ZAF (2009) 42:121–139
DOI 10.1007/s12651-009-0013-y

RESEARCH PAPER

Verknüpfung von personenbezogenen Prozess- und Befragungsdaten – Selektivität durch fehlende Zustimmung der Befragten?

Josef Hartmann · Gerhard Krug

Angenommen: 7. Januar 2009 / Online veröffentlicht: 3. April 2009

Zusammenfassung Die Verknüpfung von Daten aus verschiedenen Quellen kann das Analysepotenzial deutlich erhöhen. Voraussetzung ist, dass im Zuge der Verknüpfung keine Fehlerquellen auftreten, die den positiven Effekt der Verknüpfung zunichte machen. Der vorliegende Beitrag untersucht vor diesem Hintergrund, inwiefern solche Fehler bei der Verknüpfung von personenbezogenen Prozessdaten der Bundesagentur für Arbeit (BA) und Befragungsdaten festzustellen sind.

Die Möglichkeit, personenbezogene Prozess- und Befragungsdaten auf individueller Ebene zu verknüpfen, ist aus datenschutzrechtlichen Gründen von der Zustimmung der betroffenen Personen abhängig. Daher ist nicht auszuschließen, dass es bei Analysen mit Daten, die nur die der Verknüpfung Zustimmenden enthalten, zu verzerrten Ergebnissen kommt. Anhand einer Erhebung, in der diese Zustimmung erfragt wurde, analysieren wir die Selektivität der Zustimmungsbereitschaft von Befragungsteilnehmern im Hinblick auf eine Reihe von Einflussfaktoren, welche sich in vielen Untersuchungen als relevant für das Befragtenverhalten gezeigt haben. Wir verwenden dabei ein Mehrebenen-Probit-Modell. Es zeigt sich, dass nur zwischen wenigen Merkmalen und der Zustimmungsbereitschaft ein Zusammenhang besteht: Nichtdeutsche, Frauen, Westdeutsche und Personen mit niedrigem Einkommen stimmen seltener der Verknüpfung zu.

Um die Auswirkung eventueller Selektivitäten der Datenverknüpfung auf die Ergebnisse konkreter Analysen zu überprüfen, schlagen wir einen einfachen Test auf Basis einer „Seemingly Unrelated Estimation“ vor und setzen diesen exemplarisch in zwei Beispielen um.

J. Hartmann (✉)
TNS Infratest Sozialforschung,
Landsberger Str. 338, 80687 München, Deutschland
E-Mail: josef.hartmann@tns-infratest.com

G. Krug
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit,
Regensburger Str. 104, 90478 Nürnberg, Deutschland
E-Mail: Gerhard.krug@iab.de

Record Linkage of Register and Survey Data – is there selection bias from requiring respondents to give their consent?

Abstract Linking data from different sources can be used to compensate for the weaknesses characteristic for the single sources and therefore significantly increase the potential for analyses. This requires that linking data implies no consequences that cancel out the advantages. Considering this background the present contribution analyses to which extent we face such consequences when linking administrative data of the “Bundesagentur für Arbeit” (BA) to survey data on the level of individuals.

Due to reasons of data protection the respondents have to be asked to allow linking their survey data to their administrative data. This necessity may result in a kind of “linkage bias”, as it may be, that their willingness to admit correlates with other characteristics relevant for the research question. This may bias results. Using data of a study, in which persons were asked, whether they allow linking the survey data to administrative data, we analyse whether there is selectivity in admission regarding a number of characteristics that are known to influence respondent behaviour. Using multilevel analysis we show that only some characteristics are correlated with allowing data linking: Women, foreign persons, persons with low income and persons interested in

protecting private information as for example on income or social benefits tend to be underrepresented in the restricted sample using the linked data set.

In order to investigate the consequences of this selection for specific research questions we propose a simple test based on a „seemingly unrelated estimation“ and present two examples.

1 Problemstellung

Angesichts sinkender Ausschöpfungsquoten und begrenzter finanzieller Mittel zur Durchführung von Erhebungen gewinnt die Verknüpfung von Daten aus verschiedenen Quellen an Attraktivität. Zudem kann die Verknüpfung Schwächen der einzelnen Datenarten ausgleichen und so die Aussagekraft und Validität von Forschungsergebnissen erhöhen.

Bei der Datenverknüpfung (record linkage) kann zwischen der Ergänzung oder Datenfusion einerseits und exakter Verknüpfung andererseits unterschieden werden. Ergänzung in der rudimentärsten Form findet statt, wenn zwei Datenquellen parallel ausgewertet werden, etwa die Verwaltungsdaten zu sozialversicherungspflichtiger Beschäftigung und das „Sozio-oekonomische Panel“ (SOEP) zur gleichen Fragestellung. Darüber hinaus können Mikrodaten aber auch auf der individuellen Ebene ergänzt werden, indem zu den vorhandenen individuenbezogenen Merkmalen die Angaben von möglichst ähnlichen Individuen hinzugespielt werden (statistisches Matching; vgl. z. B. Rässler 2002).[1] Existieren zu denselben Individuen zwei Datenquellen mit verschiedenen Merkmalen, ist darüber hinausgehend ihre exakte Verknüpfung möglich. Sofern in beiden Datenquellen geeignete Identifikatoren vorliegen, bei Personen ist z. B. an Namen, Geburtsdatum und eventuell Adresse zu denken, können die Informationen aus den beiden Datenquellen direkt verknüpft werden (Jenkins et al. 2005)[2]. Für die Arbeitsmarktforschung ist neben der Verknüpfung verschiedener Prozess- oder Registerdaten der Bundesagentur für Arbeit (BA) untereinander vor allem die Verknüpfung prozessproduzierter Daten der BA mit Befragungsdaten relevant. Das hieße beispielsweise, zu erhobenen Befragungsdaten die in der Bundesagentur für Arbeit angefallenen Prozessdaten hinzuzuspielen oder umgekehrt die Registerdaten der BA mit Informationen aus Befragungen anzureichern.

Die Wichtigkeit, die der exakten Verknüpfung von Datenquellen in der wissenschaftlichen Diskussion zugeschrieben wird, spiegelt sich auch in den Empfehlungen der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik wider: „Die Kommission empfiehlt, die Möglichkeit exakter Datenverknüpfungen – ohne ausdrückliche Zustimmung aller einzelnen Befragten – für rein statistische Zwecke gesetzlich zu regeln. Eine exakte Verknüpfung von Erhebungsdaten mit Registerdaten sollte grundsätzlich ebenso beurteilt und geregelt werden“ (Kommission 2001, S. 24).

Gegenwärtig ist die Zusammenspielung von personen- und haushaltsbezogenen Prozess- und Befragungsdaten auf individueller Ebene allerdings aus datenschutzrechtlichen Gründen von der Zustimmung der betroffenen Personen abhängig.[3] Dabei kann nicht ausgeschlossen werden, dass die Bereitschaft dazu systematisch mit untersuchungsrelevanten Eigenschaften der Personen variiert, wodurch es zu einer Art Verknüpfungs-Bias kommen kann. Verweigert eine Person ihre Zustimmung, so wirkt sich dies bei (multivariaten) Analysen, die das gesamte Spektrum der Prozess- und Befragungsdaten nutzen wollen, als Totalausfall dieser Person für die entsprechenden Analysen aus (Unit Nonresponse): Es können nur Personen einbezogen werden, die keine Datenschutzbedenken haben.

Aus der Umfrageforschung ist bekannt, dass die Teilnahmebereitschaft an Befragungen von Merkmalen abhängen kann, die für die Untersuchungsfragen relevant sind, wodurch sich verzerrte Schätzungen (Bias) im Hinblick auf die Ergebnisse ergeben können. In ähnlicher Weise ist zu erwarten, dass auch die Bereitschaft, einer Zuspielung von Prozessdaten zuzustimmen, unter den Befragten nicht zufällig verteilt ist. Daher ist zu fragen, ob eine solche Beschränkung der Analysegesamtheit nicht zu Problemen führt, welche die Vorteile der Datenverknüpfung aufheben könnten.

Vor diesem Hintergrund soll in der vorliegenden Arbeit analysiert werden, in welchem Ausmaß und mit welchen Konsequenzen es bei der exakten Verknüpfung auf Basis ausdrücklicher Zustimmung der Betroffenen zur selektiven Zusammensetzung der Stichprobe kommt. Als Datenquelle dient eine Befragung, die TNS Infratest Sozialforschung im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit zur Evaluation der Kombilohnfördermaßnahme „Mainzer Modell“ durchgeführt hat (Gewiese et al. 2004). Dabei wurden Geförderte und

[1] Die Begriffe „individuell“ oder Individuum sind hier ganz allgemein zu verstehen. Es kann sich um Personen, aber auch um aggregierte Einheiten, wie z. B. Betriebe, Gemeinden oder Arbeitsagenturbezirke handeln. Die folgende Diskussion setzt den Fokus aber auf Personen. Dies liegt darin begründet, dass die unten durchgeführten Analysen auf Personendaten zurückgreifen. Grundsätzlich gelten die angeführten Vorteile einer Verknüpfung allerdings auch für Einheiten auf Aggregatebene.

[2] Selbst beim Vorliegen eindeutiger Identifikatoren treten in der Praxis jedoch nicht selten Schwierigkeiten auf.

[3] Dies gilt nicht, wenn es sich auf individueller Ebene nicht um Personen handelt. Hier hat § 13a BStatG die Grundlage zur Verknüpfung von Befragungs- und Prozessdaten geschaffen. Die Autoren danken einem anonymen Gutachter für diesen Hinweis.

Personen einer Kontrollgruppe u. a. gefragt, ob sie der Verknüpfung der Befragungsdaten mit den Prozessdaten der BA zustimmen.

Vor den entsprechenden Analysen sind wir in Abschn. 2 auf die Vorteile eingegangen, die sich durch die Verknüpfung der Informationen aus den beiden Datenquellen ergeben. Anschließend werden theoriegeleitet die Bedingungsfaktoren der Zustimmung identifiziert und Hypothesen über die Richtung ihres Einflusses abgeleitet (Abschn. 3). Nach einer kurzen Beschreibung der Stichprobe (Abschn. 4.1) wird aufbauend auf der Auseinandersetzung mit der Theorie ein Modell der Zustimmungswahrscheinlichkeit entwickelt, das den Einfluss der theoretisch identifizierten Faktoren überprüft (Abschn. 4.2). Anschließend wird untersucht, welche Auswirkung die selektive Zustimmungsbereitschaft auf konkrete Forschungsergebnisse hat. Anhand von zwei beispielhaften Forschungsfragen wird analysiert, inwieweit sich die Ergebnisse unter Einschluss und unter Ausschluss nicht zustimmender Personen unterscheiden. Im Schlussabschnitt werden die Ergebnisse zusammengefasst und hinsichtlich ihrer Konsequenzen für die Forschungsarbeit bewertet.

2 Optimierung der Datenlage durch die Verknüpfung

Die Arbeitsmarktforschung ist in der privilegierten Situation, zur Beantwortung ihrer Fragestellungen sowohl Befragungsdaten als auch sogenannte prozessproduzierte Daten heranziehen zu können. Prozessproduzierte Daten werden für Verwaltungszwecke erhoben und stehen Wissenschaftlern unter bestimmten Bedingungen aufbereitet zur Verfügung. Beispielsweise finden Prozess- bzw. Registerdaten bei der Evaluation arbeitsmarktpolitischer Maßnahmen Verwendung (Kaltenborn et al. 2005), aber auch für andere Fragestellungen, wie z. B. zur Mobilität (Windzio 2004), zu geschlechtsspezifischer Entlohnung (Achatz et al. 2005) oder zum Niedriglohnsektor (Eichhorst et al. 2005). Beispiele für solche Datensätze sind die IAB Beschäftigtenstichprobe (IABS-R01, Hamann et al. 2004), die Stichprobe der integrierten Erwerbsbiografien (IEBS, Hummel et al. 2005) oder Daten, die von der Bundesagentur für Arbeit für die Evaluation arbeitsmarktpolitischer Maßnahmen zur Verfügung gestellt werden.

