➢ Begutachteter Beitrag ➣

Fragebogentests als Mittel der Qualitätssicherung

Aktuelle Beispiele aus Mikrozensus und EU-SILC

KATRIN BAUMGARTNER
ESTHER GREUSSING
ANNELIESE OISMÜLLER
MARC PLATE
VLASTA ZUCHA

Fragebögen spielen vielfach eine zentrale Rolle im Datenerhebungsprozess. Um die Qualität und die angestrebten Gütekriterien der mit Fragebögen erhobenen Daten sicherzustellen, sind umfassende Fragebogentests unerlässlich. Der vorliegende Beitrag zeigt eine Auswahl an Testmethoden, die bei Statistik Austria in den Projekten Mikrozensus und EU-SILC angewendet wurden. Nach einer allgemeinen Skizzierung des jeweiligen Testverfahrens werden die konkrete Vorgehensweise sowie die Konsequenzen aus dem Test vorgestellt; den Abschluss bildet ein Fazit zur jeweiligen Testmethode. Ziel dieses Beitrags ist es, die Relevanz und das Potential von Fragebogentests aufzuzeigen. Dem Testen von Fragebögen soll in der amtlichen Statistik eine noch höhere Bedeutung beigemessen werden. Es wird angestrebt, das Testen als Teil der Qualitätssicherung bei Statistik Austria auch institutionell zu verankern sowie noch stärker in bestehende Arbeitsprozesse zu integrieren.

Einleitung

Bei Statistik Austria werden statistische Daten auf unterschiedliche Art und Weise erhoben. Ein zentrales Instrument zur Datenbeschaffung ist dabei der Fragebogen, der in persönlichen und telefonischen Befragungen eingesetzt wie auch von Respondenten und Respondentinnen selbst ausgefüllt wird. Da die Erstellung hochwertiger Statistiken und Analysen angestrebt wird, besteht der Anspruch, dass Fragebögen gewisse Qualitätsstandards erfüllen.[1]) Die Erfüllung bestimmter Qualitätskriterien kann sichergestellt werden, indem Fragebogentests im Zuge des Erhebungsprozesses eingesetzt werden. Sie untersuchen den Fragebogen als Messinstrument auf unterschiedliche Messfehler, die sich etwa in der Reliabilität oder Validität der Daten niederschlagen können.[2]) Fragebogentests sind damit für die Sicherstellung eines gewissen Qualitätsstandards unerlässlich.

Neben technischen Tests, die etwa Programmierfehler im Fragebogen (z.B. falsche Filterführung) aufdecken sollen, gibt es eine Vielzahl an inhaltlichen Fragebogentests. Jedes Testverfahren hat dabei eine bestimmte Schwerpunktsetzung und kann unterschiedlichen Fehlerarten nachgehen. Kognitive Interviews eignen sich beispielsweise, um relativ früh im Entstehungsprozess eines Fragebogens herauszufinden, wie der Frage-Antwort-Prozess funktioniert, d.h. wie Befragte Fragen verstehen und auf Basis welcher Überlegungen sie ihre Antworten geben. Mit den Erkenntnissen aus kognitiven Interviews können Fragen adaptiert werden, sodass Validitätsfehler reduziert und damit die Qualität der erhobenen Ergebnisse gesteigert werden. Standardisierte Testbefragungen eigenen sich hingegen, wenn bestimmte Problemlagen schon im Vorhinein vermutet werden oder aus früheren Befragungen bekannt sind. Je nach Erkenntnisinteresse finden damit unterschiedliche Fragebogentests Anwendung.

Der vorliegende Beitrag zeigt, wie das Ziel der Qualitätssicherung von Fragebögen bei Statistik Austria im Rahmen der zur Verfügung stehenden Möglichkeiten verfolgt wird. Dabei wird eine Auswahl an bisher durchgeführten Tests aus den Projekten Mikrozensus und EU-SILC vorgestellt:

- Kognitive Interviews (Mikrozensus-Arbeitskräfteerhebung)
- Respondent-Debriefing (Mikrozensus-Wohnungserhebung)
- Debriefing von Interviewern und Interviewerinnen (EU-SILC)
- Standardisierte Evaluationsbefragung (Mikrozensus-Arbeitskräfteerhebung)

Im Folgenden wird das Ziel der jeweiligen Testverfahren zunächst allgemein beschrieben, um dann einen Rahmen für die anschließende Darstellung der konkreten Vorgehensweise sowie die aus dem Test hervorgegangenen Konsequenzen darzustellen. In einer rückblickenden Zusammenfassung wird ein Fazit gezogen, welches die Stärken und Schwächen der jeweiligen Methode hervorhebt. Das letzte Kapitel des vorliegenden Beitrags zeigt schließlich eine Zusammenschau und Beurteilung aller dargestellten Testmethoden und geht auf Optimierungspotentiale hinsichtlich der Qualitätssicherung von Fragebögen ein.

Kognitive Interviews bei der Mikrozensus-Arbeitskräfteerhebung

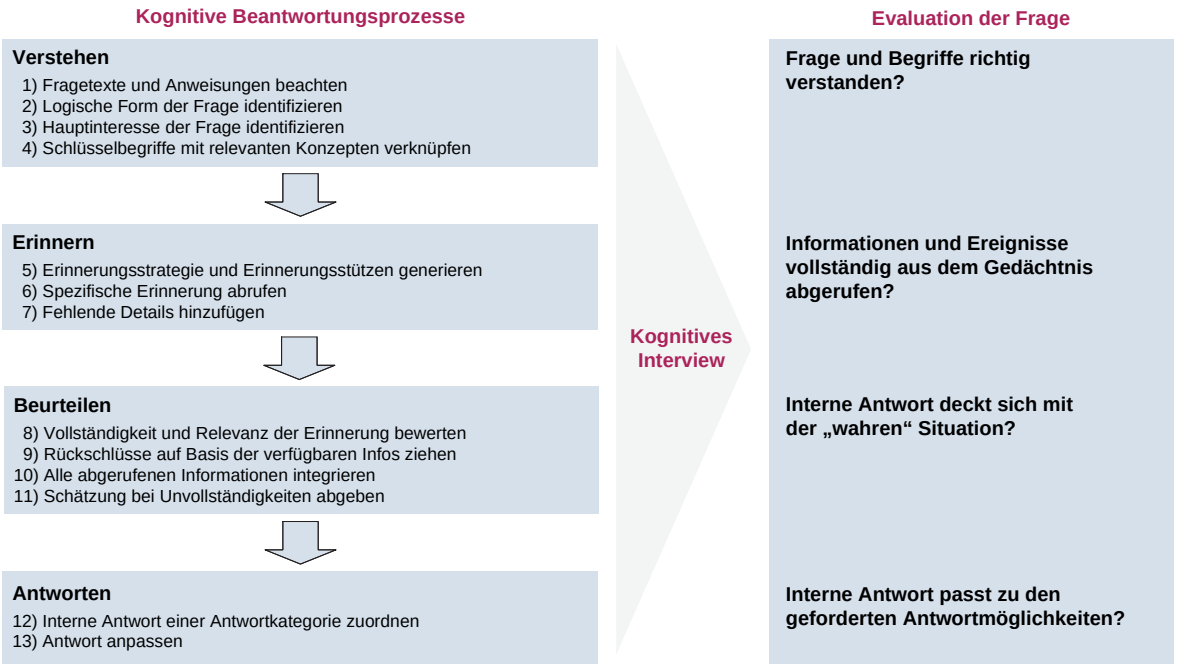
Hintergrund und Ziele

Bis eine Antwort auf eine Frage in einem standardisierten Fragebogen gefunden und schließlich gegeben wird, spielt sich im Kopf der Respondenten und Respondentinnen eine Vielzahl an gedanklichen Prozessen ab. Diese kognitiven

[1]) Siehe dazu die Qualitätsrichtlinien von Statistik Austria. Diese unterliegen einer regelmäßigen Aktualisierung und sind auf der Website von Statistik Austria einsehbar *(Statistik Austria 2012)*.

[2]) Reliable Daten sind unabhängig vom Zeitpunkt der Messung. Ein valides Messinstrument misst das, was es – gemäß der zugrundeliegenden Konzepte – messen sollte *(Schnell et al. 2008)*.

Kognitive Interviews zur Evaluation von Fragen



Q: STATISTIK AUSTRIA. Eigene Darstellung in Anlehnung an *Tourangeau et al.* (2012) und *Prüfer & Rexroth* (2005).

Beantwortungsprozesse werden unterschiedlich schnell, oft aber in nur wenigen Sekunden, durchlaufen. Viele sind den Befragten nicht bewusst, und sie laufen auch nicht immer nach exakt dem gleichen Schema ab *(Tourangeau et al. 2012)*. Dennoch lässt sich der Frage-Antwort-Prozess modellhaft verallgemeinern: Nach *Tourangeau et al.* (2012) finden bei der Beantwortung von Fragen idealtypisch 13 unterschiedliche Denkprozesse statt, welche in vier Stufen – Verstehen, Erinnern, Beurteilen und Antworten – durchlaufen werden, bevor die Antwort erfolgt *(vgl. Abbildung)*. Je nachdem, wie gut eine Frage designt ist, stellt sie eine geringe oder große Herausforderung an diese Beantwortungsprozesse. Je größer die Herausforderungen in den Beantwortungsprozessen, desto größer ist die Gefahr geringer Validität und/oder geringer Reliabilität bei einer Frage.

Kognitive Interviews sind eine qualitative Forschungsmethode mit dem Ziel, diese Denkprozesse ans Licht zu bringen. Sie ermöglichen es, sehr genau festzustellen, welche Schwierigkeiten an welcher Stelle der Frage auftreten könnten. Vorrangig eignet sich die Testmethode daher, um Messfehler im Zusammenhang mit Respondenten und Respondentinnen und deren Informationssystemen[3]) aufzudecken. Die Ergebnisse aus kognitiven Interviews liefern ein sehr detailliertes Bild über die Qualität einer konkreten Frage und haben daher auch das Potential, konkrete Handlungsempfehlungen zur Verbesserung einer Frage zu geben.

Zudem können kognitive Interviews auch dabei helfen, eine weitere Fehlerquelle für Messfehler, nämlich Defizite in der Fragebogenstruktur, aufzuzeigen. Die offen geführten Interviews bieten Raum für Anmerkungen zu Fragereihenfolgen, Filterführungen, Layouts oder sonstigen Elementen des Fragebogens *(Willis 1999, S. 19)*.

