Arbeitspapier Nr. 11:* Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

*Arbeitspapier Nr. 10:* Konstruktion und Bewertung eines ökonomischen Einkommens aus der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/F. Neher/ C. Schröder, August 2006

*Arbeitspapier Nr. 9:* Anonymising business micro data - results of a German project, R. Lenz/ M. Rosemann/D. Vorgrimler/R. Sturm, Juni 2006

*Arbeitspapier Nr. 8:* Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/H.-P. Hafner/D. Schmidt, Juni 2006

*Arbeitspapier Nr. 7:* A standard for the release of microdata, R. Lenz/ D. Vorgrimler/ M. Scheffler, Juni 2006

*Arbeitspapier Nr. 6:* Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006

*Arbeitspapier Nr. 5:* De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/M. Zwick, Oktober 2005

*Arbeitspapier Nr. 4:* Matching German turnover tax statistics, R. Lenz/D. Vorgrimler, Juni 2005

*Arbeitspapier Nr. 3:* The research data centres of the Federal Statistical Office and the statistical offices of the Länder, S. Zühlke/M. Zwick/S. Scharnhorst/T. Wende, März 2005

*Arbeitspapier Nr. 2:* Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/M. Zwick, April 2005

*Arbeitspapier Nr. 1:* Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005