Diese Daten fallen automatisch im Rahmen der Geschäftstätigkeit an und ihre Erhebung verursacht keine eigenen Kosten.[4] Sie besitzen Vorteile gegenüber Befragungsdaten, sind allerdings auch mit Beschränkungen behaftet, die für Befragungsdaten nicht in dem Maße gelten (Hakim 1983; Schmähl u. Fachinger 1994; Wirth u. Müller 2004). Probleme, wie beispielsweise die Teilnahmeverweigerung, sozial erwünschtes Antwortverhalten, Antwortverweigerung oder falsche Angaben aufgrund eingeschränkten Erinnerungsvermögens der Befragten treten im Allgemeinen nicht auf.[5] Daher ist davon auszugehen, dass diese Daten im Hinblick auf bestimmte Sachverhalte eine höhere Validität als Befragungsdaten aufweisen. Während beispielsweise bei einem Interview die Erwerbsbiografie retrospektiv erfragt werden muss, können bei den Prozessdaten alle Informationen aus den Geschäftsdaten zur sozialversicherungspflichtigen Beschäftigung verknüpft werden. Auch Effekte sozialer Erwünschtheit spielen hier keine Rolle, z. B. bei Fragen zum aktuellen Bezug von Sozialleistungen oder zum Vorliegen von Sperrzeiten. Prozessdaten sind oft für einen langen Zeitraum verfügbar und Phänomene wie das in Panelstudien beobachtbare Ausfallgeschehen im Zeitverlauf („Panelmortalität“), welches zudem oft selektiv ist, also mit inhaltlich interessierenden Merkmalen, z. B. mit bestimmten Entwicklungen im Erwerbsverlauf, zusammenhängt, treten nicht auf.[6]

Ein weiterer Vorteil von Prozessdaten liegt darin, dass sie nicht selten Vollerhebungen darstellen, bzw., wenn nötig, in Form von Zufallsstichproben ohne Ausfälle aus einer Vollerhebung gezogen werden können. Der Stichprobenumfang kann dann entsprechend groß gewählt werden, sodass auch Detailanalysen mit kleineren Bevölkerungsgruppen durchgeführt werden können, die aufgrund ihrer geringen Inzidenz in bevölkerungsrepräsentativen Befragungen nur in geringer Fallzahl vorhanden sind.

Aber auch Befragungsdaten haben bestimmte Vorteile. Zunächst müssen sich Befragungen nicht auf bestimmte Zielgruppen, beispielsweise Personen, die mit den entsprechenden Behörden, z. B. der Arbeitsverwaltung, in Kontakt kommen, beschränken, sondern sie können prinzipiell alle Zielgruppen einschließen. So liegen bei der BA beispielsweise lückenlose Informationen über die sozialversicherungspflichtigen Beschäftigungszeiten einer Person vor, nicht aber über selbstständige Tätigkeiten, über beamtete Erwerbstätigkeiten, bis zur Einführung der allgemeinen Meldepflicht 1999 auch nicht über geringfü-

[4] Das heißt natürlich nicht, dass mit der Aufbereitung solcher prozessproduzierter Daten für die Wissenschaft nicht auch Zeit und Kosten verbunden sind.

[5] Das heißt nicht, dass die Prozessdaten fehlerfrei und vollständig sind. Hier ist daran zu denken, dass die Erfassung der Informationen nur im Wege der Erfüllung anderer Aufgaben erfolgt. Die Sachbearbeiter sind keine „Erhebungsexperten“ und haben darüber hinaus durchaus (Ermessens-) Spielräume, was sie erfassen, wie sie es erfassen und teilweise sogar, ob sie überhaupt eine Information erfassen.

[6] Aber auch die Prozessdaten enthalten Informationen zu bestimmten Phasen nicht, beispielsweise wenn es sich um andere Rechtskreise handelt. Ein Beispiel ist der Bezug der damaligen laufenden Hilfe zum Lebensunterhalt (Sozialhilfe), der nicht in den Daten der BA auftauchte. Es ist anzunehmen, dass hier ebenfalls eine selektive Nichterfassung gegeben ist: Für einen bestimmten Zeitpunkt oder Zeitraum können Personen mit bestimmten Merkmalen nicht in den Analysen berücksichtigt werden.

gige Beschäftigungsverhältnisse und natürlich auch nicht über Phasen der Nichterwerbstätigkeit, in denen man nicht arbeitslos gemeldet war.

Weiterhin können in Befragungen zielgerichtet die im Rahmen der Forschungsfrage interessierenden Inhalte erhoben werden. Da die Prozessdaten im Rahmen eines Verwaltungsprozesses anfallen, sind sie in Bezug auf die Individuen, die mit diesen Daten beschrieben werden, nichtreaktiv: Die Forschung hat keinen oder geringen Einfluss auf die Erhebung der Daten, weder was den Inhalt im Allgemeinen noch was die Messung der einzelnen Merkmale im Besonderen betrifft. Beides folgt keiner wissenschaftlichen Fragestellung bzw. Operationalisierung, sondern einer administrativen Logik.[7] Über bestimmte Zielpersonen werden genau die Informationen erfasst, welche die entsprechende Behörde zur Erfüllung ihres gesetzlichen Auftrages benötigt. Gerade inhaltlich interessante Abweichungen von der Normalität können damit also mit Prozessdaten nicht untersucht werden.[8] Und auch der Kranz der erhobenen und damit verfügbaren Informationen orientiert sich streng an den administrativen Notwendigkeiten, weshalb Wirth u. Müller (2004, S. 101) Registerdaten als zum Teil von „buchhalterische(r) Zahlenmentalität und paragrafenorientierte(r) Dürre gekennzeichnet“ beschreiben.

In Befragungen folgen die erhobenen Merkmale im Gegensatz dazu dem Interesse der Forschung. Dies hat den Vorteil, dass neben Merkmalen wie Alter, Beruf oder Bildung auch solche erfasst werden können, die administrativ nicht interessant sind, wie beispielsweise Informationen zum Haushaltskontext (z. B. Höhe des Einkommens eines eventuell vorhandenen Lebenspartners, Vermögen, Hausbesitz, soziale Einbettung etc.), zu Einstellungen (etwa zu Familie und Beruf), zur Zufriedenheit (etwa mit der Tätigkeit, dem Lebensstandard oder dem Lohn) oder zum Arbeitsumfeld (z. B. betriebliche Zusatzleistungen). Die Messung dieser Merkmale folgt zudem einer am Forschungsgegenstand orientierten Operationalisierung, während die Erfassung der Informationen in Prozessdaten meist auf gesetzlichen Definitionen beruht. Dies kann zwar für die Verwaltung selbst zweckmäßig sein, ist aber für spezielle wissenschaftliche Zwecke häufig unbrauchbar oder zumindest zu grob, um daraus analytischen Nutzen zu ziehen.[9]

Schließlich können Befragungen so organisiert werden, dass sie die von den Forschern benötigten Informationen zeitnah erfassen und für Analysen zur Verfügung stellen. Die Aufbereitung von Prozessdaten ist dagegen ein zeitaufwändiges Verfahren, welches dazu führt, dass oft Jahre zwischen der Erfassung und der Bereitstellung für wissenschaftliche Analysen liegen.

Wie aus der kurzen Gegenüberstellung zu ersehen ist, liegen die Vor- und Nachteile beider Datenarten auf unterschiedlichen Gebieten. Es bietet sich daher ihre Kombination an, um zu versuchen, die Schwächen der jeweils anderen Datenquelle zu beheben (z. B. Heckman et al. 1999, S. 130). Durch die Verknüpfung von prozessproduzierten Daten mit speziell für die Forschungsfrage erhobenen Befragungsdaten soll die Aussagekraft der Analyse gestärkt und die Validität der Ergebnisse erhöht werden. Gerade, aber natürlich nicht nur, im Bereich der maßnahmeevaluierenden Arbeitsmarktforschung ist es von hoher Bedeutung, möglichst reichhaltige Informationen zu den Untersuchungseinheiten zu besitzen.[10]

Eine solche direkte Verknüpfung bringt also mehrere Vorteile mit sich:

- Zunächst stehen für jedes Individuum mehr Informationen als bei der Verwendung einer einzelnen Datenquelle zur Verfügung.
- Daneben liegen die Vor- und Nachteile beider Datenarten auf unterschiedlichen Gebieten, sodass die Merkmale jeweils aus der Datenquelle herangezogen werden können, welche die valideren und für die jeweilige Forschungsfrage passenderen Informationen liefert.[11]
- Da die Befragung zeitnah Daten liefert, die Aufbereitung von Prozessdaten zumindest als „Scientific Use File“ dagegen teilweise Jahre in Anspruch nimmt, können durch die Verknüpfung von Befragungsdaten mit bereits vorliegenden Prozessdaten früher Ergebnisse berichtet werden.
- Schließlich müssen bestimmte Informationen im Rahmen der Befragung nicht zusätzlich erhoben werden, da sie in den Prozessdaten bereits vorliegen. Die damit verbundene kürzere Interviewdauer erhöht die Bereitschaft zu

[7] Wenngleich auch in letzter Zeit zu beobachten ist, dass der Einfluss der Wissenschaft auf die Erhebung forschungsrelevanter Prozessdaten zunimmt. Für diesen Hinweis danken wir Thomas Kruppe.

[8] Ganz zu schweigen von Phänomenen wie freiwilligem Engagement oder Tätigkeiten im Bereich der Schattenwirtschaft.

[9] Daneben liegen Befragungen in der Regel in Händen von „Befragungsexperten“, was die Optimierung der Abläufe zur Datenerhebung garantiert. Dies betrifft sowohl die Entwicklung der Erhebungsinstrumente als auch die tatsächliche Befragung durch einen Interviewer.

[10] Der Grund liegt darin, dass bei der Maßnahmeevaluation kausale Effekte (Holland 1986; Heckman et al. 1999) gemessen werden sollen. Diese sind nur identifizierbar, wenn die gesamte maßnahmeergebnisrelevante Selektivität der Maßnahmeteilnahme mit den verfügbaren Daten abgebildet werden kann. Je mehr sich die Maßnahmeteilnahme an nicht mit Prozessdaten erfassbaren Faktoren (z. B. Suchintensität, Motivation, Haushaltskontext etc.) orientiert, desto wichtiger werden zusätzliche Informationen aus Befragungen.

[11] Da beide Datenarten unterschiedlichen Fehlerquellen ausgesetzt sind, kann es allerdings im Einzelfall nicht eindeutig sein, welche Information die validere ist. So sind in Befragungsdaten retrospektiv erfragte erwerbsbiografische Informationen wie z. B. die Dauer von Arbeitslosigkeitsepisoden von Erinnerungsproblemen betroffen, während in Prozessdaten administrative Regelungen die gemessene Dauer beeinflussen können, etwa wenn nach einer Krankheit der arbeitslosen Person ein neuer Arbeitslosigkeitsspell begonnen wird.

Teilnahme an Befragungen, mindert das Risiko von Interviewabbrüchen und verringert die Befragungskosten.

Allerdings müssen die Befragten aus Gründen des Datenschutzes einer solchen Datenverknüpfung zustimmen. Dabei besteht eventuell das Problem, dass die Bereitschaft dazu systematisch mit untersuchungsrelevanten Eigenschaften der Personen variieren kann, was dann zu einer Art „Verknüpfungsbias" der Untersuchung führt. Da Analysen, die das gesamte Spektrum der Prozess- und Befragungsdaten multivariat nutzen wollen, nur Personen berücksichtigen können, die der Verknüpfung zustimmen, ist zu fragen, ob eine solche Beschränkung der Analysegesamtheit nicht zu Problemen führt, welche die Vorteile der Datenverknüpfung wieder zunichte machen. Bestehen z. B. geschlechtsspezifische Unterschiede hinsichtlich der Zustimmungsbereitschaft, wird sich dies auch in einer diesbezüglichen Selektivität der resultierenden Datenverknüpfungsteilstichprobe auswirken. Diese Fragestellung wurde bisher weder systematisch verfolgt noch gab es dazu geeignete Daten.

3 Theoretische Überlegungen zum Zustimmungsverhalten

Im Folgenden sollen zunächst theoriegeleitet Faktoren identifiziert werden, welche die Zustimmungsbereitschaft der Befragten zum Zusammenspielen von Prozess- und Befragungsdaten beeinflussen und somit zu einer selektiven Zusammensetzung der verbleibenden Stichprobe führen können. Dabei orientieren wir uns an allgemeinen Analysen zum Befragtenverhalten und übertragen diese auf die Situation der Zustimmung zur Datenverknüpfung.