Zusätzlich zur Messfehler-Evaluation lässt sich mit Hilfe kognitiver Interviews auch ein Urteil über konzeptionelle Fehler bilden. Aufgrund der offenen Gesprächsführung während der Interviews kann nämlich das Expertenwissen zum Thema vergrößert werden: Die inhaltlich stark in die Tiefe gehenden Gespräche bringen in der Regel viele Details über das zu messende Thema zu Tage, die direkt aus der Erfahrungswelt der Betroffenen stammen. Das bietet den fachlichen Experten und Expertinnen die Chance, bestimmte Facetten des Themas neu oder aus anderen Perspektiven aufzunehmen und mit den zu messenden Konzepten zu vergleichen *(Willis 1999, S. 19)*.

Vorgehensweise am Beispiel Mikrozensus-Arbeitskräfteerhebung[4])

Statistik Austria führte 2013 im Rahmen des Eurostat-Grants „Quality Improvement of the LFS: Testing Labour Status" kognitive Interviews über potentiell neue Fragen zur Messung des Erwerbstatus durch.[5]) Als praktisches Beispiel für die Durchführung kognitiver Interviews soll nachfolgend die methodische Vorgehensweise dieses Pretests geschildert werden.

Schritt 1: Übersetzung des Musterfragebogen ins Deutsche

Der zu testende Fragebogenvorschlag wurde von der Task Force zur „Harmonisierung der Messung von Erwerbstätig-

[3]) Mit Informationssystemen sind hier die Quellen gemeint, aus denen die Befragten ihre Informationen beziehen. Beispielsweise das eigene Gedächtnis, schriftliche Unterlagen, elektronische Datenbanken etc.

[4]) Der Labour Force Survey (LFS) wird in Österreich im Rahmen des Mikrozensus durchgeführt *(Kytir & Stadler 2004)*.

[5]) Siehe *Statistik Austria* (2014) für den detaillierten Abschlussbericht und *Fasching (2015)* für eine Zusammenfassung der Ergebnisse.

keit und Arbeitslosigkeit“ (TF HMEU)[6]) mit dem Ziel entwickelt, bei der Datenerhebung eine größere Standardisierung über die Mitgliedstaaten zu erreichen. Die Verwendung derartiger Musterfragebögen ist als ein sinnvoller und wichtiger Schritt in Richtung einer Input-Harmonisierung bei der Datenerhebung zu werten. Da der Musterfragebogen aber auf Englisch verfasst war, mussten die zu testenden Fragen in einem ersten Schritt ins Deutsche übersetzt werden.

Aus mehreren Gründen gestaltet sich die Übersetzung des Fragebogens als komplexes Unterfangen: Erstens besteht ein Fragetext oft aus mehreren Schlüsselbegriffen. Werden diese Schlüsselbegriffe direkt übersetzt, kann es passieren, dass sie den ursprünglichen Inhalt nicht mehr auf die gleiche Weise transportieren. Zum Beispiel konnte für den englischen Schlüsselbegriff „job“ kein vergleichbar umfassender deutscher Begriff gefunden werden. Zweitens können bei einer direkten Übersetzung die deutschen Fragetexte aufgrund der Grammatik deutlich länger oder komplizierter als das englische Original werden. Drittens bestehen Fragetexte manchmal auch aus Redewendungen, die bei einer direkten Übersetzung ihren Sinn verlieren würden.

Die Übersetzung des Fragebogens wurde daher in einem iterativen Prozess von Experten und Expertinnen des Mikrozensus in Teamarbeit durchgeführt. Zuerst wurde von einer Person ein Fragetext-Entwurf entwickelt, der so nah am englischen Original wie möglich sein sollte. Dieser Entwurf wurde dann im Dreierteam so lange diskutiert, bis eine Variante entstand, die das Kriterium „einheitliches Frageverständnis über alle Respondenten und Respondentinnen hinweg“ bestmöglich erfüllen sollte. Verglichen mit dem englischen Original mussten daher bereits bei der Übersetzung Adaptionen in den Fragetexten durchgeführt werden.

Schritt 2: Identifizierung möglicher Frageprobleme und Erstellung eines Interviewleitfadens

Die ins Deutsche übersetzten Fragetexte wurden vom Testteam dahingehend untersucht, an welchen Stellen eine Gefahr für die vier Stufen des Frage-Antwort-Prozesses – Verstehen, Erinnern, Beurteilen und Antworten – drohen könnte. Bei der Identifizierung solch möglicher Messfehler waren einerseits die Erkenntnisse aus den zuvor geführten Diskussionen bei der Übersetzung von großem Nutzen. Andererseits musste dafür ein weiterer Arbeitsschritt durchgeführt werden, bei dem sich das Fragebogenteam nochmals intensiv mit den Frageentwürfen und den dahinterliegenden Messkonzepten kritisch auseinandersetzte.

Auf Basis potentieller Frageprobleme wurde nun ein Interviewleitfaden erstellt. Der Interviewleitfaden listet hierbei für jede Frage je nach Problemlage unterschiedliche Nachfragen, sog. Probes. Da beim kognitiven Interview eine Vielzahl von Probe-Varianten[7]) potentiell eingesetzt werden können, galt es in diesem Schritt, die geeignete Kombination an Probe-Varianten pro Frage zu definieren. Ziel war es, mit Hilfe der Probes ein möglichst umfassendes Bild über den Frage-Antwort-Prozess pro Frage zu gewinnen und darüber hinaus zu überprüfen, auf welche Weise die vermuteten Frageprobleme tatsächlich auftreten. Der vollständige Interviewleitfaden, inkl. der verwendeten Probe-Varianten und Probe-Formulierungen kann dem Anhang zu *Statistik Austria 2014* entnommen werden.

Als generelle Interviewtechnik wurde der Ansatz eines halboffenen qualitativen Interviews gewählt *(Froschauer & Lueger 2003)*. Demnach sollten bestimmte Probes immer gleichgestellt werden, sodass zu bestimmten Aspekten einer Frage Aussagen von allen Interviewpartnern und -partnerinnen eingeholt werden konnten. Die Reihenfolge der Probes wurde jedoch nicht genau festgelegt, sondern sollte mit dem Ziel eines guten Gesprächsflusses flexibel gehandhabt werden. Zudem sollte bei jeder Frage standardmäßig auch eine Think-Aloud-Technik, nämlich das Paraphrasieren, Verwendung finden. Darüber hinaus sollte das Interview auch mit einer großen Offenheit gegenüber den Äußerungen der Interviewpartner und -partnerinnen geführt werden: Der Interviewer bzw. die Interviewerin sollte gegebenenfalls spontan, z.B. durch genaueres Nachfragen, auf Gesprächssituationen reagieren. Diese Flexibilität in der Interviewführung ist als große Stärke von kognitiven Interviews anzuführen. Dadurch, dass beispielsweise auf ein Zögern, Stirnrunzeln oder Seufzen des Interviewpartners bzw. der Interviewpartnerin reagiert wird, lassen sich in die Tiefe gehende Einsichten im Frage-Antwort-Prozess ans Licht bringen, die mit anderen Testmethoden wohl verborgen bleiben würden.

Schritt 3: Rekrutierung der Interviewführung

Bei der Auswahl der Interviewpartner und -partnerinnen sollte einerseits eine möglichst große Varianz an unterschiedlichen Perspektiven und andererseits eine sinnvolle Gruppengröße pro Frage, insbesondere in Anbetracht der Filterführungen, erreicht werden. Aus diesem Grund wurde eine bestimmte Anzahl anhand soziodemographischer Charakteristika[8]) vorab definiert. Daraufhin wurden die entsprechenden Personen aus dem Sample der letztmalig teilnehmenden Respondenten und Respondentinnen des Mikrozensus rekrutiert. Erfreulicherweise konnte der Rekrutierungsplan nahezu perfekt erfüllt werden; die Gruppe bestand letztlich aus zehn Personen unterschiedlicher Altersgruppen zwischen 15 und 64 Jahren, je 50% Frauen und Männer. Darunter befanden sich sieben Erwerbstätige (davon eine/r selbständig, eine/r mithelfend, fünf sollten in der Referenzwoche nicht gearbeitet haben) und drei Nicht-Erwerbspersonen. Um den

[6]) Task Force on the Harmonisation of the Measurement of Employment and Unemployment, Final Report 2012.

[7]) Siehe *Willis 2005* für eine Typisierung von kognitiven Pretest-Techniken und Probe-Varianten.

[8]) Die Auswahl basierte auf den Merkmalen Alter, Geschlecht, höchster Bildungsabschluss und Erwerbsstatus.

Reiseaufwand möglichst gering zu halten, wurde eine Einschränkung auf in Wien wohnhafte Personen vorgenommen.

Die Interviews wurden von den beiden Mitgliedern des Testteams bei den Interviewpartnern und -partnerinnen zu Hause durchgeführt. Bei der Terminvereinbarung und vor Beginn des Interviews wurde darüber aufgeklärt, dass es sich um eine Testbefragung zum Zwecke der Verbesserung des Fragebogens handelt. Mit Beginn des Interviews wurden die Gespräche dann auf Tonband aufgezeichnet. Das kürzeste Interview dauerte 34 Minuten, das längste 90 Minuten; insgesamt betrug die durchschnittliche Interviewdauer 59 Minuten. Für die Teilnahme am Interview wurde eine Aufwandsentschädigung in Höhe von 30,00 € ausbezahlt.

Schritt 4: Transkribieren und Analysieren der Interviews

Auf Basis der Tonband-Aufzeichnungen wurden die Gespräche vollständig transkribiert. Das Gesagte wurde wörtlich verschriftlicht. Im Durchschnitt erstreckte sich der zeitliche Aufwand für das Transkribieren auf ca. drei Stunden pro Stunde Tonbandaufnahme und produzierte pro Interview knapp sieben Seiten Text.

Die Analyse der Interviews erfolgte auf Basis der Transkripte. Hierbei wurde das Textmaterial pro Frage und pro Interview zuerst paraphrasiert, dann generalisiert und schließlich reduziert *(Mayring 2015, S. 72)*. Die reduzierten Kernaussagen wurden in ein Analyseschema eingetragen, das pro Frage in den Zeilen alle Interviews und in den Spalten die in Schritt 2 identifizierten Problemlagen listete. Neue Problemlagen, die sich aufgrund der Interviews ergaben, wurden als neue Spalten dem Analyseschema hinzugefügt.

