Oskar Mittag[1], **Sonja Böhmer**[1], **Ruth Deck**[1], **Monika Ekkernkamp**[1], **Angelika Hüppe**[1], **Hildegard Telbis-Kankainen**[2], **Angelica Raspe**[1], **Heiner Raspe**[1]

[1] *Institut für Sozialmedizin, Lübeck*

[2] *Therapiezentrum St. Jürgen, Lübeck*

Originalartikel I Original article

Fragen über Fragen: *cognitive survey* in der Fragebogenentwicklung[1]

Summary

Questions about questions: the use of cognitive methods in the development of survey questions

Objectives: Traditional pretests are used in the development of survey items to identify technical and comprehension problems. Cognitive processes involved in answering survey questions are not the object of this kind of test.

Methods: Cognitive survey methods were used here to test a questionnaire screening for rehabilitation needs in people suffering from back pain. Essential techniques of cognitive testing (think-aloud, probing, confidence ratings) are outlined. We applied these techniques to 20 patients suffering from either acute or chronic back pain in order to test the survey.

Results: The main goal, i. e., identifying problems in item formulation by means of cognitive testing, was achieved. Almost one third of the survey questions were rephrased according to the results of the study. Some of the improvements of the questionnaire are illustrated.

Conclusions: The increased effort required to perform cognitive testing as compared to traditional pretesting pays off. The two methods have specific pros and cons and cannot replace one another.

Keywords: Cognitive survey – Questionnaire – Assessment – Evaluation – Back pain.

Gesundheitssurveys und epidemiologische Studien werden häufig mittels Fragebögen durchgeführt, in denen beispielsweise Vorerkrankungen, momentane Gesundheitsstörungen, Gesundheitsverhalten oder auch die Inanspruchnahme von Gesundheitsleistungen abgefragt werden. Aufgrund solcher Daten werden dann Massnahmen im Gesundheitswesen geplant und implementiert. Da dies mit erheblichen Konsequenzen verbunden sein kann (z.B. Kosten, Einschränkungen, Behandlungen), werden an die Verlässlichkeit der Daten hohe Massstäbe angelegt. Neben dem Problem der Stichprobenauswahl muss dabei insbesondere auch der Qualität der Erhebungsinstrumente Aufmerksamkeit gelten.

Heute liegt eine ganze Reihe von mehr oder weniger empirisch begründeten Regeln für die „richtige" Konstruktion von Fragebogenitems vor (z.B. Schnell et al. 1993). Damit ist jedoch keineswegs gewährleistet, dass entsprechend gestaltete Instrumente tatsächlich valide und reliabel sind. Die Evaluation „nach Augenschein" vom Schreibtisch aus erlaubt noch keinen Rückschluss auf die Praxistauglichkeit eines Fragebogens. Aus diesem Grund werden in aller Regel Pretests durchgeführt, um die Güte und Praktikabilität eines Fragebogens zu überprüfen und den Feldzugang zu testen. Der konventionelle Pretest beinhaltet die Vorgabe eines Fragebogens unter möglichst realistischen (Haupt-)Studienbedingungen, um technische Probleme (z.B. Handhabung, unausgefüllte Seiten bei zweiseitigem Druck), Verständlichkeit (z.B. fehlende Werte, widersprüchliche Angaben) und Akzeptanz (z.B. Rücklaufzeiten und -quoten) oder Antworttendenzen (z.B. Boden- oder Deckeneffekte) zu sondieren.

Ausser acht gelassen werden dabei die kognitiven Prozesse, die bei der Beantwortung von Fragebogenitems eine Rolle spielen. Weder wissen wir, was genau sich der Proband bei der Beantwortung einer Frage vorstellt (so kann der Begriff „Rücken" bei verschiedenen Befragten durchaus unterschiedliche anatomische Vorstellungen wachrufen), noch können wir beurteilen, welche Mechanismen der Befragte nutzt, um Informationen aus dem Gedächtnis abzurufen

[1] Aus dem Projekt " Die Abschätzung von Rehabedarf bei aktiven Mitgliedern der Gesetzlichen Rentenversicherung: Der Lübecker Algorithmus und seine Validierung" im NVRF (Projektleiter: Prof. Dr. Dr. H. Raspe; Förderer: BMBF / VDR; FKZ: 02 1 06).
Die Autoren danken ausdrücklich den beiden Gutachtern für ihre wertvollen Anregungen, die in die Überarbeitung des Manuskripts eingegangen sind.

Soz.- Präventivmed. 48 (2003) 55–64
0303-8408/03/010055–10

(ob er z.B. die Arztbesuche in den vergangenen 12 Monaten zu erinnern versucht und zählt, oder ob er lediglich einen Schätzwert nennt). Methoden des *cognitive survey* sollen näheren Aufschluss über diese Prozesse geben. *Cognitive survey* und konventioneller Pretest wollen also etwas anderes; beide Testverfahren sollten sich gegenseitig ergänzen.

Wir wollen hier über einige Grundlagen dieser Untersuchungsmethoden sowie die Geschichte der neueren kognitiven Surveymethoden berichten. Im Rahmen einer Studie setzen wir dann einige der Techniken des *cognitive survey* ein, um einen Fragebogen zur Erfassung des Bedarfs an medizinischen Rehabilitationsmassnahmen in Hinblick auf die Tauglichkeit der Items zu überprüfen. Wir berichten die wesentlichen Ergebnisse und versuchen, daraus Möglichkeiten und Grenzen dieser Methode abzuleiten und Empfehlungen für ihren weiteren Einsatz zu geben. Dabei werden wir u.a. zu dem Schluss kommen, dass sich buchstäblich die im Titel apostrophierten „Fragen über Fragen" stellen, wenn man beginnt, die Items eines Fragebogens infrage zu stellen.

Denkpsychologische Aspekte

Gedächtnis wird heute nicht mehr als einheitliches Repräsentationssystem betrachtet, sondern es werden mehrere unterschiedliche Gedächtnissysteme angenommen. Die Differenzierung in zeitabhängige Gedächtnissysteme (Ultrakurzzeit-, Kurzzeit- und Langzeitgedächtnis) wird damit um eine inhaltliche Dimension erweitert. Folgende allgemeine Teilsysteme werden unterschieden. Das *episodische Gedächtnis* bezieht sich auf persönliche (autobiographische) Erinnerungen an erlebte Ereignisse oder Szenen. Es ist kontextgebunden, d.h. zeitlich und örtlich spezifisch. Das *semantische Gedächtnis* ist dagegen kontextfrei und bezieht sich auf generalisiertes Wissen, das im Lebenslauf erlernt wurde. Das *prozedurale Gedächtnis* beinhaltet Routinen, die bei erlernten Aktivitäten (z.B. Fahrradfahren) erforderlich sind (vgl. Sudman et al. 1996; Welzer & Markowitsch 2001). Für unsere Fragestellung sind vor allem das episodische Gedächtnissystem und eventuell auch das Wissenssystem (semantisches Gedächtnis) von Bedeutung.