In der Umfrageforschung werden zur Erklärung des Befragtenverhaltens verschiedene Mechanismen diskutiert. Zunächst können Überlegungen des Modells der rationalen Wahlhandlung auf das Handeln in der Befragungssituation übertragen werden: Ausgangspunkt dieses Modells sind entscheidungstheoretische Überlegungen, die situationale Unsicherheiten und Interpretationskonflikte über das Modell des subjektiv erwarteten Nutzens berücksichtigen. Die zentrale theoretische Annahme bezieht sich auf das Grundprinzip menschlichen Handelns und besagt, dass Menschen durch ihr Handeln ihren Nutzen maximieren: Sie wählen unter den gegebenen Restriktionen die Handlung, von der sie für sich subjektiv den höchsten Nutzen erwarten.[12] Auch in der Befragungssituation sind von einer Person Entscheidungen zu treffen. So muss sie entscheiden, ob sie überhaupt an einer Befragung teilnimmt, wie sie antwortet – beispielsweise, ob sie bei Faktenfragen die Mühe langen und intensiven Nachdenkens auf sich nimmt, ob sie grundsätzlich wahrheitsgemäß antwortet oder lieber so, wie es sozial erwünscht ist – oder eben ob sie erlaubt, dass ihre Befragungsdaten mit Prozessdaten kombiniert werden. Das allgemeine Modell ist also – wie diese kurzen exemplarischen Ausführungen zeigen – prinzipiell auf alle Entscheidungen in der Befragungssituation anwendbar.

Die Höhe des Nutzens, den eine befragte Person mit der Zustimmung verbindet, ist von ihrer Erwartung positiver Konsequenzen durch das Zusammenspielen für ihre persönliche Situation abhängig. Es ist allerdings zu vermuten, dass der Nutzen einer Zustimmung für den Befragten eher gering ausfällt. Dagegen ergeben sich durch die Verweigerung für die Befragten soziale Kosten der Rechtfertigung gegenüber dem Interviewer, die sich je nach Unmittelbarkeit des Kontaktes (face-to-face, telefonisch, postalisch) unterscheiden werden. Da aus der Zustimmung wohl kaum Nutzen, aus der Ablehnung aber Kosten entstehen, ist insgesamt von einer eher hohen Zustimmung auszugehen.

Entscheidungen über die Zustimmung basieren darüber hinaus auch auf den Einstellungen der Befragten gegenüber Befragungen im Allgemeinen oder dem Untersuchungsziel und der den Auftrag gebenden Institution im Besonderen (hier beispielsweise dem Bundesministerium für Arbeit und Sozialordnung).[13] Auch aus dieser Perspektive der relevanten Einstellungen ist zunächst auf eine hohe Zustimmungsbereitschaft zu schließen, da bei der Zielperson bereits die Entscheidung zu Interviewteilnahme positiv ausgefallen war. Dies deutet einerseits auf eine im Allgemeinen positive Einstellung zur Befragung hin, andererseits auf ein gewisses Commitment, welches die Zielperson bereits eingegangen ist. Eine Zustimmung zum Zusammenspielen von Prozess- und Befragungsdaten steht in Einklang mit diesem Commitment und gewährleistet damit eine kognitive Balance, da beide Entscheidungen positiv sind. Eine Verweigerung der Zustimmung dagegen würde ein Ungleichgewicht (kognitive Dissonanz) bedeuten

[12] Angemerkt sei, dass hier keine Annahmen über die konkreten kognitiven Prozesse und auch nicht darüber, ob diese bewusst ablaufen, getroffen werden. Damit findet keine Einschränkung auf „Zweckrationalität" im Sinne von Max Weber statt. „Rationalität" bedeutet hier nur, dass sich Menschen so verhalten, als ob sie ihre Präferenzen nach einem logischen Muster ordnen und ihr Verhalten nach diesen Präferenzen ausrichten können, dass sie also systematisch auf positive und negative Anreize reagieren können (vgl. Riker u. Ordeshook 1973; Wippler 1987).

[13] Grundsätzlich gibt es verschiedene Ansichten über das Verhältnis von Kosten-Nutzen-Erwägungen und Einstellungen. Einerseits wird argumentiert, dass Einstellungen in das entscheidungstheoretische Modell der Wahlhandlung integrierbar sind, indem sie als Mittel zur Vereinfachung der Zielstruktur aufgefasst werden und somit dann an Bedeutung gewinnen, wenn es sich um „Low-Cost"-Situationen handelt (z. B. Esser 1986, 1990). Andererseits wird davon ausgegangen, dass Personen auch unabhängig von Kosten-Nutzen-Kalkulationen einstellungsbasiert handeln (Engel et al. 2004, S. 57).
Noch eine Anmerkung zum Namen des Ministeriums: Es handelt sich hier um die Bezeichnung, die zur Zeit der Befragung gültig war.

und ist daher weniger wahrscheinlich. Allerdings erkennen die Befragten die konkreten Eigenschaften der Situation (z. B. den konkreten Inhalt der Befragung) erst im Verlauf der Interaktion mit dem Interviewer, sodass nicht automatisch mit einer Zustimmung zu rechnen ist, sondern diese erfolgt „dependig on what is made salient and how much the person negatively or positively values the attribute“ (Groves et al. 2004, S. 177).

Aus Unterschieden in den Kosten-Nutzen-Erwägungen und den allgemeinen Einstellungen der Befragten können nun Zustimmselektivitäten resultieren, welche sich schließlich in einer selektiven Zusammensetzung der zur Datenverknüpfung verbleibenden Teilstichprobe auswirken können. Dabei wird im Folgenden zwischen Einflüssen unterschieden, die unabhängig von den konkreten Inhalten der Untersuchung eine Rolle spielen und für die Bereitschaft zur Datenverknüpfung relevant werden können, und solchen, die mit den konkreten Inhalten zusammenhängen.

3.1 Allgemeine Einflüsse

Neben allgemeinen Rechtfertigungskosten, die für alle Befragten ähnlich auftreten, entstehen für einige Teilgruppen der Befragten weitere Kosten. Wenn die befragte Person etwa befürchtet, dass ihre Daten nicht vertraulich behandelt werden oder dass sie eventuell sogar Kontrollzwecken dienen, so erhöht dies die Kosten, die für sie mit einer Zustimmung zum Zusammenspielen verbunden sind. Alter, Geschlecht und Bildung können in dieser Hinsicht als Proxy-Indikatoren betrachtet werden: Bedenkt man beispielsweise, dass jüngere Menschen, Männer und Befragte mit höherer Bildung mehr Erfahrungen mit der Nutzung des Internets haben, einem Medium, in dem eine Vielzahl von Daten ausgetauscht werden, so ist zu vermuten, dass diese Gruppen eher einer Verknüpfung der Daten zustimmen. Diese Vermutung wird durch die Ergebnisse einer Untersuchung von Pascale u. Mayer (2004) gestützt, bei der sich Alter, Geschlecht und Bildung als geeignete Prädiktoren für die Zustimmung zu einer ähnlich sensiblen Frage erwiesen, nämlich zu der Frage, ob für die Zwecke von Folgebefragungen die Informationen aus früheren Befragungen an andere Haushaltsmitglieder weitergegeben werden können, falls die ursprünglich befragte Person nicht angetroffen wird. Es ist zudem zu vermuten, dass die Befürchtung, dass Daten zu Kontrollzwecken verwendet werden, aufgrund der historischen Erfahrungen in Ostdeutschland höher ist.

Zu den Kosten gehört auch, dass die Befragung Zeit kostet und insgesamt eine Unannehmlichkeit darstellt, etwa weil sie als anstrengend empfunden wurde. Als Merkmale, die in diese Richtung wirken, sind neben der Dauer der Befragung und der Arbeitszeit der erwerbstätigen Befragten – die Befragung stellt für sie eine zusätzliche Belastung dar – Indikatoren zu nennen, die auf die Schwierigkeiten verweisen, die mit der Beantwortung der Fragen verbunden sein können: Zu denken ist hier an die Bildung, das Alter und die Nationalität. Personen mit niedriger Bildung oder mit Migrationshintergrund sowie ältere Befragte sollten die Befragung als anstrengender empfinden, damit mit der Befragungssituation eher unzufrieden sein, daher eher ihre Zustimmung zur Datenverknüpfung verweigern und in der Folge in der resultierenden Teilstichprobe unterrepräsentiert sein. Als weiterer Indikator für die Kosten der Interviewteilnahme kann die Bewertung der Freizeit interpretiert werden, und zwar konkret für die damit verbundenen Opportunitätskosten: Je wichtiger die Freizeit bewertet wird, umso höher sollten diese Kosten sein.

Einstellungen, die unabhängig von den konkreten Inhalten der Untersuchung eine Rolle spielen und für die Bereitschaft zur Datenverknüpfung relevant werden können, sind die Stärke des Misstrauens bzw. der Wunsch, die eigene Privatsphäre zu schützen (vgl. Singer et al. 1993). Im Normalfall sind in Erhebungen in der Arbeitsmarktforschung keine Skalen zur Messung solcher Eigenschaften vorhanden, so auch in dieser nicht. Ein geeigneter Proxy-Indikator zur Messung des Misstrauens eines Befragten gegenüber dem Interviewer bzw. der Befragung scheint aber die Beantwortung der Frage nach dem Einkommen zu sein: So ist für Personen, die bei dieser sensitiven Frage die Angabe verweigert haben, auch eine höhere Ablehnungsquote in Bezug auf das Zusammenspielen von Befragungs- und Prozessdaten zu vermuten. Diese Vermutung wird durch die Ergebnisse von Schräpler (2004, S. 125) unterstützt: In Analysen zur Teilnahme bei Wiederholungsbefragungen im Rahmen des Sozio-oekonomischen Panels (SOEP) fand dieser einen negativen Zusammenhang zwischen der Verweigerung der Frage zum Einkommen und der Wahrscheinlichkeit, an Folgewellen teilzunehmen. Dabei ist davon auszugehen, dass der Grad, bis zu dem solches Misstrauen in Bezug auf die vertrauliche Behandlung der gegebenen Informationen im Interview selbst abgebaut wird, wiederum von Eigenschaften des Interviewers und seiner Erfahrung abhängt (Pickery et al. 2001). Zu denken ist vor allem an ein freundliches und offenes Kontaktverhalten bzw. an die Basis hierfür, und zwar an die verbale Kommunikationsfähigkeit, die durch die Bildung des Interviewers abgebildet werden kann. Auch das Geschlecht kann eine Rolle spielen: Es ist beispielsweise zu vermuten, dass insbesondere die Interaktion von junger Interviewerin und älterem männlichen Befragten die Zustimmungswahrscheinlichkeit positiv beeinflusst.

Als weiterer Einflussfaktor der Teilnahme an Befragungen – hier allerdings speziell in Betrieben – identifizieren Tomaskovic-Devey et al. (1995, S. 80ff.) neben der Fähigkeit („capacity“) und der Motivation, die Fragen zu beantworten („motive to respond“), die Autorität („authority“) der Zielperson. Sie beziehen sich dabei auf „das

Ausmaß, in dem der Antwortende die formelle oder informelle Befugnis zur Antwort hat“ (Schnabel 1997, S. 126). Da mit niedrigerer Befugnis die Kosten der Teilnahme ansteigen, weil erst die Erlaubnis eingeholt werden muss, und da bei niedriger Befugnis die Wahrscheinlichkeit negativer Sanktionen höher ist, wenn das Interview ohne Erlaubnis gegeben wird, sollte die Teilnahmebereitschaft dieser Personen bei Betriebsbefragungen niedriger sein. Diese Argumentation lässt sich auch auf die Zustimmung zum Zusammenspielen der Daten in der vorliegenden Untersuchung übertragen. Konkret ist zu vermuten, dass die Befugnis, eine eigenständige Entscheidung zu treffen, mit dem Grad der Unabhängigkeit der Zielperson und diese wiederum mit ihrer Stellung im Haushaltskontext zusammenhängt. So ist anzunehmen, dass Personen unabhängiger sind, wenn sie die Hauptverdiener im Haushalt sind. Zudem ist eine Interaktion mit dem Geschlecht der befragten Person zu vermuten: Der formulierte Zusammenhang sollte stärker für weibliche Zielpersonen als für männliche gelten.