Nach Eintragung der reduzierten Kernaussagen in das Analyseschema wurden pro Frage Kategorien zur Gesamtbewertung der Frage und Kategorien zu den jeweils vorherrschenden Problemlagen gebildet. Zum Beispiel besaß die Gesamtbewertungskategorie „Frageverständnis“ die Ausprägungen „Alle Schlüsselwörter richtig wahrgenommen“, „Schlüsselwörter teilweise richtig wahrgenommen“ und „Frage falsch verstanden“. Anhand der Häufigkeiten und Inhalte der Ausprägungen ließen sich Schlussfolgerungen hinsichtlich potentieller Schwierigkeiten ziehen. Mit Hilfe der detaillierten Ausführungen zu bestimmten Ausprägungen einer Kategorie war es zudem möglich, Handlungsempfehlungen zur Verbesserung der Fragen zu liefern. Beim Festhalten der Ergebnisse galt das Vier-Augen-Prinzip: Die Schlussfolgerungen und Handlungsempfehlungen wurden von einem Teammitglied entworfen, mussten sich aber in der Diskussion mit dem zweiten Teammitglied bewähren.

Konsequenzen des Tests

Mit Hilfe des kognitiven Tests konnten zum einen detaillierte Erkenntnisse zu den einzelnen Fragen gewonnen werden:

- Bezüglich des aktuell in Österreich verwendeten Mikrozensus-Fragebogens konnten mögliche Stolperfallen abgeleitet werden. Inwiefern dies bei betroffenen Fragen zu einer Adaption der Erläuterungstexte führen sollte, wird bei zukünftigen Fragebogenaktualisierungen berücksichtigt werden.
- Bezüglich des Eurostat Musterfragebogens zeigte sich in der Gesamtschau auf die in den teilnehmenden Ländern durchgeführten Tests weiterer Adaptierungsbedarf. Die Task Force HMEU nutzte die Testergebnisse für die Weiterentwicklung des Musterfragebogens und startete 2015 eine dritte Testrunde, um die neuen Fragevarianten abschließend zu evaluieren.

Zum anderen wurden auch allgemeinere strukturelle Schwierigkeiten identifiziert. Diese betrafen überwiegend die den Fragen zugrundeliegenden Messkonzepte sowie die generelle Logik des Fragebogens:

- Es konnten Unklarheiten in den Definitionen mancher Messkonzepte aufgezeigt werden. Eurostat nahm diesbezügliche Hinweise gerne an und arbeitet an einer Schärfung dieser Konzepte.
- Auf struktureller Ebene zeigte sich ein grundsätzliches Problem bezüglich der Unterscheidung von „eine Arbeit grundsätzlich haben“ und „eine Arbeit zu einem bestimmten Zeitpunkt ausüben“. Zudem hatten einige Interviewpartner und -partnerinnen immer wieder Schwierigkeiten beim Zurückerinnern an abgefragte Zeiträume. Dies führte zu Überlegungen über tiefgreifende Änderungen in der Fragereihenfolge und im Konzept der Referenzwoche. Diese Überlegungen werden nun von Eurostat weiter verfolgt.

Darüber hinaus haben die Befunde des österreichischen Tests und den Tests aus anderen Mitgliedstaaten bei Eurostat zu einer Weiterentwicklung beim Thema Vergleichbarkeit von Datenerhebungen geführt: Auf internationaler Ebene entspann sich eine Diskussion darüber, wie Fragebögen zukünftig systematischer getestet werden könnten *(Meertens 2015)*: Sollten die Durchführung kognitiver Tests sowie die Berichterstattung der Testergebnisse stärker standardisiert werden? Und: Sollte es die Rolle eines Testkoordinators für Eurostat oder nationale Statistikämter geben?

Fazit zur Testmethode „Kognitive Interviews“

Basis für eine vielversprechende Durchführung kognitiver Interviews ist die arbeitsintensive Identifizierung potentieller Fehlerquellen bei den Fragen. Ist dieser erste Schritt getan, so lassen sich mit kognitiven Interviews die oft verborgenen Denkprozesse bei der Beantwortung einer Frage sehr gut ans Licht bringen. Aufgrund der offenen Gesprächsführung und den oftmals sehr ausführlichen Aussagen zum Frage-Antwort-Prozess können Details zu Tage gebracht werden, die der Expertengruppe neu sind oder das zu messende Konzept in einem neuen Licht erscheinen lassen. Die Methode der kognitiven Interviews eignet sich daher hervorragend, um potentielle Schwierigkeiten beim Frageverständnis, beim Zurückerinnern oder bei der Findung und Artikulation von

Antworten auf eine sehr anschauliche Weise, nämlich aus Perspektive der Respondenten und Respondentinnen, deutlich zu machen. Durch die genaue Analyse der Gespräche gelingt es, Schwierigkeiten mit der Frage auf ihre Ursachen zurückzuführen, wodurch letztlich Verbesserungsvorschläge für einzelne Fragen und den Fragebogen insgesamt gewonnen werden können.

In erster Linie sind kognitive Interviews somit anzuwenden, wenn Messfehler im Zusammenhang mit Respondenten und Respondentinnen und deren Informationssystemen aufgedeckt werden sollen. Ergänzend kann die Testmethode auch dafür genutzt werden, um Erkenntnisse über die Fragebogenstruktur und die drei Quellen für Konzeptfehler zu gewinnen.

Es liegt aber in der Natur der Testmethode, dass mit den Ergebnissen kognitiver Interviews keine Aussagen über Häufigkeitsverteilungen in der Grundgesamtheit getroffen werden können. Mit kognitiven Interviews lassen sich somit „nur" Hinweise für potentielle Schwierigkeiten finden, die Häufigkeit, mit der sie in der Gesamtbevölkerung auftreten, lässt sich nicht beurteilen. Auch muss klar sein, dass aufgrund der eher kleinen Gruppengröße an Testpersonen nicht alle Schwierigkeiten und auch nicht unbedingt alle gravierenden Schwierigkeiten gefunden werden können *(Blair & Conrad 2011)*. Bei der Durchführung kognitiver Interviews ist zudem zu beachten, dass die verwendeten qualitativen Methoden zur Interviewführung und Analyse der Gefahr der Subjektivität unterliegen. Durch die halboffene Gesprächsführung können die Aussagen der Interviewpartner und -partnerinnen relativ leicht durch die Interviewer und Interviewerinnen beeinflusst werden. So kann es passieren, dass durch eine Aktion der Interviewer und Interviewerinnen, wie z.B. ein vorschnelles Nachfragen, die Gedankengänge der Respondenten und Respondentinnen in Richtungen gelenkt werden, die in der realen Befragungssituation eigentlich nicht stattfinden würden. Umgekehrt können durch das Verpassen wichtiger Nachfragen die Aussagen lückenhaft bleiben.

Auch bei der Analyse der Interviews besteht durch das gelegentliche Auftreten widersprüchlicher oder mehrdeutiger Aussagen eine Schwierigkeit für die Reliabilität der Interpretationen. Aus diesem Grund erscheint es notwendig, kognitive Interviews immer von einem Team, bestehend aus mindestens zwei Personen, durchführen und analysieren zu lassen. Die Mitglieder des Testteams sollten dabei in einem regelmäßig stattfindenden Diskurs über gefundene Ergebnisse stehen. Auch bedarf eine qualitativ hochwertige Durchführung kognitiver Interviews eines hohen Zeitaufwands, da man sich in allen Arbeitsschritten sehr detailliert mit den Inhalten beschäftigt. Vor allem aber zeigt sich, dass neben der Kenntnis qualitativer Forschungsmethoden und Fragebogendesigns die Erfahrung der Beteiligten ein wesentlicher Faktor für die Qualität der Forschungsergebnisse ist. Es erscheint für die effiziente Durchführung kognitiver Interviews daher unerlässlich, ein Testteam organisatorisch so zu verankern, dass sich Wissen und Durchführungserfahrungen zu Fragebogendesign und -testen langfristig aufbauen können.

Respondent-Debriefing bei der Mikrozensus-Wohnungserhebung

Hintergrund und Ziele

Das Respondent-Debriefing ist jenen Testmethoden zuzuordnen, die im Zuge einer Fragebogenerhebung (meist) von Interviewern und Interviewerinnen durchgeführt werden. Es handelt sich dabei – kurz zusammengefasst – um eine Befragung der Befragten zum Fragebogen.

Das primäre Ziel des Respondent-Debriefing besteht darin, zu überprüfen, ob und in welchem Ausmaß das Begriffs- und Frageverständnis der Respondenten und Respondentinnen mit jenem der Forschenden übereinstimmt und ob die zugrundeliegenden Messkonzepte mit Hilfe der Fragen adäquat abgedeckt werden. In erster Linie wird dabei also die Validität der Fragen geprüft. Weiters können auch die von den Befragten zur Antwortgenerierung verwendeten Informationsquellen oder die Erfahrungen während der Interviewsituation mit Hilfe dieser Testmethode nachgefragt werden. Ersteres zielt beispielsweise auf das Entdecken von Messfehlern ab, die aufgrund mangelhafter, inadäquat eingesetzter oder fehlender Informationsquellen entstehen.

Zum Einsatz kommen verschiedene Arten von Nachfragetechniken, die etwa auf das Fragen- bzw. Begriffsverständnis oder auch die Informationsquellen der Befragten abzielen. Die Nachfragetechniken können sowohl quantitative als auch qualitative Informationen zum Frageverständnis hervorbringen, da verschiedene Arten von Fragen zum Einsatz kommen können – etwa offene, halb-offene oder geschlossene Nachfragen. Entscheidend bei der Auswahl und Formulierung der Nachfragen (Probes) ist das Wissen um potentiell problematische Fragen bzw. Fragenbestandteile *(Brancato et al. 2006)*. Mögliche Quellen für dieses Wissen können etwa Befunde aus Testinterviews oder Beobachtungen bereits durchgeführter, früherer Interviews sein.

Üblicherweise wird diese Testmethode – im Gegensatz zu Fokusgruppen oder Tiefeninterviews – eher zu einem späteren Zeitpunkt der Fragebogenentwicklung angewendet. Der Einsatz ist sowohl im Rahmen von Pilot-Erhebungen als auch während laufender Erhebungen möglich. Obwohl die Debriefing-Interviews üblicherweise von Interviewern und Interviewerinnen durchgeführt werden, ist grundsätzlich auch eine Befragung mittels Selbstausfüller im Anschluss an das Interview denkbar.

Obwohl ein Respondent-Debriefing in unterschiedlicher Intensität, mittels verschiedener Fragetypen oder mit Hilfe verschiedener Befragungs-Modes durchgeführt werden kann, so haben doch alle Arten des Respondent-Debriefings einige Punkte gemeinsam: (1) Das Frageverständnis wird im Zuge

einer „Echtbefragung“ nachgefragt, nicht etwa unter Laborbedingungen oder im Kontext größerer Gruppen (wie etwa Fokusgruppen). (2) Die Probes werden nach dem Interview bzw. im Anschluss an den relevanten Interviewteil gestellt. (3) Die Methode bringt die Befragten in eine neue Rolle, indem man sie um Mithilfe bei der Verbesserung des Befragungsinstruments bittet und somit das Befragungsinstrument zur Diskussion stellt. (4) Die Methode zeigt sowohl den Befragten als auch den Fragenden auf, dass Differenzen beim Frageverständnis immer wieder auftreten und im üblichen Befragungsprozess unerwähnt bleiben *(Martin 2006)*.