Menon (1994) beschreibt ein Modell der Speicherung und des Abrufs von Informationen bei unterschiedlichen Ereignis- bzw. Verhaltenstypen. Sie unterscheidet dabei regelmässiges vs. unregelmässiges Auftreten des Verhaltens einerseits und ähnliche vs. unähnliche Verhaltensweisen andererseits. Danach lassen sich vier unterschiedliche Ereignistypen bilden: regelmässig-ähnliches Verhalten (z.B. zur Arbeit gehen, Krebsfrüherkennung beim Gynäkologen, Rauchen), regelmässig-unähnliches Verhalten (z.B. Wochenendgestaltung); unregelmässig-ähnliches Verhalten (z.B. Hausarztbesuch, krankgeschrieben sein und zuhause bleiben) sowie unregelmässig-unähnliches Verhalten (z.B. Gäste haben, Krankheiten oder Verletzungen). Unähnliche Ereignisse werden eher dem episodischen Speicher zugeordnet, ähnliche eher dem semantischen. Regelmässig-ähnliches Verhalten wird anhand der Ereignisrate geschätzt. Bei unregelmässig-unähnlichem Verhalten muss die Person dagegen zunächst einzelne Ereignisse erinnern und dann zählen. Die beiden übrigen Verhaltenstypen erfordern Mischstrategien bei der Reproduktion der Information. Sudman, Bradburn und Schwarz (1996) konnten dies bestätigen; ihre Daten zeigen auch, dass Probanden das Schätzen von Ereignisraten deutlich leichter fällt als der Abruf einzelner Ereignisse aus dem episodischen Gedächtnis.

Auch Zeitaspekte spielen eine wichtige Rolle bei der Erinnerung von (autobiographischen) Gedächtnisinhalten. Die Vergessenskurve verläuft umso flacher, je wichtiger die entsprechende Information für die Person ist; sinnlose Silben, wie sie in den klassischen Experimenten von Ebbinghaus (1894; Wiederabdruck 1964) verwandt wurden, werden innerhalb weniger Tage vergessen, während die Namen von Mitschülern oder wichtige Lebensereignisse über Jahre behalten werden (s. Sudman, et al. 1996). Ein weiterer Zeiteffekt wurde erstmals von Neter und Waksberg (1964; nach Sudman et al. 1996) beschrieben. Bei zeitlich proximalen und kürzeren Intervallen (z.B. letzter Monat) werden mehr Ereignisse berichtet als bei längeren und distalen Zeiträumen (z.B. letzte 12 Monate); dieser Effekt wird als „telescoping" bezeichnet.

Geschichte der kognitiven Surveymethoden

Versuche, mehr über die kognitiven Prozesse bei der Beantwortung von Interviewfragen zu erfahren, wurden bereits recht früh unternommen. Cantril (1944) untersuchte die subjektive Bedeutung, die der Einschätzung, ob die Menschen nach dem Krieg schwerer arbeiten würden als vorher, unterlegt wird, indem er die Probanden anschliessend fragte, an welche spezielle Gruppe (z.B. Arbeiter, Unternehmer) sie bei ihrer Antwort gedacht hatten. Nuckols (1953) forderte Probanden auf, die gestellten Fragen mit eigenen Worten zu wiederholen. Mit der gleichen Methode konnte Ferber (1956) zeigen, dass Probanden Meinungsfragen beantworten, deren Inhalt sie gar nicht verstehen. Dies führte zu dem Begriff der „nonattitudes" für entsprechend zufällig zustande gekommene Antworten durch Converse (1970). Schuman (1966) überprüfte das Frageverständnis,

indem er bei zufällig ausgewählten Items Zusatzfragen zum genaueren Verständnis stellte ("random probes").
Kognitionspsychologen und Demoskopen haben seit Ende der 1970er Jahre begonnen, gemeinsam eine Reihe von Methoden (neu) zu entwickeln, die Informationen über die kognitiven Prozesse bei der Beantwortung von Fragebogen- oder Interviewfragen erlauben. In den USA fand 1980 eine erste Tagung zu Fragen kognitiver Surveymethoden statt, die seitens des Bureau of Social Science Research (BSSR) organisiert wurde; kurz zuvor hatte es in England bereits ein ähnliches Treffen gegeben. Als eigentlicher Beginn der kognitiven Bewegung in der empirischen Sozialforschung gilt aber das multidisziplinäre Advanced Research Seminar on Cognitive Aspects of Survey Methodology, das 1983 in St. Michael's, Maryland, stattfand und unter dem Akronym CASM bekannt geworden ist (vgl. Sirken et al. 1999; Sudman et al. 1996; Tanur 1991). In Deutschland wurden entsprechende Verfahren vor allem vom Zentrum für Umfragen, Methoden und Analysen (ZUMA) weiterentwickelt (z.B. Kurz et al. 1999; Prüfer & Rexroth 1996).

Kognitive Surveymethoden

Standardmethoden des *cognitive survey* sind die *think-aloud*-Technik, das *probing* zum Frageverständnis sowie die Überprüfung der Verlässlichkeit einer Antwort mittels *confidence rating*.

Think-aloud

Diese Technik gilt als die zentrale kognitive Technik (Prüfer & Rexroth 1996). Die Probanden werden aufgefordert, „laut zu denken" und dabei alle Gedanken, die zur Antwort führen bzw. geführt haben, zu formulieren. Unterschieden wird dabei, ob dies unmittelbar während der Befragung bzw. der Beantwortung des Fragebogens erfolgt (*concurrent think-aloud*), oder ob die Befragten im Nachhinein aufgefordert werden zu berichten, wie sie zu der jeweiligen Antwort gekommen sind (*retrospective think-aloud*).
Prüfer und Rexroth (1996) weisen darauf hin, dass es für die Anwendung dieser Methode keine verbindlichen Vorgaben etwa für die Anzahl der durchzuführenden Interviews, die Schulung der Interviewer oder für das praktische Vorgehen gibt. Insbesondere die Methode des *concurrent think-aloud* stellt hohe Anforderungen an die Probanden, die nur unter detaillierter und fortlaufender Anleitung überhaupt erfüllt werden können (Prüfer & Rexroth 1996; Sudman et al. 1996).
Die Technik des *think-aloud* steht in der Forschungstradition der von Max Wertheimer und Wolfgang Köhler begründeten Berliner Schule der Gestaltpsychologie. In seinen Forschungsarbeiten über das produktive Denken forderte Karl Duncker Anfang der 1930er Jahre seine Probanden auf, laut zu denken, während sie vorgegebene Probleme (z.B. die bekannte Bestrahlungsaufgabe) zu lösen versuchten (Duncker 1935). Aufgrund der so erstellten Lautlösungsprotokolle wurden dann typische Problemlöseprozesse analysiert.