3.2 Untersuchungsspezifische Einflüsse

Bei der Aktivierung von positiven oder negativen Einstellungen gegenüber dem Zusammenspiel werden bestimmte Attribute der Befragungssituation, wie das Thema der Befragung, die (voraussichtliche) Dauer, der Verwendungszweck der Daten, die durchführende Institution oder Merkmale der Interviewer, eine wichtige Rolle spielen. Diese Attribute offenbaren sich Befragten oft erst im Verlauf des Interviews. Ob Einstellungen überhaupt aktiviert werden, das heißt handlungsrelevant werden, hängt auch davon ab, inwiefern die durchführende Institution im Bewusstsein des Befragten subjektiv repräsentiert ist (Fetchenhauer 1999) und dieses ist wiederum von der Nähe des Befragten zur Institution abhängig (Engel et al. 2004, S. 66). Daher wäre eine erhöhte Zustimmungswahrscheinlichkeit dann zu erwarten, wenn die befragte Person sich in staatlicher Anstellung befindet oder sogar beamtet ist. Zudem ist aus dieser Perspektive von Personen, deren Beschäftigung staatlich gefördert wird oder wurde, eine stärkere Bereitschaft zur Zustimmung zu erwarten. Umgekehrt ist zu vermuten, dass Personen mit Merkmalen, die auf schlechte Arbeitsmarktchancen hindeuten, sich vom Staat und seinen Repräsentanten enttäuscht fühlen und daher weniger bereit sind, die Befragung und die damit verbundenen Ziele zu unterstützen. Zu denken ist hierbei an Langzeitarbeitslose, Geringqualifizierte, allein Erziehende oder an Befragte aus Ostdeutschland.

Schließlich lassen austauschtheoretische Überlegungen (Blau 1968; Homans 1964) hinsichtlich der Kosten der Zustimmungsverweigerung erwarten, dass Personen, die in irgendeiner Weise vom Staat im Allgemeinen oder vom Arbeitsamt bzw. der Arbeitsagentur im Besonderen Leistungen erhielten, einer Datenverknüpfung eher zustimmen als Personen, für die dies nicht zutrifft. Denn es ist zu vermuten, dass sich Leistungsempfänger den staatlichen Stellen gegenüber, in deren Auftrag die Befragung durchgeführt wird, verpflichtet sehen, die ihnen gewährte Hilfe mit einer Unterstützung der Befragung zu vergelten. Ein ähnliches implizites Austauschverhältnis könnte auch dann bestehen, wenn eine (eventuell mittlerweile) angetretene Stelle vom (damaligen) Arbeitsamt vermittelt wurde. Auch vor dem Hintergrund dieser Argumentation sollte die Tatsache, dass jemand eine staatliche Förderung erhielt, die Wahrscheinlichkeit einer Zustimmung erhöhen. Hier sind einerseits die Geförderten nach dem Mainzer Modell zu nennen, andererseits kann es aber auch unter den Befragten der Kontrollgruppe Personen mit Förderung geben. Im austauschtheoretischen Sinne relevant kann hierbei sein, inwiefern die Geförderten mit der Förderung zufrieden sind.

4 Empirische Analysen zur Zustimmungsselektivität

4.1 Verwendete Daten

Im Rahmen der Evaluation einer zunächst regional begrenzten, später bundesweit eingesetzten Kombilohnförderung (Mainzer Modell), wurden von TNS Infratest Sozialforschung im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit Geförderte und ungefördert erwerbstätige Vergleichspersonen befragt. Die Stichprobe umfasste Förderzugänge bzw. Abgänge aus Arbeitslosigkeit im Zeitraum von Januar 2001 bis März 2003.

Grundsätzlich stand eine Förderung nach dem Mainzer Modell allen Personen offen, die im Inland eine Beschäftigung aufnehmen dürfen. Es knüpfte nicht an personenbezogenen Merkmalen wie vorheriger Arbeitslosigkeit oder fehlender Qualifikation an. Als Vergleichsgruppe wurden Abgängerinnen bzw. Abgänger aus Arbeitslosigkeit ausgewählt, die keine abgeschlossene Berufsausbildung haben oder langzeitarbeitslos (mindestens seit einem Jahr ununterbrochen arbeitslos gemeldet) waren. Die Stichprobe wurde zeitlich und regional proportional zu den Zugängen in die Förderung gezogen. Insgesamt wurden 4.523 Personen befragt. Dabei handelt es sich um 3.080 geförderte und 1.443 nicht geförderte Befragte (Vergleichsgruppe). Unter den befragten Geförderten waren 705 Männer und 2.375 Frauen, in der Vergleichsgruppe waren es 695 Männer und 748 Frauen. In Ostdeutschland wurden insgesamt 1.549 Personen befragt, in Westdeutschland waren es 2.974.

Alle Befragten wurden um die Erlaubnis gebeten, ihre Befragungsdaten mit Prozessdaten verknüpfen zu dürfen. Insgesamt stimmten 74,4 % diesem Anliegen zu.

Tabelle 1 Theoretisch identifizierte Einflussfaktoren

Merkmalsträger	Merkmal	Wahrscheinlichkeit der Zustimmung
Allgemeine Einflüsse		
Befragte Person	Alter	−
Befragte Person	Bildung	+
Befragte Person	Geschlecht	Weiblich: −
Befragte Person	Nationalität	Deutsch: +
Befragte Person	Region	Ost: −
Befragte Person	Wichtigkeit von Freizeit	−
Befragte Person	Bruttoerwerbseinkommen	+ Keine Angabe: −
Befragte Person	Lebensform	Allein: +
Befragte Person	Erwerbstätigkeit Partner bei Befragung	−
Interview	Dauer (Zahl der Variablen)	−
Interviewer/in	Alter	+
Interviewer/in	Geschlecht	Weiblich: +
Interviewer/in	Bildung	+
Erwerbstätige Befragte	Arbeitszeit	−
Untersuchungsspezifische Einflüsse		
Befragte Person	Öffentlicher Dienst	+
Befragte Person	Arbeitslosigkeit vorher[a)]	−
Befragte Person	Ausbildung	Mit Ausbildung: +
Befragte Person	Erwerbsstatus vorher[a)]	Nicht erwerbstätig: +
Befragte Person	Vermittlung durch Arbeitsagentur	+
Befragte Person	Erwerbstätigkeit nachher[a)]	+
Befragte Person	Sozialhilfebezug vor und/oder nach Referenzzeitpunkt[a)]	Kein Bezug: −
Befragte Person	Zufriedenheit mit Kombilohnförderung	+

[a)] Vorher bzw. nachher beziehen sich auf den Zeitpunkt der Aufnahme der geförderten oder ungeförderten Beschäftigung.

4.2 Potenzielle Einflüsse auf das Zustimmungsverhalten

Auf der Basis der oben dargestellten theoretischen Überlegungen konnten verschiedene Merkmale identifiziert werden, von denen angenommen werden kann, dass sie die Zustimmung zur Verknüpfung von Befragungs- und Prozessdaten beeinflussen. In Tabelle 1 werden die Merkmalsträger, das Merkmal und die erwartete Wirkungsrichtung auf die Zustimmungswahrscheinlichkeit noch einmal zusammenfassend dargestellt.

4.3 Empirische Analyse

Um die empirische Relevanz der im vorangegangenen Abschnitt aus theoriegeleiteten Überlegungen zum Befragtenverhalten identifizierten Faktoren auf die Zustimmungsbereitschaft zur überprüfen, wurde ein Mehr-Ebenen-Probit-Modell mit zufälligen Effekten durchgeführt (z. B. Guo u. Zhao 2000)[14]. Damit wurde der Tatsache Rechnung getragen, dass neben Merkmalen der Befragten auch Merkmale der Interviewer mit der Zustimmungsbereitschaft in Zusammenhang stehen können. Das Modell geht von einer latenten Zustimmungsneigung y_{ij}^* beim i-ten Befragten, der vom j-ten Interviewer befragt wurde, aus:

$$y_{ij}^* = \beta_0 + x'_{ij}\beta_1 + z'_j\beta_2 + u_j + e_{ij}.$$

Die Zustimmungsneigung ist abhängig von Merkmalen der befragten Individuen x_{ij} und von Merkmalen der Interviewer z_j, wobei e_{ij} ein zufälliger Effekt auf der Befragtenebene und u_j ein zufälliger Effekt auf der Ebene der Interviewer ist. Beobachtet wird nicht die Zustimmungsneigung, sondern nur die tatsächliche Zustimmung oder Nichtzustimmung y_{ij}, die den Wert 1 bei Zustimmung (falls $y_{ij}^* > 0$) und den Wert 0 bei ausbleibender Zustimmung (falls $y_{ij}^* \leq 0$) annimmt. Unter der Annahme, dass e_{ij} standardnormalverteilt ist, ergibt sich ein Probit-Modell der Zustimmungswahrscheinlichkeit, wobei Φ für die Wahrscheinlichkeitsfunktion der Standardnormalverteilung steht

$$Pr(y_{ij} = 1|x, z, u) = \Phi(\beta_0 + x'_{ij}\beta_1 + z'_j\beta_2 + u_j).$$

Die Ergebnisse der Schätzung dieses Modells zeigt Tabelle 2, wobei zwischen den allgemeinen uns untersuchungsspezifischen Einflüssen unterschieden wird. Dargestellt sind zudem zwei Modelle: eines nur mit Haupteffekten, eines mit zusätzlichen Interaktionseffekten. Die Interaktionseffekte beziehen sich auf Geschlecht und Alter von Interviewerin und Befragtem sowie auf Geschlecht, Partnerschaft und Erwerbstätigkeit der Partnerin oder des

[14] Für die bivariaten Analysen vgl. Tabelle im Anhang.

Tabelle 2 Mehr-Ebenen-Probit-Modell mit zufälligen Effekten[15]

Interaktionseffekte	Ohne		Mit	
	B	Z	B	Z
Allgemeine Einflüsse				
Alter				
Bis unter 25 Jahre	–	–	–	–
25 bis 34 Jahre	0,064	0,85	0,051	0,67
35 bis 44 Jahre	0,061	0,83	0,046	0,61
45 bis 54 Jahre	0,015	0,18	−0,021	−0,24
55 Jahre oder mehr	0,006	0,05	−0,042	−0,36
Keine Angabe	−0,468	−0,88	−0,483	−0,90
Bildung				
Kein Abschluss	–	–	–	–
Volks-, Hauptschule	0,153	1,70	0,146	1,62
Mittl. Reife, POS 10. Klasse	0,128	1,36	0,121	1,29
Fachhochschulreife, Abitur	0,057	0,58	0,050	0,51
Geschlecht				
Männlich	–	–	–	–
Weiblich	−0,207	−3,79	–	–
Staatsangehörigkeit				
Nicht deutsch	–	–	–	–
Deutsch	0,145	1,91	0,148	1,94
Region				
West	–	–	–	–
Ost	0,158	2,96	0,164	3,04
Wichtigkeit von Freizeit				
Nicht wichtig	–	–	–	–
Wichtig	−0,024	−0,47	−0,024	−0,46
Bruttoerwerbseinkommen				
Nicht vorhanden	–	–	–	–
Bis 511 EURO	0,014	0,16	0,018	0,21
512–1023 EURO	0,038	0,54	0,039	0,57
1024 EURO oder mehr	0,185	2,33	0,177	2,22
Angabe fehlt	0,565	6,64	0,565	6,63
Lebensform				
Ohne Partner	–	–	–	–
Mit Partner	−0,017	−0,29	–	–
Erwerbstätigkeit Partner				
Nein	–	–	–	–
Ja	0,036	0,56	–	–
Geschlecht und Lebensform				
Frau mit erwerbstätigem Partner	–	–	–	–
Mann mit erwerbstätiger Partnerin	–	–	0,233	2,19
Frau mit nicht erwerbstätigem Partner	–	–	−0,036	−0,44
Mann mit nicht erwerbstätiger Partnerin	–	–	0,182	2,08
Frau ohne Partner	–	–	0,003	0,04
Mann ohne Partnerin	–	–	0,138	1,50
Dauer des Interviews (Zahl der beantworteten Fragen)	−0,001	−0,41	−0,001	−0,44

Partners, soweit vorhanden, sowie die Erwerbstätigkeit der befragten Person, die Arbeitszeit und Vermittlung der Stelle.