Vorgehensweise am Beispiel Mikrozensus-Wohnungserhebung

Im Rahmen der Mikrozensus-Wohnungserhebung wird laufend die Wohnsituation der österreichischen Haushalte erhoben. Ziel sind valide Aussagen über die physische und rechtliche Wohnsituation der Haushalte (wie etwa Wohnungsgröße oder Rechtsverhältnis) sowie die Darstellung von Höhe und Entwicklung der Mietkosten. Dazu werden jedes Quartal etwa 20.000 Haushalte befragt. Die Erstbefragung erfolgt persönlich, die vier Nachfolgebefragungen in der Regel telefonisch. Die erhobenen Wohnkosten fließen u.a. in die Berechnung des nationalen Verbraucherpreisindex (VPI) sowie des harmonisierten Verbraucherpreisindex der Europäischen Union (HVPI) ein und stellen somit eine wichtige Grundlage für die Berechnung der Preisentwicklung dar.

Ausgangslage

Die in der Mikrozensus-Wohnungserhebung gestellten Fragen bestehen in der jetzigen Form – mit geringfügigen Änderungen – schon seit vielen Jahren. Die meisten Wohn-Merkmale liegen somit in längeren Zeitreihen vor. Aufgrund verschiedener Entwicklungen können Änderungen bzw. Anpassungen bei den Fragen dennoch immer wieder notwendig werden. Zunächst sind Entwicklungen am Wohnungsmarkt und in der Wohnungspolitik zu beobachten, wie etwa das vermehrte Angebot an Mietobjekten mit Eigentumsoption, die Bebauung von Baurechtsgründen oder die Wiener Wohnbauoffensive – um nur einige wenige zu nennen. Diese können veränderte Abrechnungsmodalitäten zur Folge haben, mit denen die Haushalte zum Begleichen ihrer Wohnkosten oder Wohnkostenbestandteile konfrontiert sind. Weiters kann sich auch die Verwendung von Begriffen (sowohl der Befragten als auch der am Wohnungsmarkt agierenden Institutionen) aufgrund aktueller Entwicklungen ändern.

Anfang 2013 wurden daher verschiedene Maßnahmen zur Qualitätsverbesserung der Mikrozensus-Wohnungserhebung im Rahmen eines umfassenden Prozesses begonnen. Einen wesentlichen Teil davon macht die Anpassung der verwendeten Erhebungsunterlagen (Fragebogen, Erläuterungen für Erhebungspersonen etc.) aus. Die Frageninhalte sollen zum einen präzisiert und (an aktuelle Entwicklungen) angepasst werden. Zum anderen gilt es, die drei Haushaltsbefragungen Mikrozensus-Wohnungserhebung, EU-SILC und Konsumerhebung in Bezug auf wohnspezifische Inhalte zu harmonisieren (Näheres zur Harmonisierung in *Zucha & Heuberger 2014)*.

Zu beachten ist dabei, dass Änderungen der Erhebungsinstrumente einerseits Brüche in den Zeitreihen verursachen können, andererseits wirken sie sich auch auf andere Projekte und Erhebungen aus (etwa HVPI bzw. VPI, Volkswirtschaftliche Gesamtrechnung, Konsumerhebung, EU-SILC u.a.) – zu den zahlreichen Verwendungen der Mikrozensus-Wohnungserhebung siehe *Zucha 2015* sowie *Zucha & Heuberger 2014*.

Um die notwendigen Änderungen der Erhebungsunterlagen auf eine empirische Basis zu stellen, wurden mehrere Quellen zur Fundierung und Analyse verwendet: Neben den gesetzlichen Grundlagen und weiteren Unterlagen (z.B. Standarddokumentationen) wurden bereits erhobene Datensätze analysiert und Gespräche mit Experten und Expertinnen zu speziellen Problemstellungen geführt. Im Jahr 2013 wurde schließlich ein Respondent-Debriefing in der Mikrozensus-Wohnungserhebung durchgeführt.

Das Respondent-Debriefing wurde in erster Linie zur Prüfung der Validität der Fragen eingesetzt. Insgesamt waren folgende Ziele im Fokus der Entscheidung um den Einsatz dieser Testmethode:

1. Überprüfung des Begriffs- und Frageverständnisses der Befragten, Interpretation der Fragen durch die Befragten und Übereinstimmung mit den Messkonzepten der Erhebung
2. Information über die Antwortgenerierung durch die Respondenten und Respondentinnen: Nutzung von Informationsquellen durch die Befragten und Besitz relevanter Informationen zur Beantwortung der Fragen
3. Generierung von Zusatzinformationen zu einzelnen wohnbezogenen Sachverhalten.

Durchführung

Das Respondent-Debriefing wurde in der Mikrozensus-Wohnungserhebung im dritten Quartal 2013 durchgeführt. Die Zusatzbefragung erfolgte im Anschluss an den verpflichtenden Frageteil des Mikrozensus – mit dem Hinweis auf die Freiwilligkeit der Teilnahme. Als Zielhaushalte wurden jene Haushalte ausgewählt, die die Folgebefragung auf eigenen Wunsch in Form von persönlichen Interviews durchführen lassen (üblicherweise werden die Folgebefragungen telefonisch geführt). Die Interviewer und Interviewerinnen erhielten einen ausführlichen Informationsbrief zum Ziel und Ablauf des Respondent-Debriefings. Etwa 800 Haushalte stimmten der Teilnahme an der Zusatzbefragung zu. Die Befragungsdauer und die Anzahl der gestellten Fragen waren aufgrund der Filterführung der teilnehmenden Haushalte sehr unterschiedlich.

Die zu testenden Fragen bzw. Begriffe wurden zum Teil auf Grundlage der vorangegangenen Schritte – wie Dokumentenanalyse im Zuge der Harmonisierung, Item-Non-Response-Analyse und Gespräche mit Experten und Expertinnen – ausgewählt. Darüber hinaus basierte die Fragenauswahl auch auf Gesprächen mit anderen Abteilungen von Statistik Austria. Die Probes wurden daraufhin formuliert und in mehreren Probeinterviews getestet. Als Fragentypen kamen offene Fragen (v.a. zum Fragen- bzw. Begriffsverständnis), halb-offene und geschlossene Fragen zum Einsatz.

Die offenen Fragen waren nach Abschluss der Feldphase als Texteingaben im Gesamtdatensatz enthalten; die Mehrzahl der offenen Fragen wurde nach einem aufgrund der Antworten generierten Codeschema zugeordnet. Die Analyse des Respondent-Debriefing erfolgte jedoch sowohl qualitativ wie auch quantitativ. Die offenen Fragen wurden einerseits als Text (teilweise mit Zusatzinformationen zu Wohnungsmerkmalen) analysiert, andererseits wurden die vercodeten Antworten als quantitative Verteilungen und mittels Kreuztabellen ausgewertet. Zum jetzigen Zeitpunkt ist die Analyse zwar noch nicht für alle Themenbereiche vollständig abgeschlossen. Dennoch hatte das Respondent-Debriefing bereits konkrete Auswirkungen auf den Fragebogen sowie die Erhebungsunterlagen v.a. des Mikrozensus, aber auch auf die anderen beiden Erhebungen, die im Harmonisierungsprozess eingebunden waren.

Konsequenzen des Tests

Die bisherigen Ergebnisse des Respondent-Debriefing sowie die übrigen Tests, Analysen und Gespräche mit Experten und Expertinnen haben bereits zu einigen Änderungen der Erhebungsunterlagen geführt. Die inputseitigen Adaptionen umfassten nicht nur die Fragetexte, sondern auch die Hilfstexte am Bildschirm, zusätzliche Erläuterungen für Interviewer und Interviewerinnen sowie Fehlermeldungen, die während des Interviews am Bildschirm (etwa bei Eingabe unplausibler Werte) aufscheinen können. Einzelne Änderungen, wie etwa eine geringfügige Präzisierung der Fragen nach dem gesamten Wohnungsaufwand betreffen nur die Mikrozensus-Wohnungserhebung. Weitreichender waren jene Qualitätsanpassungen, die im Zuge der Harmonisierung der wohnspezifischen Fragen von Mikrozensus, EU-SILC und Konsumerhebung erfolgt sind. Diese Adaptierungen umfassten die Themenbereiche Rechtsverhältnis, Wohnungs- sowie Gebäudeausstattung und Wohnungsgröße, nicht jedoch den Bereich der Miet- und Wohnkosten.

Fazit zur Testmethode „Respondent-Debriefing"

Ein Respondent-Debriefing bringt jedenfalls eine intensive Auseinandersetzung mit den formulierten bzw. laufend im Erhebungsprozess eingesetzten Fragen. Die eingehende Beschäftigung mit dem Fragenmaterial begann im Fall der Mikrozensus-Wohnungserhebung bereits lange bevor die Zusatzfragen bzw. Probes formuliert wurden, nämlich bereits im Zuge der Entscheidung über die adäquate Testmethode sowie während der Konzeptualisierung des Respondent-Debriefings. Eine der Grundvoraussetzungen für eine Anwendung dieser Testmethode ist jedenfalls das Wissen um mögliche Schwierigkeiten in den Survey-Fragen. Ohne dieses Vorwissen können keine sinnvollen Nachfragen gestellt werden. Zielführend ist dabei, vorab möglichst eine Zuordnung der potentiellen Schwierigkeiten zu einzelnen Arten von Messfehlern vorzunehmen, sodass die Debriefing-Fragen noch zielgerichteter ausgewählt und formuliert werden können.

Bei der Formulierung der Debriefing-Fragen selbst sind jedenfalls auch die Grundregeln einer guten Frageformulierung anzuwenden *(Költringer 1997; Fowler & Cosenza 2008)*. Der Debriefing-Fragebogen bzw. die Probes sind natürlich auch vorab zu testen und gegebenenfalls anzupassen.

Manchmal kann es vorkommen – und das trifft wohl für die meisten ex-post-Testmethoden zu –, dass nicht alle Erkenntnisse unmittelbar in Fragenänderungen münden. Ein Ergebnis solcher Tests kann auch sein, längere Zeitreihen ohne Eingriff in die Frageformulierung fortzusetzen oder aufgrund anderer Projekte bzw. Erhebungen bisherige Frageformulierungen beizubehalten. Änderungen in laufenden Erhebungen stellen immer auch ein Abwägen zwischen Fragenqualität und der Darstellung der zeitlichen Entwicklung einzelner Merkmale dar.