Probing

Dabei wird die gegebene Antwort vom Interviewer durch eine oder mehrere Zusatzfragen hinterfragt, um Information zum Frageverständnis oder zur Art der Informationsgewinnung zu erhalten. Das *probing* kann ebenfalls sofort im Anschluss an die spontane Fragebeantwortung erfolgen oder auch retrospektiv nach dem Interview bzw. dem Ausfüllen des Fragebogens. Typische Fragen zum Verständnis der Frage wären: „Könnten Sie mir ein Beispiel geben für das, was Sie meinen?"; „Könnten Sie das noch näher erläutern?"; „Wie meinen Sie das." *Probing* zu Aspekten der Informationsbeschaffung wären Fragen wie: „Wie schwer fiel es Ihnen, diese Frage zu beantworten?"; „Haben Sie das geschätzt oder sind Sie in Gedanken alle Termine durchgegangen?"; „Wie sind Sie darauf gekommen?"
Es fällt schwer, die Techniken des *think-aloud* und des *probing* gegeneinander abzugrenzen. Die erstere Technik scheint in der „reinen" Form zwar zunächst deutlich nondirektiver als das *probing*, durch die Notwendigkeit begleitender verbaler Anleitungen ist dieser Unterschied jedoch zumindest infrage gestellt. Beim retrospektiven *think-aloud* verwischen sich die Grenzen zum *probing* beinahe vollständig. Willis, DeMaio und Harris-Kojetin (1999) sprechen daher auch von einem Kontinuum, das von eher allgemeinen Techniken (*think-aloud*) bis hin zu sehr spezifischen Methoden (*targeted probes*) reicht. Schechter, Blair und Vande Hey (1996) schlagen vor, beide Techniken unter dem Oberbegriff der *cognitive interviewing techniques* zusammenzufassen.

Confidence ratings

Hier geht es darum, den Grad der Verlässlichkeit einer Antwort auf Faktfragen einzuschätzen (subjektive Sicherheit). Dies geschieht zumeist mit Hilfe einer vorgegebenen Skala ("Was würden Sie sagen: Ist Ihre Angabe sehr genau, ziemlich genau, eher ungenau oder grob geschätzt?").
Als Vorteile dieser kognitiven Surveytechniken gelten schnelle Durchführbarkeit, niedrige Kosten und die Möglichkeit, Fragebögen in unterschiedlichen Entwicklungsstadien zu überprüfen; Nachteile sind die Beschränkung auf einzelne Items sowie das hohe Unsicherheitsrisiko bezüglich der Generalisierbarkeit der Ergebnisse aufgrund der

zwangsläufig geringen Fallzahl (Prüfer & Rexroth 1996). Angesichts des vergleichsweise hohen personellen und Arbeitsaufwands, der mit einer qualitativen Testung verbunden ist, muss das Zeit- und Kostenargument für die kognitiven Methoden aber wohl relativiert werden. Aktuelle Übersichten über die Methoden des *cognitive survey* geben die Werke von Sirken et al. (1999), Schwarz und Sudman (1996) sowie Sudman, Bradburn und Schwarz (1996), die zur vertiefenden Lektüre empfohlen werden.

Beispiele für den Einsatz kognitiver Surveymethoden

Groves, Fultz und Martin (1991) überprüften mögliche Gendereffekte bei der Beantwortung einer Frage nach dem allgemeinen Gesundheitszustand, indem sie die Frage anschlossen: „Als Sie die Frage beantwortet haben, woran haben Sie da gedacht?". Sie fanden u. a., dass bei Probanden, die an die letzten Jahre dachten, keine Geschlechtsunterschiede auftraten, während bei denjenigen, die sich auf proximale Zeiträume bezogen, die Frauen eine deutlich schlechtere Gesundheit angaben als Männer. Loftus et al. (1991) überprüften in mehreren Experimenten, wodurch die Erinnerung an die Zahl der Arztbesuche innerhalb der letzten 12 Monate beeinflusst wird; sie fanden, dass Probanden, die Arztbesuche konkret mit Datum benennen mussten, die wahre Anzahl um über 50% unterschätzen, während globale Schätzungen zu wesentlich realistischeren Angaben führten. Willis, Roystone und Bercini (1991) berichten über die Ergebnisse eines kognitiven Survey und geben anschauliche Beispiele dafür, wie einzelne Items eines Gesundheitssurveys aufgrund der Ergebnisse umformuliert wurden. So wurden Fragen umformuliert, das Format von Antwortkategorien wurde verändert, und einzelne Items wurden ganz ausgeschieden. Jobe und Mingay (1990) überprüften einen Funktionsfragebogen bei 18 älteren Probanden. Sie setzten dabei die Technik des *think-aloud* sowie gezielte Nachfragen (*probes*) ein. Weitere Beispiele für Pretestprotokolle, die verschiedene kognitive Surveytechniken beinhalten, finden sich bei Kurz, Prüfer und Rexroth (1999) sowie bei Means et al. (1989).

Stichprobe und Methoden

Wir setzten einige der oben genannten Methoden ein, um die Items eines Screeningbogens für Rehabedarf bei aktiven Mitgliedern der GRV mit Dorsopathien in Hinblick auf semantischen Gehalt und Verlässlichkeit näher zu untersuchen. Die 20 Probanden rekrutierten sich als Gelegenheitsstichprobe aus drei Gruppen: zwei „Rückengesunde" aus dem Umfeld des Instituts, 11 ambulante Rehapatienten aus einem ambulanten Therapiezentrum[2] ("akute Rückenschmerzen") sowie sieben chronische Schmerzpatienten aus einer orthopädischen Praxis ("chronische Rückenschmerzen"). 14 der Probanden sind weiblich (70%); das Durchschnittsalter betrug 45 Jahre (SD = 8,29; Bereich = 31–60 Jahre). Bei den Schulabschlüssen überwiegen Haupt- und Realschule (acht bzw. sechs Probanden), gefolgt vom Abitur (drei Probanden); zwei Probanden gaben andere Schulabschlüsse an, ein *missing value*. sechs Probanden gaben als (letzte) Berufsstellung „Arbeiter" an, neun gaben „Angestellter" an, je ein Proband war selbständig bzw. gab „Sonstiges" an, ein *missing value*. Acht Probanden waren aktuell erwerbstätig, die übrigen befanden sich in Ausbildung (1), waren arbeitslos (3), berentet (1), Hausfrauen (5) oder kreuzten „Sonstiges" an (2).

Der 15-seitige Fragebogen wurde unter standardisierten Bedingungen entweder im Institut oder im Therapiezentrum St. Jürgen vorgegeben. Die Probanden wurden nach erfolgter Aufklärung über die Studie zunächst gebeten, den Fragebogen auszufüllen und dabei alles zu sagen, was ihnen durch den Kopf geht. Die Instruktion für dieses *think-aloud* lautete: "Wir wollen wissen, wie gut unser Fragebogen verständlich ist und welche Probleme möglicherweise beim Ausfüllen entstehen. Deshalb bitten wir Sie, beim Ausfüllen des Fragebogens **möglichst alles zu sagen, was Ihnen durch den Kopf geht**. Sie sollen also beim Ausfüllen „**laut denken**". Natürlich können Sie auch nachfragen, wenn irgend etwas unklar ist."