Im Hinblick auf die allgemeinen Einflüsse zeigt sich beim Alter und der Bildung keine signifikante Selektivität des Zustimmungsverhaltens. Beide Merkmale waren als Indikatoren für die Belastung durch das Interview eingeführt worden. Zudem war argumentiert worden, dass sie als Proxy-Indikatoren für das Vertrauen in die Gewährleistung der Anonymität der Daten interpretiert werden können. Beide Argumentationsstränge waren auch für die Staatsangehörigkeit angeführt worden. Hier bestätigen sich allerdings die Vermutungen: Deutsche Zielpersonen stimmen der Verknüpfung eher zu als nichtdeutsche, weshalb in einer verknüpften Stichprobe mit diesbezüglicher Selektivität zu rechnen ist.

Bezüglich des Geschlechts und der Wohnregion bestehen ebenfalls signifikante Selektivitäten der Zustimmung.

[15]Die Referenzkategorien sind jeweils genannt und enthalten keine Einträge

Tabelle 2 Fortsetzung

Interaktionseffekte	Ohne		Mit	
	B	Z	B	Z
Alter Interviewer				
Bis unter 25 Jahre	−0,058	−0,60	−0,056	−0,58
25 bis 34 Jahre	−0,020	−0,15	−0,015	−0,12
35 bis 44 Jahre	0,065	0,42	0,077	0,49
45 bis 54 Jahre	0,421	2,68	0,432	2,73
55 Jahre oder mehr	–	–	–	–
Geschlecht Interviewer				
Männlich	–	–	–	–
Weiblich	−0,018	−0,22	−0,027	−0,32
Interaktionseffekt: Interviewer – Befragter				
Andere Konstellationen	–	–	–	–
Junge Interviewerin – älterer Befragter	–	–	0,157	1,15
Bildung Interviewer				
Anderes	–	–	–	–
Fachhochschulreife, Abitur	0,157	1,97	0,156	1,94
Untersuchungsspezifische Einflüsse				
Öffentlicher Dienst				
Nein	–	–	–	–
Ja	−0,039	−0,41	−0,038	−0,40
Arbeitslosigkeit vor (geförderter) Arbeitsaufnahme				
Nicht arbeitlos	–	–	–	–
Bis unter 12 Monate	−0,150	−1,79	−0,146	−1,75
12 Monate oder mehr	−0,222	−2,38	−0,221	−2,36
Angaben fehlen	0,067	0,45	0,062	0,42
Ausbildung				
Keine	–	–	–	–
Ausbildung/Studium	−0,010	−0,19	−0,015	−0,28
Erwerbsstatus vor (geförderter) Arbeitsaufnahme				
Erwerbstätig	–	–	–	–
Anderes	0,044	0,55	0,044	0,55
Erwerbstätigkeit befragte Person				
Nein	–	–	–	–
Teilzeit, Vermittlung durch Arbeitsagentur	0,132	1,16	0,129	1,13
Teilzeit, Vermittlung nicht durch Arbeitsagentur	0,065	0,72	0,064	0,70
Vollzeit, Vermittlung durch Arbeitsagentur	0,016	0,14	0,012	0,11
Vollzeit, Vermittlung nicht durch Arbeitsagentur	0,014	0,15	0,012	0,14
Sozialhilfe (HLU) vor und/oder nach Arbeitsaufnahme	–	–	–	–
Weder vor noch nach	–	–	–	–
Durchgängig HLU	0,294	3,27	0,285	3,17
Einstieg: HLU nach Arbeitssaufnahme	0,156	0,81	0,161	0,83
Ende der HLU durch Arbeitssaufnahme	0,127	1,68	0,123	1,61
Mindestens einmal keine Angabe	−0,477	−2,93	−0,474	−2,90
Zufriedenheit mit Höhe der CAST-Förderung				
Nicht CAST-gefördert	–	–	–	–
Weniger zufrieden, keine Angabe	−0,082	−1,10	−0,081	−1,08
Sehr zufrieden, zufrieden	0,191	2,97	0,187	2,90
Konstante	0,416	1,81	0,251	1,02
Var at level 1	0,173	–	0,173	–
Level 2	0,092	–	0,094	–
Fallzahl	4.523	–	4.523	–
Gruppen	142	–	142	–
−2 Log Likelihood Nullmodell	5.248,0	–	5.248,0	–
−2 Log Likelihood volles Modell	5.010,5	–	5.008,1	–
Differenz −2LL Nullmodell und volles Modell	237,5	–	239,9	–
Freiheitsgrade	41		44	
P-Wert der Differenz	p <	0,001	p <	0,001
Pseudo R^2	0,05	–	0,05	–

Weibliche Zielpersonen stimmen seltener als Männer und Befragte aus Ostdeutschland eher als Westdeutsche dem Zusammenspielen zu. Letzteres widerspricht den oben formulierten Erwartungen, dass die historischen Erfahrungen in Ostdeutschland eher zu geringerer Zustimmungsbereitschaft führen würden. Damit sind sowohl Frauen als auch westdeutsche Befragungsteilnehmer im verknüpften Datensatz unterrepräsentiert. Die Einstellung der Zielpersonen zur Freizeit hat dagegen keine Konsequenz für das Zustimmungsverhalten zur Datenverknüpfung.

Personen, die ein hohes Erwerbseinkommen erzielen, stimmen der Datenverknüpfung eher zu. Die Resultate stehen in Einklang mit der Vermutung, dass der Wunsch nach Schutz der Privatsphäre sich darin äußert, dass bestimmte Fragen nicht beantwortet werden und aus dem gleichen Grund auch eher die Erlaubnis zur Datenverknüpfung nicht gegeben wird: Personen, die die Angabe zum Einkommen im Monat nach dem Referenzzeitpunkt verweigern, lassen mit geringerer Wahrscheinlichkeit das Zusammenspielen der Daten zu. Weder ob eine Zielperson einen Lebenspartner hat noch ob dieser erwerbstätig ist, wirkt sich auf das Zustimmungsverhalten aus. Auch in der Interaktion der Merkmale Geschlecht, Lebensform und Erwerbstätigkeit des Partners (Modell mit Interaktionseffekten) zeigt sich lediglich die bereits oben angesprochene Selektivität in Bezug auf das Geschlecht.

Entgegen der Erwartungen zeigt sich kein Einfluss der Dauer des gerade gegebenen Interviews. Bezüglich Alter und Geschlecht der Interviewer war argumentiert worden, dass die Zielperson ältere und weibliche Interviewerinnen eher als vertrauenswürdig einschätzt und daher eher die Zustimmung zum Zusammenspielen von Befragungs- und Prozessdaten gibt. Für das Alter sprechen die Ergebnisse für die Hypothese, für das Geschlecht ist kein Effekt feststellbar. Ebenfalls nicht bestätigt wird die Vermutung, dass ältere Befragte bei einer jungen Interviewerin eher die Datenverknüpfung erlauben. Der Interaktionseffekt zwischen Alter und Geschlecht von Interviewerin und Befragtem ist nicht signifikant. Mit höherer Bildung der Interviewer geht dagegen eine erhöhte Zustimmung der Befragten einher. Dies war mit dem Argument, dass mit der Bildung die Kommunikationsstärke steigt, erwartet worden.

Bei den untersuchungsspezifischen Einflüssen kann keine Selektivität der Zustimmung in Bezug auf die Zugehörigkeit zum öffentlichen Dienst festgestellt werden. Für Zielpersonen ohne Berufsausbildung und solche, die vor dem Referenzzeitpunkt arbeitslos waren, war erwartet worden, dass sie die Datenverknüpfung eher nicht erlauben, weil sie vom Arbeitsmarkt und den arbeitsmarktpolitischen Institutionen enttäuscht sind. Die Hypothese wird für Arbeitslose bestätigt, für Personen ohne Berufsausbildung allerdings nicht. Ebenso wenig hängt die Tatsache, dass die Zielpersonen vor dem Referenzzeitpunkt einer Erwerbstätigkeit nachgingen, mit der Zustimmung zusammen.[15]

Auch die Tatsache, dass die befragte Person zum Befragungszeitpunkt erwerbstätig ist, führt nicht zu selektivem Zustimmungsverhalten.[16] Interaktionseffekte der Erwerbstätigkeit mit der Arbeitszeit und der Vermittlung der Stelle durch das Arbeitsamt sind ebenfalls nicht feststellbar. Ob also eine angetretene Stelle vom Arbeitsamt vermittelt wurde, beeinflusst nicht die Bereitschaft zur Datenverknüpfung. Betrachtet man den Bezug von Sozialhilfeleistungen (Hilfe zum Lebensunterhalt, HLU), so sprechen die Ergebnisse für die Vermutung, dass der Bezug von staatlichen Unterstützungsleistungen die Bereitschaft, die Datenverknüpfung zu erlauben, in der vorliegenden Untersuchung erhöht. Im Übrigen zeigt sich auch hier, dass Personen, die bei eher sensiblen Fragen die Auskunft verweigern, in verknüpften Stichproben erwartungsgemäß unterrepräsentiert sind. Schließlich zeigt sich, dass genau diejenigen geförderten Personen, welche aus der Förderung einen Vorteil ziehen, der Datenverknüpfung eher zustimmen. Damit werden die austauschtheoretischen Überlegungen gestützt.

Zusammenfassend kann festgehalten werden: Die Resultate der multivariaten Analysen zeigen, dass zumindest teilweise mit einer Selektivität der Zustimmung zur Datenverknüpfung zu rechnen ist. So werden Frauen und nichtdeutsche Staatsbürger wohl tendenziell in verknüpften Datensätzen unterrepräsentiert sein. Ebenso unterrepräsentiert werden Personen in unteren Einkommensgruppen sein und solche, die generell eher am Schutz privater Informationen interessiert sind, wie z. B. zum Einkommen oder zum Bezug von Sozialleistungen. Die Ergebnisse zum Einfluss der Interviewereigenschaften verweisen darauf, dass der gezielte Einsatz von älteren Interviewern und solchen mit hoher Bildung die allgemeine Zustimmungsbereitschaft noch erhöhen könnte. Die Dauer des Interviews scheint hingegen keinerlei Einfluss auf die Bereitschaft zur Datenverknüpfung zu haben.

Bei den untersuchungsspezifischen Einflüssen zeigte sich, dass Personen, die mit der Förderung zufrieden sind, in der verknüpften Stichprobe überrepräsentiert sind. Ebenso stimmen Sozialleistungsbezieher häufiger der Verknüpfung zu. Dies begründet zumindest die Vermutung, dass Personen, die eine positive Einstellung zum Untersuchungsgegenstand haben, etwa weil sie positive Erfahrungen mit der auftraggebenden Institution haben, in verknüpften Stichproben überrepräsentiert sind. Je nach

[15] Es sei darauf hingewiesen, dass es beim Arbeitsamt bzw. mittlerweile der Arbeitsagentur arbeitslos Gemeldeten innerhalb gewisser Grenzen erlaubt ist, eine Erwerbstätigkeit auszuüben.

[16] Das Modell mit dem entsprechenden Haupteffekt wurde aus Platzgründen hier nicht dargestellt.

Tabelle 3 Lineare Regression der Arbeitszeit

Analyse 1: Arbeitszeit	Modell 1: vollständige Stichprobe		Modell 2: eingeschränkte Stichprobe	
	Coef.	*t-Wert*	Coef.	*t-Wert*
Alter	0,053	0,33	−0,007	−0,04
Alter (quadr.)	−0,001	−0,44	0,000	0,00
Stellung im Beruf	–	–	–	–
Arbeiter	–	–	–	–
Angestellter	−1,809	−4,48	−2,039	−4,55
sonst.	−1,204	−0,90	−1,918	−1,28
Branche				
Leih-/Zeitarbeitsfirma	–	–	–	–
Reinigungsgewerbe	−6,597	−9,24	−5,900	−7,19
Hotel- und Gaststättengewerbe	−0,687	−0,82	−0,749	−0,79
Call-Center	−2,315	−2,51	−1,570	−1,53
Energie- und Wasser; Bergbau, Verarb. Gewerbe	0,335	0,56	0,184	0,28
Handel	−2,558	−4,18	−2,766	−4,05
Verkehr/Nachrichtenüberm., Banken, Versicherungen	0,344	0,26	0,705	0,48
Öffentlicher Dienst/Sozialversicherung	−4,395	−5,84	−4,554	−5,51
Bereich andere Dienstleistungen, k. A.	−1,792	−3,18	−1,681	−2,68
Geschlecht, Lebensform und Partnererwerbstätigkeit				
Frau, allein lebend	–	–	–	–
Mann, allein lebend	4,383	4,84	4,139	3,91
Frau, allein erziehend	−1,325	−1,93	−1,738	−2,16
Mann, allein erziehend	7,221	3,13	7,284	3,49
Frau, Partner nicht erwerbstätig	−0,516	−0,66	−0,604	−0,65
Frau, Partner erwerbstätig	−0,350	−0,45	−0,933	−1,04
Mann, Partner nicht erwerbstätig	10,143	11,58	10,045	10,05
Mann, Partner erwerbstätig	10,329	10,70	10,201	9,15
Zahl der Kinder (unter 18 Jahre im Haushalt)	−0,469	−2,18	−0,357	−1,51
Nebeneinkommen vorhanden	−3,212	−3,70	−3,373	−3,51
Erfolg im Beruf sehr wichtig	−0,356	−0,90	−0,357	−0,80
Einkommen sehr wichtig	0,460	1,27	0,309	0,76
Arbeit sehr wichtig	1,012	2,81	1,000	2,46
Freizeit sehr wichtig	0,013	0,03	0,617	1,38
Familie sehr wichtig	−0,742	−1,65	−1,090	−2,20
Region: Ost	5,343	14,41	5,241	12,41
N	3.408		2.638	
R^2	0,3056		0,3211	
Wald-Test auf Gleichheit aller Koeffizienten in Modell 1 und 2	$\text{chi}^2(27) = 30{,}58$ $\text{Prob} > \text{chi}^2 = 0{,}2886$			

Legende: $*p < 0{,}1$; $**p < 0{,}05$; $***p < 0{,}01$

Fokus der Befragung bzw. den mit dieser durchgeführten Analysen kann dies problematisch sein oder auch nicht.