Das Durchführen eines Respondent-Debriefings bedeutet nicht nur einen finanziellen (Zusatz-)Aufwand, sondern erfordert auch zeitliche Ressourcen und längere Vorlaufzeiten. Diese sind im Rahmen einer Testreihe, in der eventuell geänderte Fragen nochmals getestet werden sollen, entsprechend zu berücksichtigen. Weiters kann auch der Einsatz offener Probes zwar sehr erkenntnisgewinnend, aber gerade in der Phase der Auswertung auch sehr zeitintensiv sein.

Die in der Mikrozensus-Wohnungserhebung bereits durchgeführten Änderungen stützen sich auf die erwähnten Datenquellen. Die Ergebnisse des Respondent-Debriefings zeigten, dass der Einsatz weiterer, ergänzender Testmethoden zu den Wohnfragen sinnvoll wäre. Um die Sichtweise und Erfahrungen der Interviewer und Interviewerinnen besser einbeziehen zu können, wäre etwa ein Interviewer- und Interviewerinnen-Debriefing hilfreich. Qualitative Zusatzinformationen zu Teilergebnissen des Respondent-Debriefings der Mikrozensus-Wohnungserhebung könnten auch mit kognitiven Tests hilfreich ergänzt werden – insbesondere zu den Wohnkosten-Fragen.

Interviewer- und Interviewerinnen-Debriefing bei EU-SILC

Hintergrund und Ziele

Im Zentrum dieser Testmethode stehen die Erfahrungen der Interviewer und Interviewerinnen. Ziel dieser Methode ist es, den gesamten Erhebungsprozess – über das Design des

Fragebogens bis hin zur Bearbeitung der Stichprobe – aus Perspektive der die Erhebung durchführenden Interviewer und Interviewerinnen zu evaluieren. Es wird dadurch ein Bereich der Datenerhebung einsehbar, der für die Projektverantwortlichen üblicherweise nicht direkt zugänglich ist *(Carley-Baxter 2008)*. Die Ergebnisse der Debriefings können Verbesserungspotentiale aufzeigen (z.B. hinsichtlich Frageformulierungen, Plausibilisierungschecks oder in Bezug auf die Feldsteuerung) sowie Anstoß für weitere Forschungsaktivitäten (z.B. zur Evaluierung bestimmter Fragen mittels kognitiver Tests) geben. Die Methode des Debriefings von Interviewern und Interviewerinnen wird innerhalb des europäischen statistischen Systems empfohlen *(Brancato et al. 2006)* und auch als bewährte Testmethode ausgewiesen *(ABS 2001; Henningsson 2002)*.

Hintergrund solcher Debriefings ist, dass Interviewern und Interviewerinnen eine zentrale Rolle im Datenerhebungsprozess einnehmen. Ihr Einsatz kann zwar einerseits mit Fehlerquellen im Datenerhebungsprozess verbunden sein (z.B. durch Abweichungen des Interviewers bzw. der Interviewerin vom vorgegebenen Fragewortlaut oder durch geändertes Antwortverhalten der Befragten aufgrund persönlicher Merkmale des Interviewers bzw. der Interviewerin wie Geschlecht oder Alter), andererseits können sie auch Fehler des Fragebogens ausgleichen, indem sie etwa den Respondenten und Respondentinnen Hilfestellungen und Motivation bei der Beantwortung bestimmter Fragen anbieten. Debriefings von Interviewern und Interviewerinnen geben Raum für das Feedback der Erhebungspersonen und tragen so zur Minimierung von Messfehlern bei.

Vorgehensweise am Beispiel EU-SILC

In Österreich werden für EU-SILC (EU-Statistics on Income and Living Conditions) jährlich rund 6.000 Haushalte entweder persönlich (CAPI – Computer Assisted Personal Interviewing) oder telefonisch (CATI – Computer Assisted Telephone Interviewing) befragt. Die Debriefings werden als standardmäßige Testmethode ebenso jährlich mit den zwölf CATI-Interviewern und -Interviewerinnen, welche für die Dauer der Feldarbeit von ca. vier Monaten für EU-SILC tätig sind, durchgeführt.[9]) Bei den in EU-SILC angewendeten CATI-Debriefings können zwei verschiedene Formen unterschieden werden: Begleitende Maßnahmen während der Feldzeit sowie leitfadengestützte Gruppendiskussionen zu Feldende.[10])

[9]) Da die CAPI-Erhebungspersonen in ganz Österreich aktiv sind, werden in EU-SILC derzeit keine Debriefings mit ihnen abgehalten. Möglich wäre jedoch ein Debriefing in Form von (teilstrukturierten) Fragebögen, die schriftlich (auch per E-Mail oder webbasiert) beantwortet werden können *(Brancato et al. 2006)*.

[10]) Standardisierte Fragebögen wären eine weitere Form des Interviewer- und Interviewerinnen-Debriefings *(Brancato et al. 2006)*; sie werden in EU-SILC nicht verwendet. Darüber hinaus wird angemerkt, dass Debriefings grundsätzlich zu jedem Zeitpunkt eines Erhebungsprozesses eingesetzt werden können, allerdings ist die Zielsetzung jeweils eine andere.

Begleitende Maßnahmen während der Feldzeit

Um Raum für das Feedback der CATI-Interviewer und -Interviewerinnen bereits während der Feldarbeitszeit zu schaffen, werden in EU-SILC folgende Debriefing-Maßnahmen gesetzt:

- Regelmäßige Teammeetings: In diesen zunächst wöchentlich, nach der Einarbeitungszeit 14-tägig durchgeführten Teammeetings erhalten die Interviewer und Interviewerinnen die Möglichkeit, sich zu verschiedenen Bereichen der Feldarbeit – darunter auch zum Fragebogen – zu äußern. Vorab werden sie dazu angehalten, ihre Beobachtungen zu notieren, und zwar möglichst direkt nach den Befragungen. Damit soll eine weitgehend vollständige und unverzerrte Berichterstattung erreicht werden. Unterstützend wird dazu bei Bedarf auch ein Dokumentationsleitfaden mit allgemeinen Hilfestellungen an die Interviewer und Interviewerinnen ausgegeben.

- Schriftliche Dokumentation von Besonderheiten: Bei Bedarf, vor allem aber zur Evaluierung neu eingeführter Fragen, werden die Interviewer und Interviewerinnen gebeten, Besonderheiten, die im Zusammenhang mit bestimmten Fragen auftreten, schriftlich zu dokumentieren. Rückmeldungen und Reaktionen der Respondenten und Respondentinnen sollen dabei ebenso wie Probleme (z.B. in der Zuordnung von Antworten) oder Verbesserungsvorschläge von Seiten der Interviewer und Interviewerinnen festgehalten werden. Im Gegensatz zur zuvor dargestellten Methode, die den mündlichen Austausch zwischen Team- und Fachgruppenmitgliedern fokussiert, liegt hier der Schwerpunkt auf der schriftlichen Dokumentation, die darauf abzielt, auch scheinbar Nebensächliches festzuhalten.

- Kollektive Sammlung von Fragen und Antworten: Alle Fragen und Unklarheiten, die im Zusammenhang mit dem Fragebogen genannt werden (entweder in Einzelgesprächen oder in den Teammeetings), werden in einem für alle Interviewer und Interviewerinnen zugänglichen Dokument gesammelt. Nach Prüfung der Anmerkungen durch die fachliche Expertengruppe wird für jede Frage eine Antwort bereitgestellt. Diese Sammlung von Fragen und Antworten dient während der Feldarbeitszeit zur Information der Interviewer und Interviewerinnen und wird in der Vorbereitung der nächsten Erhebungswelle als Ausgangspunkt für Adaptierungen des Fragebogens herangezogen.

Leitfadengestützte Gruppendiskussionen zu Feldende

Die begleitenden Maßnahmen während der Feldzeit werden zu Feldende durch die ausführlichen Debriefings in Form von leitfadengestützten Gruppendiskussionen abgerundet. Diese Form des Debriefings von Interviewern und Interviewerinnen ist die am weitesten verbreitete *(Brancato et al. 2006)* und wird für EU-SILC in Gruppen von etwa jeweils sechs Personen an zwei Terminen abgehalten. Im Vorfeld wird – wie von *Brancato et al.* (2006) empfohlen – ein Leitfaden

erstellt, der für EU-SILC folgende Themenbereiche umfasst:

- Fragebogen (allgemein und bestimmte Schwerpunkte, z.B. Fragen des jährlich wechselnden Moduls, neu eingeführte oder geänderte Fragen, Fragen mit wiederkehrenden Auffälligkeiten oder Problemen)
- Infrastruktur und technische Aspekte der Fragebogen-Navigation
- Schulung und Betreuung/Supervision
- Feldsteuerung und Bearbeitung der Stichprobe

Der Output der Diskussion wird durch den Leitfaden bereits im Vorhinein strukturiert, wobei die Diskussion auch neuen Input zulassen sollte. Die Interviewer und Interviewerinnen werden daher zumeist aufgefordert, ihre Gedanken zum Thema spontan zu äußern und erst danach werden die Details durch gezieltes Nachfragen abgehandelt *(Brancato et al. 2006)*. Eine wichtige Rolle nimmt dabei die Moderation ein, die Raum für neue Themen geben, Abweichungen vom Thema erkennen, alle teilnehmenden Personen miteinbeziehen und an richtigen Stellen gezielte Nachfragen stellen sollte *(Carley-Baxter 2008)*. Von einer weiteren teilnehmenden Person der Fachgruppe wird die Diskussion schriftlich protokolliert.

Konsequenzen des Tests

Nach Abschluss der Feldphase und Durchführung der Debriefings von Interviewern und Interviewerinnen werden alle zur Verfügung stehenden Dokumente (Protokolle der Gruppendiskussionen, Dokumentationen von Besonderheiten, Sammlung von Fragen und Antworten) einer inhaltlichen Prüfung unterzogen. Diese können sich auf die verschiedensten Bereiche der Feldarbeit (Fragebogen, Infrastruktur, Supervision, Feldsteuerung etc.) beziehen und werden in der Planung und Durchführung der nächsten Feldphase so weit wie möglich berücksichtigt.

Bei der Auswertung der Rückmeldungen zum Fragebogen, die im Fokus dieses Beitrags stehen, sind mehrere Aspekte zu beachten, wie beispielsweise: Sind vom genannten Problem auch andere Erhebungen von Statistik Austria betroffen? Kann das Problem mit den bereits vorhandenen Richtlinien (z.B. durch die Beschreibung der Zielvariablen von Eurostat) gelöst werden oder sind weitere Recherchen und Rückfragen zur Konkretisierung des zu messenden Konzepts notwendig? Welche Elemente des Fragebogens (z.B. Fragetext, Erläuterungen, Anweisungen für Interviewer und Interviewerinnen) sind betroffen?