Alle Äusserungen wurden handschriftlich bei dem jeweils entsprechenden Item in einem zweiten Exemplar des Fragebogens protokolliert (Lautlösungsprotokolle). Wenn ein Proband eine ganze Seite ausgefüllt hatte, ohne einen Kommentar zu geben, wurde er zu Beginn der nächsten Seite noch einmal an die Instruktion erinnert.

Anschliessend wurden die Probanden nach einem standardisierten Protokoll zu einzelnen Items des Fragebogens befragt. Die Auswahl der Items richtete sich einerseits danach, zu welchen Items wir zusätzliche Information erhofften bzw. welche Items wir als potentiell kritisch in Hinsicht auf das Verständnis ansahen, andererseits wollten wir möglichst verschiedene kognitive Techniken anwenden, um deren Praktikabilität zu überprüfen. Es wurden Techniken des *probing* sowie auch *confidence ratings* angewandt. Dabei wurden die Probanden auch gebeten, die von ihnen angegebenen aktuellen Rückenschmerzen in einem vorgegebenen Rückenmannequin ohne Schraffur des gemeinten Bereiches zu lokalisieren.

[1] Wir danken dem Geschäftsführer des Therapiezentrums St. Jürgen (Lübeck), Herrn J.-U. Schmidt, sowie den Mitarbeiterinnen und Mitarbeitern für ihre freundliche Unterstützung bei der Durchführung der Untersuchung.

Die Untersuchung dauerte zwischen 60 und 90 Minuten. In einem Fall musste sie abgebrochen werden, da die Probandin nach 90 Minuten erst knapp die Hälfte des Fragebogens ausgefüllt hatte. Die Untersuchung wurde von drei Mitarbeitern des Instituts durchgeführt.

Ergebnisse

Die Anwendung der verschiedenen Methoden des *cognitive survey* auf die Fragebogenitems und wesentliche Ergebnisse werden im Folgenden einzeln dargestellt und anhand von Beispielen illustriert.

Think-aloud

In einem Item des Fragebogens wird nach Besuchen bei verschiedenen Fachärzten gefragt. Hier wurde die vorgesehene Kategorie „nein" von vielen Probanden ignoriert; angekreuzt wurden meist nur diejenigen Ärzte, die auch tatsächlich innerhalb der letzten 12 Monate aufgesucht wurden. Die Auswertung der Fragebögen bestätigt dies. Der Anteil der „fehlenden Werte" liegt ausser für den Orthopäden zwischen 30 und 60%; bei dem Orthopäden hat lediglich ein Proband (5%) keine der vorgegebenen Kategorien angekreuzt. Das Format der Antwortkategorien wurde daraufhin so geändert, dass lediglich der „positive" Arztbesuch angegeben werden muss (Abb. 1).

Bei demselben Item trat zusätzlich das Problem auf, dass mindestens eine Probandin eine Routineuntersuchung beim Gynäkologen angab, die nicht mit Rückenschmerzen in Zusammenhang stand. Mindestens ein weiterer Proband gab einen Arztbesuch an, der ebenfalls nicht mit Rückenschmerzen in Verbindung stand. Die Instruktion wurde daraufhin noch einmal überarbeitet, Abbildung 1 zeigt das neue neben dem alten Itemformat.

In Item 17 wird danach gefragt, ob der Rücken (Wirbelsäule) in den letzten 12 Monaten geröntgt wurde; drei Probanden fragten hier spontan, ob sie auch ein CT bzw. ein MRT angeben sollten. Da es bei der Frage nicht um das Röntgen, sondern um die Differentialdiagnostik mittels bildgebender Verfahren geht, wurden diese beiden Untersuchungsmethoden in Klammern hinzugesetzt.

Item 19 fragt nach der Einnahme von Schmerz- oder Rheumamitteln. Fünf Probanden betonen hier spontan ihre kritische Haltung zu Medikamenten und geben „nie" an, obwohl sie zumindest zeitweise Medikamente genommen hatten. Aus den begleitenden Kommentaren ist ausserdem ersichtlich, dass ein Proband die Einnahme von Aspirin wegen Kopfschmerzen angibt; das anschliessende *probing* zeigt, dass ausserdem zwei weitere Probanden ebenfalls Aspirin benannt haben, allerdings gleichzeitig auch Schmerz- oder Rheumamittel i.e.S. Drei Probanden fragen spontan danach, welcher Zeitraum gemeint ist bzw. beziehen die Frage

16. Welche der folgenden Ärzte haben Sie **in den letzten 12 Monaten** wegen oder im Zusammenhang mit Ihren **Rückenschmerzen** aufgesucht und wie häufig? Hier geht es **nicht** um Arztkontakte im Rahmen von Klinikaufenthalten und Rehamassnahmen. *(Sie können mehrere Kästchen ankreuzen.)*

	nein	**ja, einmal**	**mehrfach**
Praktischer Arzt / Arzt für Allgemeinmedizin	☐	☐	☐
Internist	☐	☐	☐
Frauenarzt (Gynäkologe)	☐	☐	☐
Orthopäde	☐	☐	☐
Neurologe/Psychiater	☐	☐	☐
Chirurg	☐	☐	☐

sonstige, und zwar: _____

16. Welche der folgenden Ärzte haben Sie **in den letzten 12 Monaten wegen oder im Zusammenhang mit Ihren Rückenschmerzen** aufgesucht und wie häufig? Hier geht es **nicht** um Arztkontakte im Rahmen von Klinikaufenthalten und Rehamassnahmen. *(Sie können mehrere Kästchen ankreuzen.)*

	einmal (wegen Rückenschmerzen!)	**mehrfach** (wegen Rückenschmerzen!)
Praktischer Arzt / Arzt für Allgemeinmedizin	☐	☐
Internist	☐	☐
Frauenarzt (Gynäkologe)	☐	☐
Orthopäde	☐	☐
Neurologe/Psychiater	☐	☐
Chirurg	☐	☐

sonstige (wegen Rückenschmerzen), und zwar: _____

Abbildung 1 Das Item „Arztkontakte" vor dem *cognitive survey* (oben) und nach erfolgter Neuformulierung (unten)

auf die letzten 12 Monate. Die Frageformulierung wurde daraufhin geändert: (1) die Reihenfolge „Schmerz- oder Rheumamittel“ wurde vertauscht, um den Akzent auf die Antirheumatika zu legen, (2) es wurde der Zusatz „…wegen Ihrer Rückenschmerzen“ hinzugefügt, und (3) es wurde ein Zeitfenster von 12 Monaten eingeführt. Abbildung 2 zeigt die alte und die neue Itemformulierung.

Wir hatten in dem Fragebogen zum Pretest aus Gründen der Einheitlichkeit bei der Skala „Katastrophisieren“ (PRSS) von Flor und Turk (1992) die Ziffern „1“ bis „6“ in den Antwortkästchen weggelassen. Neun Probanden äusserten spontan, ihnen falle es schwer, sich zwischen den Antwortkategorien zu orientieren, eine weitere Probandin meinte: „Da denke ich mir „1” bis „5” (sic!)…“. Wir haben daraufhin die Ziffern wieder eingesetzt.