Trotz allem verweisen die Ergebnisse der Regressionsanalyse jedoch auf eine eher geringe Selektivität des Zustimmungsverhaltens. Die in den Modellen enthaltenen Merkmale erklären nur wenig von der Zustimmungsbereitschaft, wobei anzumerken ist, dass meist nur Proxy-Variablen für die theoretisch interessierenden Konzepte verwendet werden konnten. Insofern können erst weitere Analysen, die auf bessere Indikatoren zugreifen können, die Frage des Ausmaßes der Selektivität genauer klären. Bis zum Vorliegen solcher Analysen sind die folgenden Vergleiche zwischen Modellen mit und ohne Zustimmer als eine weitere Annäherung zu sehen.

4.4 Auswirkung der Selektivität auf Forschungsergebnisse – ein einfacher Test

Um näher zu untersuchen, ob die Tatsache, dass nur Beobachtungen von Personen ohne Datenschutzbedenken für die Analysen zur Verfügung stehen, das Ergebnis konkreter Forschungsarbeiten beeinflusst, wird im Folgenden zunächst eine Regression mit der Arbeitszeit als zu erklärendes Merkmal durchgeführt. Zeigen sich in den Modellen unterschiedliche Regressionskoeffizienten, kann davon ausgegangen werden, dass die Selektivität hinreichend groß ist, um solche Unterschiede zu erzeugen. Ergeben sich keine Unterschiede, ist die Selektivität der Zustimmungsbereitschaft in diesem konkreten Fall ohne Einfluss. Allerdings

Tabelle 4 Logistische Regression der Zufriedenheit mit der Tätigkeit

Analyse 2: Zufriedenheit mit der Tätigkeit	Modell 1: vollständige Stichprobe		Modell 2: eingeschränkte Stichprobe	
	Odds Ratio	*z-Wert*	Odds Ratio	*z-Wert*
Stundenlohn	1,039	1,73	1,040	1,55
Stundenlohn (quadr.)	1,000	−1,53	1,000	−1,42
Förderung Mainzer Modell	–	–	–	–
Geförderte (qualifiziert und nicht langzeitarbeitslos)	–	–	–	–
Geförderte (gering qualifiziert u./o. langzeitarbeitslos)	0,759	−2,04	0,724	−2,08
Vergleichsgruppe (gering qualifizierte u./o. langzeitarbeitslos)	0,892	−0,61	0,854	−0,73
Zufriedenheit mit Verdienst	3,935	10,78	4,513	10,14
Alter				
Bis unter 25 Jahre	–	–	–	–
25 bis 34 Jahre	0,854	−0,72	0,997	−0,01
35 bis 44 Jahre	0,989	−0,04	1,018	0,06
45 bis 54 Jahre	0,826	−0,61	0,763	−0,74
55 Jahre oder mehr	3,080	1,94	3,686	1,88
Bildung				
Kein Abschluss	–	–	–	–
Volks-, Hauptschule	0,706	−1,37	0,545	−1,96
Mittl. Reife, POS 10. Klasse	0,538	−2,34	0,423	−2,69
Fachhochschulreife, Abitur	0,408	−3,21	0,370	−2,96
Staatsangehörigkeit deutsch	1,651	2,56	2,013	3,06
Erwerbsstatus vor Referenzzeitpunkt				
Erwerbstätig	–	–	–	–
Anderes	0,956	−0,24	0,956	−0,21
Stellung im Beruf				
Arbeiter	–	–	–	–
Angestellter	1,018	0,14	1,040	0,26
Sonst.	0,657	−1,28	0,563	−1,54
Befristung				
Nicht befristet/trifft nicht zu	–	–	–	–
Befristet	1,188	1,51	1,260	1,75
Ausbildung				
Keine	–	–	–	–
Ausbildung/Studium	1,051	0,34	1,000	0,00
Berufserfahrung	0,987	−1,24	0,995	−0,40
Branche				
Leih-/Zeitarbeitsfirma	–	–	–	–
Reinigungsgewerbe	1,032	0,13	0,903	−0,37
Hotel- und Gaststättengewerbe	1,253	0,97	1,123	0,44
Call-Center	1,253	0,70	1,309	0,70
Energie- und Wasser; Bergbau, Verarb. Gewerbe	1,340	1,42	1,183	0,70
Handel	1,415	1,77	1,279	1,08
Verkehr/Nachrichtenüberm,, Banken, Versicherungen	2,249	2,10	2,403	1,84
Öffentlicher Dienst/Sozialversicherung	1,873	2,11	1,172	0,49
Andere Dienstleistungen, k. A.	1,685	2,92	1,382	1,58
Geschlecht	1,405	2,23	1,678	2,95
Lebensform und Partnererwerbstätigkeit				
Allein lebend	–	–	–	–
Mann, allein erziehend	1,170	0,75	1,098	0,38
Frau, allein erziehend	1,085	0,42	1,018	0,08
Partner, nicht erwerbstätig	1,299	1,28	1,289	1,06
Partner, erwerbstätig	0,985	−0,26	0,997	−0,04
Lohnersatzleistungen				
Arbeitslosenhilfe	–	–	–	–
Arbeitslosengeld	1,380	1,97	1,342	1,55
Keine Leistungen	0,920	−0,51	0,969	−0,17
Sozialhilfe (auch ergänzend)	0,967	−0,20	0,929	−0,39
Nebeneinkommen vorhanden	1,275	0,71	0,974	−0,07
Arbeit sehr wichtig	1,062	0,52	1,003	0,03
Einkommen sehr wichtig	1,044	0,36	1,127	0,89
Beschäftigung ist Übergangslösung	0,293	−10,17	0,269	−9,38
Region Ost	1,325	1,93	1,355	1,83
Konstante	2,681	2,20	2,519	1,75

lassen sich die so gewonnenen Ergebnisse nicht auf andere Analysen übertragen, da sich bei dem Vorgehen stets nur die Selektivität auswirkt, die auch mit der Arbeitszeit bzw. den Regressoren korreliert ist. Um zumindest ein breiteres Spektrum an inhaltlichen Fragestellungen abzudecken, wurde neben der linearen Regression der Arbeitszeit auch eine logistische Regression zur Zufriedenheit mit der Tätigkeit berechnet, eine Analyse also, in der es weniger um „harte" Informationen als vielmehr um Einstellungsfragen geht.

Im Folgenden werden die beiden Fragestellungen jeweils einmal mit allen verfügbaren Fällen bearbeitet und einmal so, als hätte die Datenverknüpfung stattgefunden, also ohne die Nichtzustimmer. Für die Analysen werden allerdings nur Befragungsdaten verwendet, da bei den Personen, die die Erlaubnis zum Zusammenspielen nicht gegeben haben, keine Prozessdaten zugespielt werden dürfen. Modell 1 wird mit allen verfügbaren Fällen in der Stichprobe durchgeführt, Modell 2 dagegen nur mit solchen Fällen, die der Datenzusammenspielung zugestimmt haben. Es werden die resultierenden Modelle dahingehend betrachtet, ob sie zu unterschiedlichen Schlussfolgerungen bei der Identifikation statistisch signifikanter Einflussfaktoren auf die Arbeitszeit bzw. die Zufriedenheit mit der Tätigkeit führen. Da die Modelle 1 und 2 auf sich überschneidenden Stichproben beruhen, werden sie als „Seemingly Unrelated Regression" (Greene 2003 S. 340ff.) durchgeführt und anschließend werden mittels eines Wald-Tests jeweils für beide Fragestellungen die Koeffizienten des Modells 1 mit denen des Modells 2 verglichen.

Tabelle 3 und Tabelle 4 weisen die Ergebnisse der beiden Analysen aus, jeweils differenziert nach Modell 1 und Modell 2. Betrachtet man die Regressoren im Einzelnen, dann zeigt sich in der Analyse zur Arbeitszeit bei drei Regressoren eine unterschiedliche Schlussfolgerung zur Signifikanz (bei einem Signifikanzniveau 0,05 bzw. t-Wert von 1,96). Beim Branchendummy „Callcenter" wird der kritische Wert von 1,96 im zweiten Modell unterschritten und ebenso wird der Koeffizienten der „Kinderzahl" nun im Gegensatz zum Referenzmodell 1 nicht mehr als signifikant ausgewiesen.

Auch in der zweiten Analyse zur Zufriedenheit (Tabelle 4) treten einige Unterschiede zwischen Modell 1 und 2 auf. So ist der Einfluss des Stundenlohns auf die Zufriedenheit mit der Tätigkeit ebenso wie der Einfluss einiger Branchen und der Einfluss des vorherigen Bezuges von Arbeitslosengeld im Vergleich zu Arbeitslosenhilfe nur in Modell 1 signifikant. Umgekehrt werden nun in Modell 2 vorher insignifikante Effekte als signifikant ausgewiesen. Dies sind der Effekt eines Volks- oder Hauptschulabschlusses im Vergleich zu keinem Schulabschluss und der Einfluss der Befristung des Arbeitsvertrages auf die Zufriedenheit.

Wie der Vergleich der Modelle 1 und 2 zeigt, gäbe es bei der Betrachtung der Modelle ohne Zustimmer zum Teil abweichende Interpretationen der Einflussfaktoren auf die Arbeitszeit und die Zufriedenheit mit der Tätigkeit. Ob sich die Modelle 1 und 2 tatsächlich, das heißt statistisch signifikant unterscheiden, oder ob Unterschiede in Größenordnung und Signifikanz der Koeffizienten eher auf den Fallzahlverlust bzw. zufällige Abweichungen zurückgehen, zeigt ein Test auf Gleichheit der Koeffizienten. In beiden Analysen führt der Wald-Test der Hypothese der Gleichheit aller Koeffizienten des Modells 2 mit denen in Modell 1 nicht zum Verwerfen dieser Hypothese. Daher gibt es keinen Grund anzunehmen, dass sich in den beiden beispielhaften Analysen durch die Beschränkung auf die Befragten ohne Datenschutzbedenken systematische Veränderungen der Schlussfolgerungen ergeben würden. Wie zu sehen war, bedeutet dies nicht, dass im Einzelfall ein Forscher nicht zu unterschiedlichen Schlussfolgerungen kommen würde. Die Unterschiede entstehen aber – zumindest in den hier betrachteten Anwendungsbeispielen – durch die Orientierung an einem relativ willkürlichen Schwellenwert des Prüfmaßes. So weichen denn auch die t-Werte zwischen den in Modell 1 signifikanten und in Modell 2 insignifikanten Merkmalen (bzw. umgekehrt) nur wenig voneinander ab.