Die anschließende Recherche in vorhanden Richtlinien und Vorgaben sowie die Befragung von Experten und Expertinnen kann in unterschiedlichen Konsequenzen für den Fragebogen münden. Einige Beispiele, die Folge von Debriefings in EU-SILC sind, werden nachfolgend dargestellt:

- Standardisierte Evaluationsbefragung zu ausländischen Bildungsabschlüssen: Debriefings von Interviewern und Interviewerinnen brachten Probleme im Zusammenhang mit der Vercodung von im Ausland erworbenen Bildungsabschlüssen zur Sprache. Interne Diskussionen mit Experten und Expertinnen des Bildungsbereichs sowie anderer Erhebungen konnten zwar erste Hilfsmittel zur Verfügung stellen, es wurde jedoch in kurzer Zeit deutlich, dass dieses Problem ein separates Testprojekt verlangt. Eine standardisierte Evaluationsbefragung wurde schließlich auch realisiert. Durch die Rückmeldung der Interviewer und Interviewerinnen wurde die Bearbeitung dieses zwar zuvor bekannten, aber noch nicht im Detail bearbeiteten Problems vorangetrieben.

- Kognitive Tests zur Leistbarkeit von Ausstattungsmerkmalen und Teilhabe (darunter die Merkmale der Deprivation; *siehe Statistik Austria 2015)*: Die Rückmeldungen im Rahmen der Debriefings von Interviewern und Interviewerinnen waren die Ausgangsbasis für eine erneute Review der Eurostat-Richtlinien und einen Vergleich mit dem SILC-Fragebogen aus Deutschland. Es stellte sich heraus, dass bestimmte Aspekte der Fragen zu praktischen Problemen führen können. Da Auswirkungen auf die Ergebnisse bei einer Änderung der Fragen anzunehmen sind, sollten Änderungen nur unter der Kenntnis erfolgen, was mit welcher Methode gemessen wird und ob alternative Fragestellungen den Anforderungen gemäß den Richtlinien für EU-SILC entsprechen. Demnach sind kognitive Tests, die nunmehr angestrebt werden, das einzige Mittel, dies herauszufinden.

- Input für den Musterfragebogen zur Inanspruchnahme medizinischer Leistungen: Debriefings von Interviewern und Interviewerinnen aus mehreren Jahren brachten Verständnisschwierigkeiten mit diesen Fragen zur Sprache; aufgrund der gemeinsamen europäischen Richtlinien konnten diese allerdings nicht vollständig behoben werden. Seit 2015 sind jedoch überarbeitete Fragen im Einsatz, wobei dieser Frageadaptierung eine längere Überarbeitungsphase von Seiten Eurostats vorangegangen war. Im Zuge der Überarbeitung des Fragedesigns wurden die teilnehmenden Länder aufgefordert, Testergebnisse bereitzustellen. Die Debriefings dieses Jahres mit Schwerpunkt auf die Inanspruchnahme medizinischer Leistungen waren damit ein wertvoller Beitrag zur Diskussion um den Musterfragebogen.

- „Finetuning" des Fragebogens: Das Feedback der Interviewer und Interviewerinnen führt vielfach zu einer Verbesserung einzelner Frageelemente. Erläuterungen zu einzelnen Fragen, die für die Interviewer und Interviewerinnen direkt bei der jeweiligen Frage als Hilfetext abrufbar sind, können auf Basis der Rückmeldungen häufig erweitert oder konkretisiert werden. Plausibilisierungschecks können ebenso von Änderungen betroffen sein, indem etwa neue Prüfungen eingeführt werden oder das Wording von Plausibilisierungstexten verbessert wird. Eine weitere Konsequenz des Feedbacks kann die Adaptierung der Anweisungen für Interviewer und Interviewerinnen (ebenfalls direkt bei der Frage ersichtlich) sein.

Fazit zur Testmethode „Interviewer- und Interviewerinnen-Debriefing"

Die vorangegangenen Beispiele zeigen, dass die Ergebnisse aus den Debriefings von Interviewern und Interviewerinnen vielfach Ausgangspunkt für Adaptierungen des Fragebogens sind. Sie leisten damit oftmals einen entscheidenden Beitrag zur Verbesserung des Erhebungsinstruments. Neben dem Input zum Fragebogen ist auch jener zu anderen Bereichen der Feldarbeit zu betonen. Schulungsmaßnahmen können ebenso evaluiert werden wie technische Aspekte der Fragebogengestaltung oder Aspekte der Feldsteuerung und der Bearbeitung der Stichprobe. Debriefings sind zudem Kommunikationskanäle, die einen intensiven Austausch zwischen den Interviewern und Interviewerinnen und den Projektverantwortlichen ermöglichen und fördern.

Bei der Interpretation des Outputs aus den Debriefings von Interviewern und Interviewerinnen ist jedoch zu beachten, dass es sich um subjektive Informationen handelt. Die Rückmeldung spiegelt möglicherweise eher die Sichtweise des Interviewers bzw. der Interviewerin als jene der Befragten wider. Zudem kann aus den Rückmeldungen teilweise nicht eruiert werden, ob die berichteten Schwierigkeiten auf nur wenige oder viele Befragte zutreffen *(Brancato et al. 2006)*. Hinsichtlich der inhaltlichen Ausrichtung des Outputs ist zu beachten, dass sich dieser in gewissem Ausmaß nach der Auswahl der Themen, die sich im Leitfaden finden, richtet. Bei entsprechend offener Moderation sollte dieses Problem aber nachrangig sein. Für die Situation von EU-SILC ist letztlich noch darauf Rücksicht zu nehmen, dass zurzeit ausführliche Debriefings nur mit den CATI-Interviewern und -Interviewerinnen abgewickelt werden. Die Perspektive der CAPI-Erhebungspersonen ist damit zum aktuellen Zeitpunkt nur eingeschränkt vorhanden.

Standardisierte Evaluationsbefragung

Hintergrund und Ziele

Evaluationen werden eingesetzt, um soziale Interventionsprogramme (z.B. Maßnahmen zur Reduktion von Armut oder Ausgrenzung) auf ihre Funktionalität und Wirksamkeit hin zu überprüfen *(Bortz & Döring 2002)*. Bevölkerungsumfragen sind keine sozialen Interventionsprogramme, das Konzept der Evaluation kann aber trotzdem auf das Testen von Fragen übertragen werden. Denn das Ziel ist jeweils dasselbe: die Bewertung des untersuchten Gegenstands mit Hilfe empirischer Forschungsmethoden. Im Folgenden wird eine standardisierte Evaluationsbefragung zur Bestimmung der Datenqualität einzelner Erhebungsmerkmale vorgestellt.

Das Design der standardisierten Evaluationsbefragung ähnelt der Paralleltest-Methode, wie sie bei Reliabilitäts- und Validitätstests eingesetzt wird. Paralleltests bestehen aus zwei Fragebögen, die das gleiche Konstrukt mit vergleichbaren Items erfassen. Beide Fragebögen werden zum selben Befragungszeitpunkt an einer Stichprobe erhoben und die individuellen Werte im Anschluss miteinander korreliert *(Rammstedt 2004)*. Bei der standardisierten Evaluationsbefragung wird ebenfalls ein zweiter Fragebogen eingesetzt. Dieser versucht jedoch ein einzelnes Erhebungsmerkmal nicht nur vergleichbar, sondern detaillierter und somit näher an der Realität abzubilden, um die Zuordnung jedes Wertes im ersten Fragebogen überprüfen zu können. Die Befragungen werden außerdem zu verschiedenen Zeitpunkten durchgeführt.

Die hier präsentierte Evaluationsbefragung ist ein standardisiertes Erhebungsinstrument. Im Unterschied zu qualitativen (insbesondere kognitiven) Testverfahren wird nicht nachgefragt, wie einzelne Antworten zustande kommen, sondern es wird überprüft, wie gut sich Respondenten und Respondentinnen in vorgegebene Antwortkategorien einordnen können. Das standardisierte Design ermöglicht die Evaluation auf Basis einer großen Stichprobe und die Berechnung des Anteils nicht korrekt zugeordneter Werte. Sie eignet sich deshalb besonders für Tests von Erhebungsmerkmalen, deren Zielgruppen in mehrere Subgruppen (z.B. unterschieden nach Herkunftsland) unterteilt sind. Außerdem können bereits spezifische Frageformulierungen getestet werden, falls die Evaluationsergebnisse eine Adaption der untersuchten Frage nahelegen.

Die Testmethode der standardisierten Evaluationsbefragung verfolgt zwei Ziele. Sie will (1) die Qualität der Zuordnung zu einzelnen Antwortkategorien einer Frage überprüfen und bewerten und (2) Hinweise finden, durch welche Adaptionen und Hilfsmittel die Datenqualität der untersuchten Frage in Zukunft verbessert werden kann.

Vorgehensweise am Beispiel Mikrozensus-Arbeitskräfteerhebung

In der österreichischen Mikrozensus-Arbeitskräfteerhebung (und allen dahingehend harmonisierten Erhebungen wie beispielsweise EU-SILC) wird Bildung anhand der erworbenen Qualifikationen innerhalb des formalen Bildungssystems erhoben. Die Respondenten und Respondentinnen werden gebeten, ihre höchste erfolgreich abgeschlossene Schulbildung zu nennen[11]) und den vorgegebenen Antwortmöglichkeiten zuzuordnen. Die Antwortkategorien repräsentieren das österreichische Bildungssystem. Auf Basis der Mikrozensus-Arbeitskräfteerhebung ist davon auszugehen, dass im Jahr 2014 ca. 13% der österreichischen Bevölkerung ab 15 Jahren den höchsten Bildungsabschluss im Ausland[12]) erworben haben. Auch sie werden gebeten, ihren höchsten Bildungsabschluss einer Kategorie des österreichischen Bildungssys-

[11]) „Was ist Ihre höchste erfolgreich abgeschlossene Schulbildung? Bitte ordnen Sie sich selbst einer der folgenden Antwortmöglichkeiten zu." (Pflichtschule / Lehre mit Berufsschule / Fach- oder Handelsschule / Matura / Abschluss an einer Universität / (Fach-)Hochschule / Anderer Abschluss nach Matura). Es folgen weiterführende Fragen zum jeweiligen Bildungsabschluss sowie zu zusätzlichen Ausbildungen.