Interessanterweise erhielten wir aus den Lautlösungsprotokollen auch Hinweise auf Probleme mit einzelnen Items aus langjährig eingeführten und psychometrisch gut überprüften Instrumenten. So reagierten einige Probanden auf das Item: „Können Sie sich im Bett aus der Rückenlage aufsetzen?“, aus dem FFbH-R (Kohlmann & Raspe 1996) spontan mit Kommentaren wie: „Das tue ich natürlich nicht – das ist doch ungesund!“ Hier scheinen 10 Jahre konsequenter Rückenschule einen an dieser Stelle unerwarteten Erfolg zu zeigen, der bei der Fragebogenkonstruktion noch nicht abzusehen war. Veränderungen in entsprechend eingeführten, standardisierten Instrumenten wurden hier aber nicht angestrebt.

Zunächst wurden aufgrund der Ergebnisse des *cognitive survey* 20 Items identifiziert, bei denen Änderungsbedarf vermutet wurde. Nach eingehenden Diskussionen innerhalb der Forschungs- und Autorengruppe wurden dann 12 Items tatsächlich geändert. Das sind 27% der Items des Fragebogens bzw. 29%, sofern man die standardisierten Instrumente ausnimmt, bei denen Änderungen von vornherein nicht angestrebt wurden. In sieben Fällen handelte es sich um inhaltliche bzw. Formulierungsänderungen, in sechs Fällen um formale Änderungen der Antwortkategorien oder des Formats und in einem Fall wurde das ganze Item entfernt, da sich herausstellte, dass es bei der Zielgruppe in keinem Fall zutraf (“Gründe für Nichtbesuch eines Arztes im 12-Monatszeitraum“). Die vorstehenden Zahlen beinhalten z.T. inhaltliche und formale Änderungen bei ein und demselben Item; mehrfache Änderungen innerhalb einer Kategorie wurden dagegen nur einmal gezählt.

Probing

Bei einigen Items des Fragebogens wurde der Bedeutungshorizont mittels *probing* ausgelotet. Ein Item erfragt, welches unter einer Reihe zuvor abgefragter Gesundheitsprobleme „derzeit“ das Hauptproblem sei. Im *cognitive survey* wurden die Probanden nun gebeten zu erläutern, welchen Zeitraum sie mit der Formulierung „derzeit“ verbinden. Dabei wurde noch einmal unterschieden, wie weit „derzeit“ in die Vergangenheit reicht, und ob darunter auch ein Zeitraum bis in die Zukunft verstanden wird. Es zeigte sich, dass die Probanden mit der Frage, was „derzeit“ ihr Hauptproblem sei (Item 15), Zeiträume verbinden, die zwischen „einigen Tagen“ (25%) bis zu „einigen Jahren“ (30%) in die Vergangenheit und zwischen „überhaupt nicht“ (20%) und „einige Jahre“ (25%) auch in die Zukunft reichen.

Der Abstand zwischen den Antwortkategorien „mehrmals“ und „fast immer“ in einem Item zu weiteren Rückenschmerzen in den vergangenen 12 Monaten wurde von einigen Probanden im *think-aloud* als zu gross empfunden; betrachtet man die Antworten der Probanden auf die Probingfrage, was für sie „einmal“, „mehrmals“ und „fast immer“ bedeute (s. Tab. 1), so wird aber deutlich, dass die vorgegebenen

19. Wie häufig nehmen Sie **Schmerz- oder Rheumamittel?**
praktisch jeden Tag . . . . . ☐
mehrmals in der Woche (an bis zu 5 Tagen) . . . . . ☐
mehrmals im Monat (an bis zu 10 Tagen) . . . . . ☐
mehrmals im Jahr (an bis zu 20 Tagen) . . . . . ☐
nie . . . . . ☐

19. Wie häufig haben Sie **in den letzten 12 Monaten Rheuma- oder Schmerzmittel** genommen? *(Nur wegen Ihrer Rückenschmerzen!)*
praktisch jeden Tag . . . . . ☐
mehrmals in der Woche (an bis zu 5 Tagen) . . . . . ☐
mehrmals im Monat (an bis zu 10 Tagen) . . . . . ☐
mehrmals im Jahr (an bis zu 20 Tagen) . . . . . ☐
nie . . . . . ☐

Abbildung 2 Das Item „Schmerzmitteleinnahme” vor dem *cognitive survey* (oben) und nach erfolgter Neuformulierung (unten)

Tabelle 1 Ergebnisse des *probing* beim Item "Hatten Sie (ausserdem) in den letzten 12 Monaten Rückenschmerzen?"

Antwort-kategorie	„Was bedeutet für Sie …"
einmal	„Eine Episode vor 12 Wochen"
mehrmals	„Ein- bis zweimal im Monat" „Eine Woche lang" „An manchen Tagen - mehrmals" „Einmal kommt nicht hin" „Mehrmals, paarmal in den letzten 12 Monaten, bei der Arbeit, ein bis zwei Wochen im Jahr" „Vor einem Jahr, dann wieder vor sechs Wochen" „Weiss ich nicht" „Wöchentlich"
fast immer	„Beim Spazierengehen zum Beispiel" „Immer" „Immer" „In Ruhe ist alles OK; Rückenschmerzen in Bewegung" „Jeden Tag ausser am Wochenende" „Keinen Tag und keine Nacht ohne Schmerzen" „Tag und Nacht" „Täglich, pausenlos, ausser nachts"

Antwortkategorien recht genau zwischen Probanden ohne Rückenschmerzen im vorgesehenen Zeitfenster, mit einmaligen Ereignissen, mit rezidivierenden Rückenschmerzen und dem Gefühl, immer unter Rückenschmerzen zu leiden, unterscheiden. Genau dies wurde mit dem Item intendiert und von einer Änderung wurde deswegen abgesehen.

Ausserdem erlaubt das *probing*, die Korrektheit des Frageverständnisses zu überprüfen; zwei Beispiele seien genannt: Die Lokalisation der aktuellen Rückenschmerzen auf dem „nackten" Rückenmannequin erfolgt in allen Fällen innerhalb des „richtigen" Bereiches; allerdings markieren etliche Probanden darüber hinaus noch weitere Stellen, die ausserhalb des im Fragebogen schraffiert vorgegebenen Bereiches liegen (vgl. Abb. 3). Teilweise handelt es sich dabei um ausstrahlende Schmerzen, die aller Wahrscheinlichkeit nach nicht vom Rücken sondern der Halswirbelsäule kommen. Man kann wohl davon ausgehen, dass die Probanden nicht zwischen diesen Schmerzen und den „eigentlichen" Rückenschmerzen unterscheiden.