5 Zusammenfassung und Diskussion

Das Zusammenspielen von Prozess- und Befragungsdaten kann eine Reihe von Vorteilen für die Forschung haben. Diese Vorteile können allerdings nicht zum Tragen kommen, wenn es aufgrund selektiver Zustimmung zum Zusammenspielen zu verzerrten Stichproben kommt. Eine Beschränkung der Analysegesamtheit auf die Fälle, die einer Verknüpfung zustimmen, ließe dann andere Ergebnisse erwarten als eine Einbeziehung aller Fälle in die Analysen. Der vorliegende Beitrag untersucht dieses Problem unter Verwendung der Daten einer Befragung zur Wirkung eines finanziellen Zuschusses zum Lohn.

Es wurde zunächst analysiert, welche Faktoren die Zustimmung zum Zusammenspielen von Befragungs- und Prozessdaten beeinflussen und so zu Selektivitäten im zusammengespielten Datensatz führen würden. Unterschieden wurde zwischen allgemeinen Faktoren, die unabhängig vom konkreten Gegenstandsbereich der Befragung wirken können und untersuchungsspezifischen Faktoren. Aus der Untersuchung ergeben sich Hinweise, dass in verknüpften Datensätzen nichtdeutsche Staatsbürger, Frauen, Westdeutsche und Personen mit eher niedrigem Einkommen unterrepräsentiert sein würden. Bei den untersuchungsspezifischen Faktoren zeigten sich Anhaltspunkte dafür, dass Personen, die dem Untersuchungsgegenstand (Förderung) bzw. der durchführenden Institution positiv gegenüberstehen, in Datensätzen mit Datenverknüpfung auf Basis expliziter Zustimmung der Betroffenen stärker vertreten

sein würden. Dies ist zu beachten, wenn solche Daten zur Bewertung bestimmter staatlicher Maßnahmen verwendet werden.

Die vorliegenden Ergebnisse sprechen allerdings dafür, dass das Ausmaß der Selektivität und entsprechende Auswirkungen auf konkrete Analysen eher gering sind. Zwar weist der Chi-Quadrat-Test auf eine signifikante Modellverbesserung durch die eingeführten erklärenden Variablen hin, allerdings zeigen die zwei beispielhaft durchgeführten Analysen, dass trotz des signifikanten Einflusses bestimmter Merkmale die Selektivität nicht unbedingt hoch genug ist, um sich verzerrend auf die Ergebnisse auszuwirken. Es hat sich gezeigt, dass die Nichtberücksichtigung von Personen mit Datenschutzbedenken zumindest in den beiden hier berichteten Analysen keinen systematischen Einfluss auf die Analyseergebnisse hat. Obwohl sich die Ergebnisse nicht unbedingt dahingehend verallgemeinern lassen, dass das Zusammenspielen von Prozess- und Befragungsdaten nie zu selektiven Ausfällen führen wird, sprechen sie doch dafür, dass sich das Problem zumindest in Grenzen hält. Weitere Untersuchungen, eventuell mit eigens zur Untersuchung der Zustimmungsselektivität konzipierten Befragungen, wären nötig, um zu Aussagen mit höherem Allgemeinheitsgrad zu gelangen. Bis dahin können Selektivitätsanalysen wie die hier durchgeführten in jedem Fall helfen, das Ausmaß der Selektivität abzuschätzen.

Kurzfassung

Die Verknüpfung von Daten aus verschiedenen Quellen kann Schwächen der einzelnen Datenarten ausgleichen und so das Analysepotenzial deutlich erhöhen. Angesichts begrenzter finanzieller Mittel zur Durchführung von Erhebungen und sinkender Ausschöpfungsquoten gewinnt die Verknüpfung von Daten aus verschiedenen Quellen zusätzlich an Attraktivität.

Für die Arbeitsmarktforschung ist neben der Verknüpfung verschiedener Prozess- oder Registerdaten der Bundesagentur für Arbeit (BA) untereinander vor allem die Verknüpfung prozessproduzierter Daten der BA mit Befragungsdaten relevant. Das hieße beispielsweise, zu erhobenen Befragungsdaten die in der Bundesagentur für Arbeit angefallenen Prozessdaten hinzuzuspielen oder umgekehrt die Registerdaten der BA mit Informationen aus Befragungen anzureichern.

Vorteile einer solchen Verknüpfung von Informationen aus Prozess- und Befragungsdaten auf Ebene des Individuums (exakte Verknüpfung) sind:

- Zunächst stehen für jedes Individuum mehr Informationen als bei der Verwendung einer einzelnen Datenquelle zur Verfügung.
- Daneben liegen die Vor- und Nachteile beider Datenarten auf unterschiedlichen Gebieten, sodass die Merkmale jeweils aus der Datenquelle herangezogen werden können, welche die valideren und für die jeweilige Forschungsfrage passenderen Informationen liefert.
- Da die Befragung zeitnah Daten liefert, die Aufbereitung von Prozessdaten zumindest als „Scientific Use File“ dagegen teilweise Jahre in Anspruch nimmt, können durch die Verknüpfung von Befragungsdaten mit bereits vorliegenden Prozessdaten früher Ergebnisse berichtet werden.
- Schließlich müssen bestimmte Informationen im Rahmen der Befragung nicht zusätzlich erhoben werden, da sie in den Prozessdaten bereits vorliegen. Die damit verbundene kürzere Interviewdauer erhöht die Bereitschaft zu Teilnahme an Befragungen, mindert das Risiko von Interviewabbrüchen und verringert die Befragungskosten.

Allerdings müssen die Befragten aus Gründen des Datenschutzes einer solchen Datenverknüpfung zustimmen. Dabei besteht eventuell das Problem, dass die Bereitschaft dazu systematisch mit untersuchungsrelevanten Eigenschaften der Personen variiert, was dann zu einer Art „Verknüpfungsbias“ der Untersuchung führt. Bestehen z. B. geschlechtsspezifische Unterschiede hinsichtlich der Zustimmungsbereitschaft, wird sich dies auch in einer diesbezüglichen Selektivität der resultierenden Teilstichprobe auswirken. Da Analysen, die das gesamte Spektrum der Prozess- und Befragungsdaten multivariat nutzen wollen, nur Personen berücksichtigen können, die der Verknüpfung zustimmen, ist zu fragen, ob eine solche Beschränkung der Analysegesamtheit nicht zu Problemen führt, welche die Vorteile der Datenverknüpfung wieder zunichte machen. Diese Fragestellung wurde bisher weder systematisch verfolgt, noch gab es dazu geeignete Daten.

Vor diesem Hintergrund wird in der vorliegenden Arbeit analysiert, in welchem Ausmaß und mit welchen Konsequenzen es bei der exakten Verknüpfung auf Basis ausdrücklicher Zustimmung der Betroffenen zur selektiven Zusammensetzung der Stichprobe kommt. Als Datenquelle dient eine Befragung, die TNS Infratest Sozialforschung im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit zur Evaluation der Kombilohnmaßnahme „Mainzer Modell“ durchgeführt hat. Dabei wurden Geförderte und Personen einer Kontrollgruppe u. a. gefragt, ob sie der Verknüpfung der Befragungsdaten mit den Prozessdaten der BA zustimmen.

Unter Verwendung einer statistischen Mehrebenenanalyse zeigen wir, dass zumindest teilweise mit einer Selektivität der Zustimmung zur Datenverknüpfung zu rechnen ist. So werden Frauen und nichtdeutsche Staatsbürger wohl tendenziell in verknüpften Datensätzen ebenso unterrepräsentiert sein wie Personen in unteren Einkommensgruppen

und solche, die generell eher am Schutz privater Informationen interessiert sind, wie z. B. zum Einkommen oder zum Bezug von Sozialleistungen. Die Ergebnisse zum Einfluss der Interviewereigenschaften verweisen darauf, dass der gezielte Einsatz von älteren Interviewern und solchen mit hoher Bildung die allgemeine Zustimmungsbereitschaft noch erhöhen könnte. Die Dauer des Interviews scheint hingegen keinerlei Einfluss auf die Bereitschaft zur Datenverknüpfung zu haben.

Bei den untersuchungsspezifischen Einflüssen zeigte sich, dass Personen, die mit der Förderung zufrieden sind, in der verknüpften Stichprobe überrepräsentiert sind. Ebenso stimmen Sozialleistungsbezieher häufiger der Verknüpfung zu. Dies begründet zumindest die Vermutung, dass Personen, die eine positive Einstellung zum Untersuchungsgegenstand haben, etwa weil sie positive Erfahrungen mit der auftraggebenden Institution gemacht haben, in verknüpften Stichproben überrepräsentiert sind. Je nach Fokus der Befragung bzw. den mit den Daten durchgeführten Analysen kann dies problematisch sein oder auch nicht.

Die Ergebnisse der Analyse sprechen allerdings auch dafür, dass das Ausmaß der Selektivität und entsprechende Auswirkungen auf konkrete Analysen eher gering sind. Zwar weist der Chi-Quadrat-Test auf eine signifikante Modellverbesserung durch die eingeführten erklärenden Variablen hin, allerdings zeigen zwei beispielhaft durchgeführte Analysen zu inhaltlichen Fragestellungen, dass trotz des signifikanten Einflusses bestimmter Merkmale auf die Zuspielerlaubnis die Selektivität nicht unbedingt hoch genug ist, um sich verzerrend auf die inhaltlichen Ergebnisse auszuwirken. Es hat sich also gezeigt, dass die Nichtberücksichtigung von Personen mit Datenschutzbedenken zumindest in den beiden hier berichteten Analysen keinen systematischen Einfluss auf die Analyseergebnisse hat. Obwohl sich die Ergebnisse nicht unbedingt dahingehend verallgemeinern lassen, dass das Zusammenspielen von Prozess- und Befragungsdaten nie zu selektiven Ausfällen führen wird, sprechen sie doch dafür, dass sich das Problem zumindest in Grenzen hält. Weitere Untersuchungen, eventuell mit eigens zur Untersuchung der Zustimmungsselektivität konzipierten Befragungen, wären nötig, um zu Aussagen mit höherem Allgemeinheitsgrad zu gelangen. Bis dahin können Selektivitätsanalysen wie die hier durchgeführten in jedem Fall helfen, das Ausmaß der Selektivität abzuschätzen.

Executive summary

Linking data from different sources can be used to compensate for the weaknesses characteristic for the single sources and can therefore significantly increase the potential for analyses. In addition, due to financial restrictions regarding the realization of surveys and facing declining response rates linking data becomes more and more attractive.

For labour market research administrative data based on different processes in the "Bundesagentur für Arbeit" (BA) can be linked, but besides that the possibility to link administrative data to survey data is very important. Both directions are possible: Linking administrative data to survey data and vice versa.

Advantages of linking data from different sources on an individual level ("record linkage") are:

- There is more information per individual, compared to using only one data set.
- Different data sources have different advantages and disadvantages. So each time information can used from the source delivering better information regarding the research question of interest.
- Surveys deliver data just in time, whereas generating scientific use files out of administrative data may take several years. So linking survey data to already existing administrative data helps reporting results earlier.
- As certain information already exists in administrative data it has not to be collected in the survey. So the survey interview is shorter, what increases the willingness to participate, reduces the risk of termination, and so reduces the survey costs.

Due to reasons of data protection the respondents have to be asked to allow linking their survey data to their administrative data. This necessity may result in a kind of "linkage bias", as it may be, that their willingness to admit correlates with other characteristics relevant for the research question. As especially multivariate analyses use a wide range of administrative and survey data they are restricted to the respondents that allowed linking data. If for example there are gender-specific differences in allowing linking data the restricted sample that can be used for the multivariate analyses reflects this bias. This leads to the question, whether this restriction implies consequences that cancel out the advantages of record linkage. As adequate data were missing, so far there is no systematic research on this question.

Considering this background the present contribution analyses to which extent and with which consequences the necessity to ask for admission for record linkage results in a bias regarding the restricted sample. The survey data used here originate from a study aiming at evaluating the effectiveness of in-work benefits in order to integrate unemployed people characterised by one or more obstacles to placement ("Mainzer Modell"). The survey was realised by TNS Infratest on behalf of the "Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit". In this study persons subsidized and persons of a control group among others were asked, whether they allow linking the survey data to administrative data.

Using multilevel analysis we show that at least in part there is selection in allowing data linking. So women, foreign persons, persons with low income and persons interested in protecting private information as for example on income or social benefits tend to be underrepresented in the restricted sample using the record linked data set. Regarding interviewer characteristics we find that older and higher educated interviewers reach higher admission rates. The duration of the interview seems not to influence the willingness to admit linking of data.