[12]) Als ausländische Bildungsabschlüsse gelten in diesem Zusammenhang alle Abschlüsse, die Personen vor ihrer Zuwanderung nach Österreich oder im Jahr der Zuwanderung erlangt haben.

tems zuzuordnen. Abschlüsse aus dem Ausland lassen sich aber nicht problemlos vergleichen und in österreichische Kategorien überführen, da sich die einzelnen Bildungssysteme in ihrer Struktur und in ihren Inhalten zum Teil deutlich unterscheiden *(Europäische Kommission 2014)*. Die Folgen sind Unsicherheiten in der Befragungssituation und ein erhöhtes Risiko fehlender oder nicht korrekt erfasster Werte.

Interviewer und Interviewerinnen haben unter anderem im Rahmen von Debriefings vermehrt auf die Problematik der Zuweisung ausländischer Bildungsabschüsse zu österreichischen Kategorien hingewiesen. Um festzustellen, wie hoch der Anteil der nicht adäquat zugeordneten ausländischen Bildungsabschlüsse in der Mikrozensus-Arbeitskräfteerhebung tatsächlich ist und welche Maßnahmen zur langfristigen Verbesserung der Datenqualität möglich sind, wurde ein Erhebungsinstrument für eine standardisierte Evaluationsbefragung entwickelt.

Zielgruppe

Die Zielgruppe der standardisierten Evaluationsbefragung umfasst alle Personen mit ausländischem Bildungsabschluss. Da in der Mikrozensus-Arbeitskräfteerhebung nicht abgefragt wird, in welchem Land der Bildungsabschluss erworben wurde, werden alle Personen um Teilnahme gebeten, die ihre höchste Ausbildung vor ihrer Zuwanderung nach Österreich oder im Jahr ihrer Zuwanderung abgeschlossen haben. Die standardisierte Evaluationsbefragung wird für die Dauer von zwei Monaten an alle Folgebefragungen (2. bis 5. Welle) angehängt.[13]) In diesem Zeitraum umfasst die Zielgruppe ca. 800 Personen. Die Teilnahme ist freiwillig, es wird aber von einer Ausschöpfung von ca. 80% ausgegangen. Um die Antwortqualität zu erhöhen, werden – wie auch im Standardprogramm des Mikrozensus – neben dem deutschen Fragebogen Übersetzungen auf Türkisch, Bosnisch/Kroatisch/Serbisch und Englisch eingesetzt. Es werden nur Selbstauskünfte in die standardisierte Evaluationsbefragung einbezogen.

Durchführung

Die Erfassung der höchsten erfolgreich abgeschlossenen Schulbildung nach österreichischen Kategorien erfolgt in der Mikrozensus-Arbeitskräfteerhebung in der Regel in der persönlichen Erstbefragung. In den telefonischen Folgebefragungen in den darauffolgenden vier Quartalen werden bei diesem Merkmal nur noch etwaige Veränderungen erhoben. Kernelement der standardisierten Evaluationsbefragung ist der Vergleich zweier Fragen bzw. Fragebögen, die dasselbe Konzept auf unterschiedliche Weise erfassen. Es wurde deshalb ein Erhebungsinstrument mit sechs Fragen entwickelt, die in Kombination eine detaillierte Beschreibung des jeweiligen Abschlusses ergeben. Indem diese Fragen im Anschluss an die Folgebefragungen eines Quartals gestellt wurden, können sie – entsprechend der Testmethode – nachher mit den Angaben der Erstbefragung verglichen werden. Der Vergleich gibt Aufschluss über die Qualität der Antworten der Erstbefragung.

Das Erhebungsinstrument besteht aus zwei offenen und vier geschlossenen Fragen, die unter Mithilfe von Bildungsexperten und -expertinnen von Statistik Austria formuliert wurden. Sie orientieren sich an den ergänzenden Dimensionen der Internationalen Standardklassifikation im Bildungswesen (ISCED)[14]) zur Spezifizierung der einzelnen Qualifikationsniveaus *(OECD 1999; UNESCO-UIS 2012)*.
Folgende Merkmale werden erfasst:

- Land des höchsten Bildungsabschlusses (geschlossene Frage)
- Kumulative Beschulungsdauer (geschlossene Frage): Sie erlaubt eine grobe Einstufung in ein niedriges, mittleres oder hohes Bildungsniveau und wird als Kontrolle eingesetzt, falls die weiteren Angaben der Befragten unplausibel oder unvollständig sind.
- Art der Schule, Hochschule oder Institution, in welcher der höchste Bildungsabschluss erlangt wurde (offene Frage)
- Art des Zeugnisses, Zertifikats oder Diploms, welches als Bestätigung für den höchsten Bildungsabschluss ausgestellt wurde (offene Frage): Sowohl der Schultyp als auch das Zeugnis, Zertifikat oder Diplom sollen von den Befragten in ihrer Landessprache und wenn möglich auch auf Deutsch genannt werden. Diese Klartexteinträge stellen den Kern der standardisierten Evaluationsbefragung dar.
- Ausbildung in einer Schule und in einem Betrieb (geschlossene Frage): Eine Besonderheit des österreichischen Bildungssystems ist die Lehre als duale Berufsausbildung. In manchen Ländern kann ein Lehrabschluss auch ohne Besuch einer Berufsschule erlangt werden, was jedoch nicht der Antwortkategorie „Lehre mit Berufsschule“ entspricht.
- Berechtigung, an einer Universität oder Hochschule zu studieren (geschlossene Frage): Mit dieser letzten Frage können Bildungsabschlüsse auf Maturaniveau identifiziert werden. Sie dient ebenfalls als Kontrolle, falls die Klartexteinträge unzureichend sind.

Auswertung

Die Auswertung der standardisierten Evaluationsbefragung erfolgt in zwei Schritten:

Durch die Kombination offener und geschlossener Fragen sowie durch die Erhebung zentraler Merkmale in der Landessprache können die einzelnen im Ausland erworbenen Bildungsabschlüsse zuerst auf Einzelfallebene bestimmt und

[13]) Die Evaluation wurde im 3. Quartal 2015 durchgeführt. Als dieser Artikel verfasst wurde, war sie noch nicht abgeschlossen, weshalb keine Ergebnisse präsentiert werden können.

[14]) Die International Standard Classification of Education (ISCED) ist ein vom UNSECO-Institut für Statistik erstelltes Klassifikationsschema, das zur Übersetzung nationaler Bildungsqualifikationen in international vergleichbare Kategorien dient.

einem österreichischen Qualifikationsniveau zugeordnet werden. Die Zuordnung erfolgt mit Hilfe einer Korrespondenztabelle, die ausländische Bildungsabschlüsse ihren österreichischen Äquivalenten gegenüberstellt. Sie wurde auf Basis einer umfangreichen Literaturrecherche erstellt und umfasst sowohl die aktuellen als auch bereits ausgelaufene Bildungsprogramme der wichtigsten Zuwanderungsländer des österreichischen Mikrozensus.

Anschließend wird diese Zuordnung mit den Antworten der Respondenten und Respondentinnen auf die Frage nach ihrer höchsten erfolgreich abgeschlossenen Schulbildung aus der Erstbefragung verglichen. Auf aggregierter Ebene kann nun berechnet werden, wieviele Abschlüsse nicht korrekt zugewiesen wurden und ob der ausländische Bildungsabschluss eher über- oder unterschätzt wird.

Mögliche Konsequenzen des Tests

Sollten sich die Angaben aus der Erstbefragung deutlich von den Angaben aus der Evaluationsbefragung unterscheiden, weist dies auf Schwächen bei der Erfassung ausländischer Bildungsabschlüsse in der österreichischen Mikrozensus-Arbeitskräfteerhebung (und allen dahingehend harmonisierten Erhebungen) hin. Die Ursachen dafür können in verschiedenen Bereichen der Erhebung liegen.[15]) Die angewendete Testmethode kann aber Hinweise auf mögliche Maßnahmen liefern, um die Befragungssituation zu vereinfachen und die Datenqualität langfristig zu verbessern: (1) Erstellung detaillierter Erläuterungen zu den einzelnen österreichischen Bildungsabschlüssen, (2) Erstellung von Korrespondenztabellen oder Hilfetexten zu den Bildungssystemen der häufigsten Zuwanderungsländer oder (3) Einführung zusätzlicher Fragen, die im Zuge der Datenaufbereitung als Kontrolle herangezogen werden können.

Bei der Anwendung der Testmethode der standardisierten Evaluationsbefragung auf ausländische Bildungsabschlüsse müssen auch kritische Aspekte beachtet werden. Die Respondenten und Respondentinnen werden gebeten, die beiden offenen Fragen sowohl in ihrer Landessprache als auch auf Deutsch zu beantworten. Da die Interviewer und Interviewerinnen diese Sprache in den meisten Fällen nicht sprechen, wird die Bezeichnung des Schultyps und des Zeugnisses, Zertifikats oder Diploms buchstabiert. Dabei müssen die Interviewer und Interviewerinnen darauf achten, dass die Antworten möglichst detailliert und eindeutig sind. Die Bearbeitung dieser Fragen dauert dadurch länger und beansprucht die Aufmerksamkeit der Respondenten und Respondentinnen stärker, wodurch das Risiko steigt, unzureichende oder fehlerbehaftete Antworten zu erhalten. Ein weiterer Aspekt bezieht sich auf die Auswertung: Sie muss teilweise auf Einzelfallbasis erfolgen, was einen hohen Aufwand an zeitlichen Ressourcen bedeutet.

[15]) Um die genaue Ursache festzustellen empfiehlt sich die Durchführung zusätzlicher Tests.

Fazit zur Testmethode „Standardisierte Evaluationsbefragung"

Standardisierte Evaluationsbefragungen haben zum Ziel, die Zuordnung von Antworten zu einzelnen Antwortkategorien einer Frage zu überprüfen und zu bewerten sowie Hinweise zu erlangen, durch welche Adaptionen und Hilfsmittel die Datenqualität der untersuchten Frage in Zukunft verbessert werden kann. Im Zentrum des Tests stehen zwei Fragen bzw. Fragebögen, die dasselbe Merkmal auf unterschiedliche Weise erfassen. Durch das standardisierte Design können die Antworten der beiden Fragebögen verglichen und der Anteil der nicht korrekt zugeordneten Werte berechnet werden. Standardisierte Evaluationsbefragungen eignen sich für große Stichproben und lassen sich leicht in bestehende Erhebungen implementieren – Eigenschaften, die vor allem für Zielgruppen, die aus mehreren Subgruppen bestehen, wichtig sind. Darüber hinaus können bereits spezifische Formulierungen (z.B. für zukünftige Zusatzfragen) getestet werden, falls die Ergebnisse eine Veränderung der untersuchten Frage nahelegen.