Die 12 Probanden, die eine Ausstrahlung der Rückenschmerzen bis zum Knie bzw. bis unterhalb des Knies angegeben hatten ("Ischiasschmerz"), wurden aufgefordert, das exakte Ausstrahlungsgebiet anzugeben; 10 von ihnen gaben dabei typische Beschwerden an, zwei eher Beschwerden mit Schmerzen an der Vorderseite des Oberschenkels. Überwiegend wurde das Item also richtig verstanden und beantwortet.

Schliesslich bildet das *probing* ein gutes Verfahren zur Hypothesengenerierung; ein Beispiel: Im Fragebogen wird die „allgemeine Gesundheit" auf einer fünfstufigen Skala von „sehr gut" bis „schlecht" eingestuft. Auf die Probingfrage:

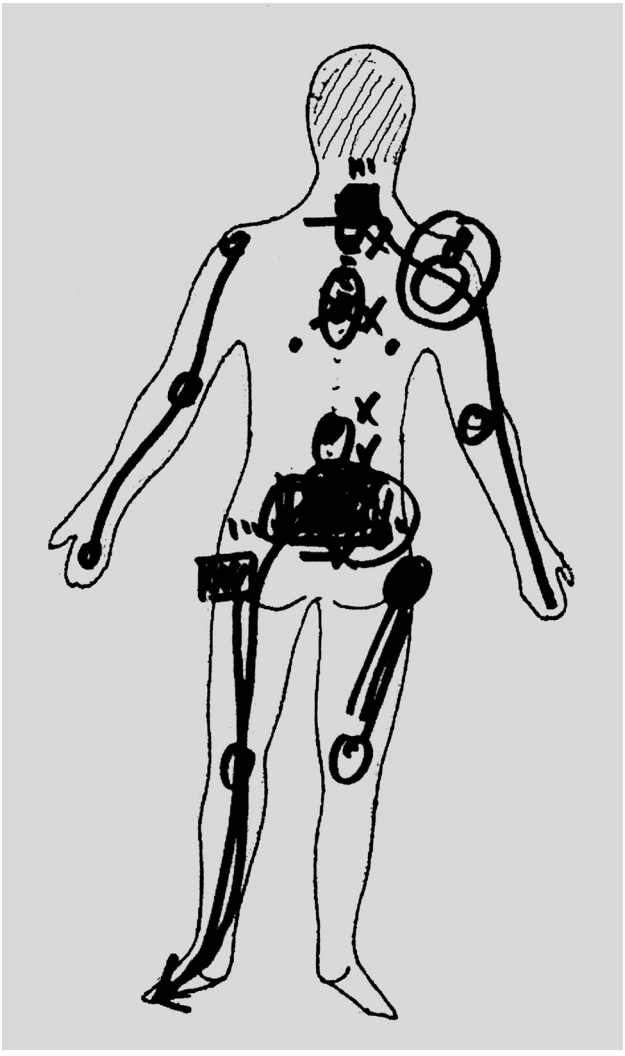

Abbildung 3 Übereinanderprojizierte Schmerzlokalisationen ("Rückenschmerzen") von den 12 Probanden, die am Untersuchungstag Rückenschmerzen angaben

„Was bedeutet für Sie sehr gut/gut/zufriedenstellend/weniger gut/schlecht?", nennen Probanden mit akuten Rückenschmerzen zu 60% Inhalte, die mit Bewegung oder Aktivitäten zu tun haben, und zu 30% Schmerzen; dagegen nennen nur 20% der chronischen Schmerzpatienten Bewegung und 60% Schmerzen (Tab. 2). Daraus lässt sich die Hypothese ableiten, dass für die Probanden mit akuten Rückenschmerzen die (Bewegungs-)Einschränkungen in Alltag, Freizeit oder Beruf im Vordergrund stehen, bei den chronischen Patienten dagegen eher die Schmerzen.

3. Confidence rating

Das Verfahren erlaubt eine grobe Abschätzung der Genauigkeit/Zuverlässigkeit von Angaben bei Faktfragen; einige Beispiele:

Tabelle 2 Ergebnisse des *probing* beim Item "allgemeine Gesundheit"; subjektive Bedeutung der Antworten nach den Kategorien Bewegungseinschränkung vs. Schmerz

Inhalte der Probandenäusserung	akute RS (N = 10)	chronische RS (N = 5)
Bewegung, Sport, Arbeit, Alltagsaktivitäten	6	1
(Rücken-)Schmerzen	3	3
sonstige	2	1

Anmerkung: In der Gruppe mit akuten Rückenschmerzen (RS) findet sich eine Doppelnennung (Bewegung und Schmerzen); 1 (akute RS) bzw. 2 (chronische RS) missing values.

Bei den Angaben zu Krankschreibungszeiten und zu Rehamassnahmen geben fast alle Probanden an, die Angaben seien „sehr genau" (92% bzw. 100%) und sie hätten sich erinnert, nicht geschätzt (jeweils 100%). Angaben zu Körpergrösse und -gewicht werden überwiegend als „sehr genau" (39% bzw. 58%) oder „ziemlich genau" (50% bzw. 32%) eingeschätzt, obwohl der Zeitpunkt der letzten Messung stark differierte (bei Grösse zwischen „vor 3 Tagen" bis „im Pass" bzw. „1989 beim Bund" und bei Gewicht zwischen „täglich" bis „6 Monate").

Diskussion

Unser Ziel, Probleme bei der Beantwortung des Fragebogens zu identifizieren, die mit der Formulierung oder dem Format einzelner Items bzw. der Antwortvorgaben zusammenhängen, wurde erreicht. Immerhin wurden aufgrund der Ergebnisse des *cognitive survey* annähernd ein Drittel der Fragebogenitems verändert; bei einer weniger konservativen Vorgehensweise hätte sich dieser Anteil noch bis auf 50% vergrössern lassen. Dies erscheint als eine gute „Ausbeute", und das Ergebnis lässt den vergleichsweise hohen zeitlichen und personellen Aufwand für die Durchführung des *cognitive survey* insgesamt gerechtfertigt erscheinen. Berücksichtigt man jedoch, dass in einem zweiten Gang nunmehr auch die revidierten Items erneut einem *cognitive survey* unterzogen werden sollten, was hier unterblieben ist, so stellt sich doch die Frage nach dem Verhältnis von personellem und Zeitaufwand einerseits und Effekt andererseits noch einmal anders. Dieses Argument gilt allerdings genauso für den konventionellen Pretest.

Hinweise auf konkreten Änderungsbedarf bei einzelnen Items ergaben sich vor allem aus den Lautlösungsprotokollen. Die vorgenommenen Veränderungen betreffen sowohl die Frageformulierungen als auch formale Aspekte insbesondere bei der Gestaltung der Antwortkategorien. Das *probing* erbrachte wichtige Zusatzinformation zu einzelnen Items (Bedeutungshorizont) und eignet sich nach unserem Eindruck auch gut zum Generieren von Hypothesen. Die *confidence ratings* schliesslich gaben vorsichtige, erste Hinweise auf die Reliabilität der Messung.