Regarding study-specific characteristics our results show that persons satisfied with the subsidy as well as persons receiving social security benefits more often agree. These results may be due to positive experience with the BA, the institution on behalf of which the study was conducted and may reflect a positive attitude towards BA. Depending on the research question this may be a problem or not.

However the results also show that the extent of the selection bias as well as the consequences for specific analyses are rather small. Although a chi square test indicates a significant model improvement in explaining the admission of data linking, two examples with specific research questions show that selectivity is not necessarily high enough to distort research results. That is, in the research examples excluding persons that didn't admit the record linkage had no systematic influence on the results. Although these findings cannot be generalised in the sense, that linking survey and administrative data never leads to biased results, they suggest that the problem is restricted. Further investigation using specifically designed studies are necessary to enable more generalisable results. Until then analyses regarding selectivity like the ones presented in this contribution can help to estimate the extent of the bias.

Danksagung Wir danken Johannes Ludsteck, Roman Lutz, Susanne Rässler, Helmut Rudolph, Mark Trappmann und Katja Wolf sowie den zwei anonymen Gutachtern dieser Zeitschrift für anregende Diskussionen und hilfreiche Kommentare.

Anhang A

	Mit Zusammenspielen einverstanden		Test auf Gleichheit der Mittelwerte
	Ja	Nein	t-Wert
Interviewer:			
Geschlecht weiblich	0,712	0,703	0,58
Alter:			
Bis unter 25 Jahre	–	–	–
25 bis 34 Jahre	0,193	0,218	−1,85
35 bis 44 Jahre	0,098	0,106	−0,77
45 bis 54 Jahre	0,073	0,068	0,54
55 Jahre oder mehr	0,141	0,079	5,49
Bildung			
Sonst	–	–	–
FHR, Abitur	0,509	0,484	1,49
Befragte:			
Alter			
Bis unter 25 Jahre	–	–	–
25 bis 34 Jahre	0,263	0,241	1,54
35 bis 44 Jahre	0,386	0,366	1,22
45 bis 54 Jahre	0,157	0,170	−1,05
55 Jahre oder mehr	0,048	0,066	−2,35
Angabe fehlt	0,001	0,002	−1,12
Deutsche Staatsangehörigkeit	0,921	0,900	2,24
Bildung			
Kein Hauptschulabschluss	–	–	–
Volks-/Hauptschule	0,368	0,352	1,02
Mittlere Reife/POS 10. Klasse	0,417	0,403	0,85
Abi/FHS	0,158	0,178	−1,59
Berufsausbildung: Ausbildung/Studium	0,698	0,691	0,51
Arbeitslosigkeit vor Referenzzeitpunkt	–	–	–
Nicht arbeitslos	–	–	–
Bis unter 12 Monate	0,581	0,595	−0,84
12 Monate oder mehr	0,246	0,285	−2,66
Angaben fehlen	0,045	0,028	2,41
Erwerbsstatus Beschäftigungsaufnahme: Erwerbstätig	0,900	0,908	−0,73
Nicht kombilohngefördert	–	–	–
Weniger zufrieden mit Förderhöhe, keine Angabe	0,122	0,156	−3,00
Sehr/zufrieden mit Förderhöhe	0,539	0,421	7,01
Bruttoerwerbseinkommen Befr. nach Referenzzeitpunkt			
Nicht vorhanden	–	–	–
1–325 EURO	–	–	–
326–511 EURO	0,108	0,102	0,62
512–1023 EURO	0,368	0,322	2,84
1024 EURO oder mehr	0,208	0,159	3,69
Angabe fehlt	0,052	0,140	−9,90
HLU vor/nach Referenzzeitpunkt			
Weder-noch	–	–	–
Vor und nach	0,098	0,059	3,98
Nur nach	0,013	0,010	0,73
Nur vor	0,134	0,098	3,14
Mindestens einmal keine Angabe	0,006	0,032	−6,73
Geschlecht weiblich	0,685	0,707	−1,40
Lebensform: mit Partner	0,449	0,467	−1,08
Erwerbstätigkeit befragte Person: ja	0,841	0,797	3,44
Öffentlicher Dienst: ja	0,060	0,055	0,57
Erwerbstätigkeit Partner/in: ja	0,197	0,209	−0,82
Region Ostdeutschland	0,347	0,328	1,17

Literatur

Achatz, J., Gartner, H., Glück, T.: Bonus oder Bias? Mechanismen geschlechtsspezifischer Entlohnung. Kölner Z. Soziol. Sozialpsychologie **57**(3), 466–493 (2005)

Blau, P.M.: Social Exchange. In: Sills, D.L. (ed) International Encyclopedia of the Social Sciences, vol. 7, pp. 452–457. Macmillan, New York (1968)

Eichhorst, W., Gartner, H., Krug, G., Rhein, T., Wiedemann, E.: Niedriglohnbeschäftigung in Deutschland und im internationalen Vergleich. In: Allmendinger, J., Eichhorst, W., Walwei, U. (eds) IAB Handbuch Arbeitsmarkt: Analysen, Daten, Fakten, pp. 107–142. Campus Verlag, Frankfurt am Main (2005)

Engel, U., Pötschke, M., Schnabel, C., Simonson, J.: Nonresponse und Stichprobenqualität. Ausschöpfung in Umfragen der Markt- und Sozialforschung. Verlagsgruppe Deutscher Fachverlag, Frankfurt am Main (2004)

Esser, H.: Über die Teilnahme an Befragungen. ZUMA-Nachrichten **18**, 38–47 (1986)

Esser, H.: „Habits", „Frames" und „Rational Choice". Die Reichweite der Theorie der rationalen Wahl (am Beispiel der Erklärung des Befragtenverhaltens). Z. Soziol. **19**, 231–247 (1990)

Fetchenhauer, D.: Möglichkeiten und Grenzen von Rational Choice – Erklärungen für betrügerisches Verhalten am Beispiel des Versicherungsbetruges. Kölner Z. Soziol. Sozialpsychologie **51**(2), 283–312 (1999)

Gewiese, T., Hartmann, J., Krug, G., Rudolph, H.: Das Mainzer Modell aus der Sicht der Arbeitnehmer und Betriebe. Befunde aus der Begleitforschung. http://doku.iab.de/externe/2004/k040823w09.pdf (Stand: 19. September 2008)

Greene, W.H.: Econometric Analysis. Pearsons Education, Inc., Upper Saddle River, New Jersey (2003)

Groves, R.M., Fowler, F.J., Couper, M., Lepkowski, J.M., Singer, E., Tourangeau, R.: Survey Methodology. John Wiley & Sons, Hoboken, New Jersey (2004)

Guo, G., Zhao, H.: Multilevel Modelling for Binary Data. Annu. Rev. Sociol. **26**, 441–462 (2000)

Hakim, C.: Research based on administrative records. Sociol. Rev. New Series **31**, 489–519 (1983)

Hamann, S., Krug, G., Köhler, M., Ludwig-Mayerhofer, W., Hacket, A.: Die IAB-Regionalstichprobe 1975–2001: IABS-RO1. ZA-Information **55**, 34–59 (2004)

Heckman, J.J., LaLonde, R.J., Smith, J.A.: The Economics and Econometrics of Active Labor Market Programs. In: Ashenfelter, O., Card, D. (eds) Handbook of Labor Economics, vol. 3A. Elsevier Science B. V., Amsterdam (1999)

Holland, P.W.: Statistics and causal inference. J. Am. Statistical Assoc. **81**(396), 945–960 (1986)

Homans, G.C.: Bringing man back. Am Sociol Rev **29**, 809–818 (1964)

Hummel, E., Jacobebbinghaus, P., Kohlmann, A., Oertel, M., Wübbeke, C., Ziegerer, M.: Stichprobe der Integrierten Erwerbsbiografien: IEBS 1.0, Handbuch-Version 1.0.0. FDZ Datenreport Nr. 06/2005, Nürnberg (2005)

Jenkins, S.P., Lynn, P., Jäckle, A., Sala, E.: Linking household survey and administrative record data: what should the matching variables be? DIW Discussion Papers 489. German Institute of Economic Research, Berlin (2005)

Kaltenborn, B., Krug, G., Rudolph, H., Weinkopf, C., Wiedemann, E.: Evaluierung der arbeitsmarktpolitischen Sonderprogramme CAST und Mainzer Modell, Forschungsbericht Nr. 552. Bundesministerium für Wirtschaft und Arbeit, Kaltenborn, Berlin (2005)

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik: Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik. Nomos, Baden-Baden (2001)

Pascale, J., Mayer, T.S.: Exploring confidality issues related to dependent interviewing: preliminary findings. J. Off. Statistics **20**, pp. 357–377 (2004)

Pickery, J., Loosveldt, G., Carton, A.: The effects of interviewer and rtespondent characteristics on response behavior in panel surveys. Sociol. Methods. Res. **29**, 509–523 (2001)

Rässler, S.: Statistical Matching. A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York Berlin (2002)

Riker, W.H., Ordeshook, P.C.: An Introduction to Positive Political Theory. Prentice-Hall, Englewood Cliffs, NJ (1973)

Schmähl, W., Fachinger, U.: Prozessproduzierte Daten als Grundlage für sozial- und verteilungspolitische Analyse – Erfahrungen mit Daten der Rentenversicherungsträger für Längsschnittanalysen. In: Deutsche Forschungsgemeinschaft (eds) Mikroanalytische Grundlagen der Gesellschaftspolitik, vol. 2, pp. 179–200. Akademie Verlag, Berlin (1994)

Schnabel, A.: Teilnahmeverhalten bei Unternehmensbefragungen. Arbeit **6**(2), 154–172 (1997)

Schräpler, J.-P.: Respondent behavior in panel studies. A case study for income-nonresponse by means of the German socio-economic panel (SOEP). Sociol. Methods Res. **33**, 118–156 (2004)

Singer, E., Mathiowetz, N., Couper, M.: The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. census. Public Opinion Quarterly **57**, 465–482 (1993)

Tomaskovic-Devey, D., Leiter, J., Thompson, S.: Item nonresponse in organizational surveys. Sociol Methodol **25**, 77–110 (1995)

Windzio, M.: Kann der regionale Kontext zur „Arbeitslosenfalle" werden? Kölner Z. Soziol. Sozialpsychologie **56**, 257–278 (2004)

Wippler, R.: Kulturelle Ressourcen, gesellschaftlicher Erfolg und Lebensqualität. In: Giesen, B., Haferkamp, H. (eds) Soziologie der sozialen Ungleichheit, pp. 221–254. Westdeutscher Verlag, Opladen (1987)

Wirth, H., Müller, W.: Mikrodaten der amtlichen Statistik als eine Datengrundlage der empirischen Sozialforschung. In: Diekmann, A. (eds) Methoden der Sozialforschung (Sonderheft 44 der Kölner Z Soziol Sozialpsychologie), pp. 93–127. VS Verlag für Sozialwissenschaften, Wiesbaden (2004)

Dr. Josef Hartmann, Studium der Soziologie an der Universität Mannheim, 1990 Abschluss als Diplom-Soziologe. Von 1991 bis 1996 wissenschaftlicher Mitarbeiter an der Universität Mannheim und von 1996 bis 1997 an der Ruprecht-Karls-Universität Heidelberg.

2002 Promotion zum Dr. phil. an der Ruprecht-Karls-Universität Heidelberg.

Seit 1997 Studienleiter bei TNS Infratest Sozialforschung im Bereich „Arbeitsmarktforschung", seit 2008 Leiter des Bereichs „Arbeitsmarktforschung".

Forschungsfelder: Studien zur Evaluation arbeitsmarktpolitischer Programme und arbeitsrechtlicher Normen, Untersuchungen zu Arbeitsbedingungen, insbesondere Belastungen und Arbeitszeit, Untersuchungen zur Bereitstellung arbeitsmarktstatistischer Grunddaten zur Ergänzung der amtlichen Statistik sowie Analysen zu methodologischen Fragen.
josef.hartmann@tns-infratest.com

Gerhard Krug, Studium der Soziologie an der Universität Bamberg, 2003 Abschluss als Diplom-Soziologe. Von 2003–2004 und seit 2006 wissenschaftlicher Mitarbeiter im Institut für Arbeitsmarkt. 2005–2007 Stipendiat im IAB-Graduiertenprogramm.

Forschungsfelder: Quantitative Methoden der Arbeitsmarktforschung; Messung kausaler Effekte; soziologische Handlungstheorie.
Gerhard.krug@iab.de