Das methodische Design unterliegt jedoch auch Einschränkungen. So muss das Erhebungsinstrument der standardisierten Evaluationsbefragung selbst genau getestet werden, um sicherzugehen, dass bei beiden Fragebögen dasselbe Merkmal erfasst wird. Die Ergebnisse der Evaluation weisen außerdem nur darauf hin, dass beim untersuchten Erhebungsmerkmal Probleme bestehen. Wie diese im Einzelfall zustande kommen, muss durch weitere (z.B. kognitive) Tests geklärt werden. Letztlich muss berücksichtigt werden, dass die Forschenden entscheiden, ob die Respondenten und Respondentinnen ihre Antworten den korrekten Antwortkategorien zugeordnet haben, was eine hohe inhaltliche Kompetenz im Hinblick auf das untersuchte Merkmal voraussetzt.

Schlussbemerkungen

Ein guter Fragebogen ist die Basis für hohe Datenqualität. Aus diesem Grund ist es notwendig, jeden Fragebogen systematisch zu testen. Denn nur durch Fragebogentests lässt sich vollständig beurteilen, wie gut ein Fragebogen wirklich ist. Die Anwendung von Fragebogentestmethoden sollte daher fixer Bestandteil im Prozess der Fragebogenerstellung sein.

Im vorliegenden Beitrag wurden vier Testmethoden vorgestellt, die bei den Erhebungen Mikrozensus und EU-SILC als Mittel zur kontinuierlichen Verbesserung des Fragebogens Anwendung finden. Es konnte gezeigt werden, dass durch die Verwendung der Testmethoden Verbesserungspotentiale zutage gebracht werden können, die ansonsten verborgen bleiben würden. Es zeigte sich aber auch, dass keine der verwendeten Testmethoden für sich allein ein vollständiges Bild über die Qualität des Fragebogens liefern kann. Vielmehr legt jede Testmethode einen eigenen Schwerpunkt in der Qualitätsbetrachtung und unterliegt in der Durchführung jeweils unterschiedlichen Vor- und Nachteilen.

Die Stärke kognitiver Interviews liegt in der Detailgenauigkeit, durch die ein Einblick in den Frage-Antwort-Prozess möglich wird. Zudem zeichnen sich kognitive Interviews durch ihre große Offenheit gegenüber noch unbekannten Aspekten des Fragebogens aus. Sie eignen sich daher besonders in einer frühen Phase der Fragebogenentwicklung, um den Fragebogenentwurf hinsichtlich Validität sowie Messfehler im Zusammenhang mit den Respondenten und Respondentinnen und ihren Informationssystemen zu bewerten.

Dagegen sind die anderen drei vorgestellten Testmethoden eher in späteren Phasen der Fragebogenentwicklung angesiedelt. Sie können nach der ersten „finalen" Fragebogenversion in Pilot-Studien oder bei der Post-Evaluation eingesetzt werden, was besonders bei Paneldesigns sinnvoll ist. Respondent-Debriefings sind empfehlenswert, wenn Informationen über eine überschaubare Anzahl gut bekannter Schwierigkeiten mit einzelnen Fragen in Erfahrung gebracht werden sollen. Durch das standardisierte Nachfragen haben die Forschenden die Möglichkeit zu steuern, auf welchen potentiellen Fehlerquellen (Konzept-, Mess- oder Non-Response-Fehler) der Fokus liegen soll. Debriefings von Interviewern und Interviewerinnen sind deutlich offener; ihre Stärke liegt darin, die Facetten der Feldrealität in Erfahrung zu bringen. Diese Testmethode eignet sich daher dafür, den Erhebungsablauf insgesamt zu bewerten und einzelne, noch unbekannte Details des Fragebogens im Zusammenhang mit Konzept- und Messfehlern oder Non-Response ans Licht zu bringen. Standardisierte Evaluationsbefragungen ermöglichen es, die Validität der Messung eines bestimmten Konzepts zu quantifizieren. Diese Testmethode ist anwendbar, wenn für ein Konzept eine detailliertere Messung mit alternativen Fragen vorstellbar ist und davon ausgegangen werden kann, dass sie das zu messende Phänomen genauer abbilden als die bereits bestehenden Fragen im Fragebogen. Sie eignet sich aber auch dazu, unterschiedliche Frageformulierungen gegeneinander abzuwägen.

Für zukünftige Arbeiten im Bereich der Fragebogenerstellung und -verbesserung bleibt festzuhalten, dass die Anwendung einer einzelnen Testmethode wohl selten ausreicht, um ein vollständiges Bild über die Qualität eines Fragebogens zu erhalten. Zur Bildung eines Gesamturteils über den Fragebogen bedarf es im Optimalfall eines iterativen Testdesigns, bei dem unterschiedliche Methoden miteinander kombiniert werden. Durch den Methodenmix ist es darüber hinaus möglich, Nachteile einer Testmethode durch die Vorteile einer anderen abzufedern.

Darüber hinaus erscheint die stärkere Institutionalisierung von Fragebogentests – sowohl innerhalb als auch zwischen Organisationen – als wichtiges Ziel für die Zukunft. Die verschiedenen Testmethoden sowie deren Kombination erfordern entsprechendes Know-how, welches organisatorisch verankert werden sollte. Die durchgeführten Tests zeigten auch, dass es einer intensiven und systematischen Weiterverarbeitung der Ergebnisse bedarf, damit sie zu den erforderlichen Adaptionen im Fragebogen führen können. Nicht zuletzt wurde deutlich, dass aus den Testergebnissen auch allgemeine Befunde ableitbar sind, die durchaus auf andere Erhebungen anwendbar sind.

Mit einer verstärkten Institutionalisierung von Fragebogentests in der amtlichen Statistik kann das Testen von Fragebögen als Querschnittsthema im Rahmen der Qualitätssicherung institutionell verankert werden und die Fragebogenentwicklung noch stärker auf einem evidenzbasierten Fundament aufbauen.

Literatur

ABS (2001): "Pretesting in survey development. An Australian Bureau of Statistics perspective". Research Paper. Canberra: Australian Bureau of Statistics.

Blair, J. / Conrad, F.G. (2011): "Sample Size for Cognitive Interview Pretesting". Public Opinion Quarterly, 75(4), S. 636–658.

Bortz, J. / Döring, N. (2002): „Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler". Berlin: Springer.

Brancato, G. et al. (2006): "Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System. European Commission Grant Agreement 2004103000002".

Carley-Baxter, L. (2008): "Interviewer debriefing". In P. Lavrakas (Hrsg.): Encyclopedia of survey research methods (S. 369-370). Thousand Oaks: SAGE Publications.

Europäische Kommission (2014): „Struktur der europäischen Bildungssysteme 2014/15: Schematische Diagramme". http://eacea.ec.europa.eu > Eurydice > More about Eurydice > Publications.

Fasching, M. (2015): „Kognitive Tests zur Erfassung des Erwerbsstatus". Statistische Nachrichten, 9/2015, S. 693-701.

Fowler, F.J. / Cosenza, C. (2008): "Writing effective questions". In: E. De Leeuw / J.J. Hox / D.A. Dillman (Hrsg.): International Handbook of Survey Methodology (S. 136-160). New York: Lawrence Erlbaum.

Froschauer, U. / Lueger, M. (2003): „Das qualitative Interview. Zur Praxis interpretativer Analyse sozialer Systeme". Wien: WUV.

Henningsson, B. (2002): "Interviewer Debriefing by E-Mail". Statistics Sweden.

Költringer, R. (1997): „Richtig fragen heißt besser messen. Optimale Formulierungstechniken für Umfragen". Mannheim: FRG.

Kytir, J. / Stadler, B. (2004): „Die kontinuierliche Arbeitskräfteerhebung im Rahmen des neuen Mikrozensus. Vom „alten" zum „neuen" Mikrozensus". Statistische Nachrichten, 6/2014, S. 511-518.

Martin, E. (2006): "Vignettes and Respondent Debriefings for Questionnaire Design and Evaluation. Research Report Series (Survey Methodology #2006-8)". www.census.gov/srd/papers/pdf/rsm2006-08.pdf.

Mayring, P. (2015): „Qualitative Inhaltsanalyse. Grundlagen und Techniken". Weinheim, Basel: Beltz.

Meertens, V. (2015): "Pre-testing LFS model questionnaires." 10th Workshop on Labor Force Survey Methodology. Prag.

OECD (1999): "Classifying Educational Programmes. Manual für ISCED-97. Implementation in OECD Countries". 1999 Edition.

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

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

Schnell, R. et al. (2008): „Methoden der empirischen Sozialforschung". München: Oldenbourg.

Statistik Austria (2015): „Lebensbedingungen in Österreich – ein Blick auf Erwachsene, Kinder und Jugendliche sowie (Mehrfach-)Ausgrenzungsgefährdete". Wien.

Statistik Austria (2014): „Cognitive Testing of Proposed LFS-Questions Measuring the Employment Status". Unter Mitarbeit von Marc Plate und Melitta Fasching. Wien. (zuletzt geprüft am 08.07.2015)

Statistik Austria (2012): „Qualitätsrichtlinien der Statistik Austria. Version 1.2. Stand 31.12.2012". Wien.

Tourangeau, R., et al. (2012): "The psychology of survey response". Cambridge u.a: Cambridge Univ. Press.

UNESCO-UIS (2012): "International Standard Classification of Education: ISCED 2011". Montreal: UNESCO Institute for Statistics.

Willis, G. (1999): "Cognitive Interviewing. A "How to" Guide". Research Triangle Institute.

Willis, G. (2005): "Cognitive Interviewing. A tool for improving questionnaire design". Sage.

Zucha, V. / Heuberger, R. (2014): "Toward harmonisation of survey questions on housing". Präsentiert bei der European Conference on Quality in Official Statistics, Wien. http://www.q2014.at > Papers & Presentations > Session 19 (2.-5. Juni 2014).

Zucha, V. (2015): "Mixing Methods for Quality Assessment and Harmonisation of Survey Questions". Präsentiert bei der 6th Conference of the European Survey Research Association (ESRA), Reykjavik. (13.-17. Juli 2015).

Summary

Questionnaires are a key part in data collection processes. In order to ensure high data quality, they need to undergo comprehensive tests and evaluations. This paper shows a selection of testing methods conducted by Statistics Austria as part of the projects Microcensus and EU-SILC. For each method, a general description as well as detailed information on implementation and outcome is provided. Further implications are discussed.

The aim of this article is to show the importance of testing and evaluating survey instruments. Overall there should be a greater emphasis assigned to this topic by providing an institutional foundation for testing questionnaires as part of the overall quality management as well as by better integrating tests in existing workflows.