Betrachtet man die vorgenommenen Veränderungen der Items im Detail, so wird deutlich, dass es sich in den meisten Fällen eher um ein „Abschleifen der Kanten" handelt und weniger um grundlegende Veränderungen. Dies entspricht weitgehend den Beispielen, die von Willis et al. (1991) angeführt werden. Inwieweit diese Änderungen in der Formulierung einzelner Fragen bzw. dem Antwortformat eine praktische (messbare) Relevanz haben, können wir anhand unserer Daten nicht entscheiden. Damit ist ein grundlegendes Problem beim Einsatz kognitiver Methoden bei der Fragebogenentwicklung angesprochen: Wir wissen nicht, ob die Probleme, die wir bei einzelnen Items entdecken, tatsächlich „real" oder lediglich akademisch sind.

Ausserdem gibt es zumindest keine objektiven Kriterien dafür, ab wann ein Item verändert werden sollte oder nicht. Wo geht es darum, potentielle Probleme lediglich im Vorfeld der Befragung zu identifizieren (*flagging*), und wo besteht tatsächlich Änderungsbedarf? Weder die Zahl der Probanden, die Probleme anmelden, noch die Art des Problems selbst kommen als alleinige Entscheidungsgrundlage infrage. So könnten Probleme, die ausschliesslich von einem Probanden angemeldet werden, später eine wichtige Subgruppe betreffen, oder aber es könnten allein aufgrund der Untersuchungssituation Schwierigkeiten berichtet werden, die „im Feld" gar nicht auftreten. Es fehlt ein „Goldstandard", anhand dessen sich valide und reliable Entscheidungskriterien ableiten liessen (vgl. Willis et al. 1999).

Nicht unterschätzt werden darf auch, dass mit der Aufgabenstellung des lauten Denkens auch ein erheblicher Aufforderungscharakter verbunden ist. Wir wissen wenig darüber, wie die Untersuchungssituation die Probanden beeinflusst (Willis et al. 1999). In unserer Untersuchung entstand der Eindruck, dass einige der Probanden die Situation durchaus nutzten, um Frustrationen aus früheren Befragungen „loszuwerden". Dies aber mag dazu führen, dass Probleme genannt werden, die vielleicht keine wirkliche Relevanz für die spätere Befragung haben. Erst der Vergleich mit den Daten aus dem konventionellen Pretest würde erlauben, die tatsächlichen Problemitems zu identifizieren.

Schliesslich ist wenig darüber bekannt, welcher Stellenwert den kognitiven Verfahren im Vergleich zum konventionellen Pretest zukommt (vgl. Willis et al. 1999). Beide Verfahren haben zumindest teilweise unterschiedliche Ziele und sind keinesfalls alternativ zu betrachten. Das *cognitive survey* soll den üblichen Pretest ja nicht ersetzen, sondern ergänzen. Der konventionelle Pretest kann durch kognitive

Verfahren allein deswegen nicht ersetzt werden, weil es nur so möglich ist, etwa die Forschungslogistik, den Feldzugang oder die Rücklaufquoten zu überprüfen. Wenn die kognitiven Methoden aber eine wirklich relevante Ergänzung zum Pretest darstellen, wäre der Nachweis zu fordern, dass der zusätzliche Aufwand durch einen tatsächlichen (messbaren) Nutzen gerechtfertigt ist.

Darüber hinaus gibt es eine ganze Reihe weiterer offener Fragen und Probleme. Was ist bei der Stichprobenauswahl zu beachten und welche Stichprobengrösse ist anzustreben? Hierzu finden sich keine einheitlichen Vorstellungen in der vorliegenden Literatur (vgl. Prüfer & Rexroth 1996). Wir haben versucht, Frauen und Männer aus verschiedenen Sozialschichten einzubeziehen, die sich in Schwere und Dauer der Rückenbeschwerden unterscheiden, uns bei der Stichprobenkonstruktion also eher von theoretischen Erwägungen leiten lassen. Das muss aber nicht unbedingt zu einer für die zugrunde liegende Fragestellung günstigen Stichprobe geführt haben. Das Ziel ist hier ja, mögliche Probleme möglichst frühzeitig im Prozess der Fragebogenkonstruktion zu identifizieren. Damit ist es weniger Repräsentativität der Stichprobe, die angestrebt werden sollte, sondern eher die Berücksichtigung möglicher Problemgruppen.

Und schliesslich: Nach welchem System sollen die von den Probanden benannten Probleme kodiert werden? Wir haben uns für die groben Kategorien „Formulierung/Inhalt", „Format" und „Weglassen" entschieden; andere Möglichkeiten finden sich z.B. bei Willis et al. (1999). Und ist es wirklich notwendig, den gesamten Fragebogen vorzugeben, wie wir dies getan haben, oder würde es nicht ausreichen, lediglich diejenigen Items einem *cognitive survey* zu unterziehen, die potentiell auch geändert werden sollen (also z. B. bereits eingeführte, standardisierte Instrumente auszunehmen)? Eine Lösung hierfür bietet der Einsatz von Checklisten für mögliche Problemitems, wie er von Lessler und Forsyth (1996) vorgeschlagen wird. Allerdings darf bei einer Auswahl von Items nicht der Gesamtkontext des späteren Fragebogens verlorengehen (Reihenfolgeeffekte). Darüber hinaus ist es interessant herauszufinden, wo auch etablierte Instrumente, die ja oft nicht so getestet wurden, ihre Probleme haben. Zumindest in einem Fall (s.o.) ist dies in unserer Untersuchung der Fall gewesen.

Grenzen der vorliegenden Untersuchung sehen wir einmal darin, dass unsere Stichprobe einen im Vergleich zu der späteren Zielpopulation (Versicherte der Arbeiterrentenversicherung) zu geringen Anteil von Hauptschulabsolventen aufweist. Mögliche Verständnisprobleme wurden dadurch eventuell unterschätzt. Zum anderen erscheint uns die von uns gewählte handschriftliche Protokollierung zumindest prinzipiell als problematisch. Ein Tonbandmitschnitt der Probandenäusserungen würde die Objektivität der Untersuchung deutlich erhöhen und sicherlich eine differenziertere Auswertung erlauben. Einer der beiden Gutachter berichtet in diesem Zusammenhang von der eigenen Praxis, die Interviews aufzunehmen und die anschliessende Vercodung anhand der Bandmitschnitte vorzunehmen.

Zusammenfassung

Fragestellung: Der konventionelle Pretest in der Fragebogenentwicklung erlaubt Aussagen zu technischen Problemen der Itemkonstruktion. Ausser acht gelassen werden dabei die kognitiven Prozesse, die bei der Beantwortung von Fragebogenitems eine Rolle spielen.

Methoden: Im Rahmen eines *cognitive survey* an 20 Probanden werden die Items eines Screeningbogens für Rehabedarf wegen Dorsopathien in Hinblick auf semantischen Gehalt und Verlässlichkeit untersucht. Einige wesentliche Techniken (*think-aloud, probing, confidence ratings*) werden beschrieben. Sie werden dann in einer standardisierten Untersuchungssituation bei einer Gelegenheitsstichprobe von 20 Patienten mit akuten und chronischen Rückenschmerzen eingesetzt, um den Fragebogen zu überprüfen.

Ergebnisse: Das Ziel, Probleme bei der Beantwortung einzelner Items zu identifizieren, wurde erreicht. Annähernd ein Drittel der Fragebogenitems wurden aufgrund der Ergebnisse verändert. Als besonders aussagekräftig erwies sich die Technik des *think-aloud*. Die Veränderungen des Fragebogens werden an Beispielen illustriert.

Schlussfolgerungen: Trotz des im Vergleich zum konventionellen Pretest vergleichbar hohen Aufwandes erscheint die Methode des *cognitive survey* lohnend. Die beiden Methoden haben spezifische Vor- und Nachteile und ergänzen sich.

Soz.- Präventivmed. 48 (2003) 55–64

Résumé

Questions à propos de questions: la méthode de l'enquête cognitive dans le développement des items d´un questionnaire
Objectifs: Le pré-test réalisé habituellement lors de l'élaboration d'un questionnaire permet d'obtenir des informations sur les problèmes techniques en rapport avec la construction des items. Les phénomènes d'ordre cognitif qui jouent un rôle dans la réponse aux items ne sont alors pas pris en considération.
Méthodes: Le contenu sémantique et la validité des items d'un questionnaire pour le dépistage du besoin en réadaptation dans les dorsalgies ont été examinés dans le cadre d'une enquête cognitive pratiquée auprès de 20 sujets. Nous décrivons certaines techniques essentielles de ce type d'enquête (*think-aloud, probing, confidence ratings*), qui ont été utilisées auprès d'un échantillon de 20 malades atteints de dorsalgies aiguës ou chroniques dans le but de tester le questionnaire.
Résultats: L'objectif d'identifier les problèmes survenant lors de la réponse à certains items du questionnaire a été atteint. Un tiers environ des items a été modifié à la suite des résultats. La technique du *think-aloud* s'est avérée particulièrement intéressante. L'article illustre à l'aide d'exemples les modifications du questionnaire.
Conclusions: La méthode de l'enquête cognitive s'avère avantageuse, malgré un surcroît de travail par rapport aux pré-tests traditionnels. Les deux méthodes ont des avantages et des inconvénients spécifiques et se complètent.

Literaturverzeichnis

Cantril H (1944). Gauging public opinion. Princeton: Princeton University Press.

Converse R (1970). Attitudes and nonattitudes: continuation of a dialogue. In: Tufte ER, ed. The quantitative anlaysis of social problems. Reading: Addison-Wesley: 168–89.

Duncker K (1935). Zur Psychologie des produktiven Denkens. Berlin: Springer.

Ebbinghaus H (1964). Memory: a contribution to experimental psychology. New York: Dover (Original 1894 veröffentl.).

Ferber R (1956). The effect of respondent ignorance on survey results. J Am Stat Assoc *51*: 576–86.

Flor H, Turk CD (1992). Chronic back pain and rheumatoid arthritis: predicting pain and disability from cognitive variables. J Behav Med *11*: 251–65.

Groves RM, Fultz NH, Martin E (1991). Direct questioning about comprehension in a survey setting. In: Tanur JM, ed. Questions about questions. New York: Russel Sage: 49–61.

Jobe JB, Mingay DJ (1990). Cognitive laboratory approach to designing questionaires for surveys of the elderly. Public Health Rep *105*: 518–24.

Kohlmann T, Raspe H (1996). Der Funktionsfragebogen Hannover zur alltagsnahen Diagnostik der Funktionsbeeinträchtigung durch Rückenschmerzen (FFbH-R). Rehabilitation *35*: I–VIII.

Kurz K, Prüfer P, Rexroth M (1999). Zur Validität von Fragen in standardisierten Erhebungen: Ergebnisse des Einsatzes eines kognitiven Pretestinterviews. ZUMA-Nachrichten *44*: 83–107.

Lessler JT, Forsyth BH (1996). A coding system for appraising questionaires. In: Schwarz N, Sudman S, eds. Answering questions: methodology for determining cognitive and communicative processes in survey research. San Francisco: Jossey-Bass: 259–91.

Loftus EF, Smith KD, Klinger MR, Fiedler J (1991). Memory and mismemory of health events. In: Tanur JM ed. Questions about questions. New York: Russel Sage: 102–37.

Means B, Swan GE, Jobe JB, Esposito JL, Loftus EF (1989). Recall strategies for estimation of smoking levels in health surveys. American Statistical Association: Proceedings of the Section on Survey Research Methods: 421–4.

Menon G (1994). Judgements of behavioral frequencies: memory search and retrieval strategies. In: Schwarz N, Sudman S, eds. Autobiographical memory and the validity of retrospective reports. New York: Springer: 161–72.

Nuckols R (1953). A note on pre-testing public opinion questions. J Appl Psychol 37: 119–20.

Prüfer P, Rexroth M (1996). Verfahren zur Evaluation von Survey-Fragen: ein Überblick. Mannheim: ZUMA. (ZUMA-Arbeitsbericht; Nr. 96/05).

Schechter S, Blair J, Vande Hey J. (1996). Conducting cognitive interviews to test self-administered and telephone surveys: which method should we use? American Statistical Association: Proceedings of the Section of Survey Research Methods: 10–7.

Schnell R, Hill PB, Esser E (1993). Methoden der empirischen Sozialforschung. 4., überarb. Aufl. München; Wien: R. Oldenbourg.

Schumann H (1966). The random probe: a technique for evaluating the validity of closed questions. Am Sociol Rev *31*: 218–22.

Schwarz N, Sudman S, eds. (1996). Answering questions: methodology for determining cognitive and communicative processes in survey research. San Francisco: Jossey-Bass.

Sirken MG, Herrmann DJ, Schechter S, Schwarz N, Tanur JM, Tourangeau R, eds. (1999). Cognition and survey research. New York: Wiley.

Sudman S, Bradburn NM, Schwarz N (1996). Thinking about answers. San Francisco: Jossey-Bass.

Tanur JM, ed. (1991). Questions about questions. New York: Russel Sage.

Welzer H, Markowitsch HJ (2001). Umrisse einer interdisziplinären Gedächtnisforschung. Psychol Runds *52*: 205–14.

Willis GB, DeMaio TJ, Harris-Kojetin BH (1999). Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. In: Sirken MG, Herrmann DJ, Schechter S, Schwarz N, Tanur JM, Tourangeau R, eds. Cognition and survey research. New York: Wiley: 133–53.

Willis GB, Royston P, Bercini D (1991). The use of verbal report methods in the development and testing of survey questionaires. Appl Cogn Psychol *5*: 251–67.

Korrespondenzadresse

Dr. Oskar Mittag
Institut für Sozialmedizin
Beckergrube 43-47
D-23552 Lübeck
Tel.: +49 451 7992515
Fax: +49 451 7992522
e-mail: oskar.mittag@sozmed.mu-luebeck.de

Soz.- Präventivmed. 48 (2003) 55–64