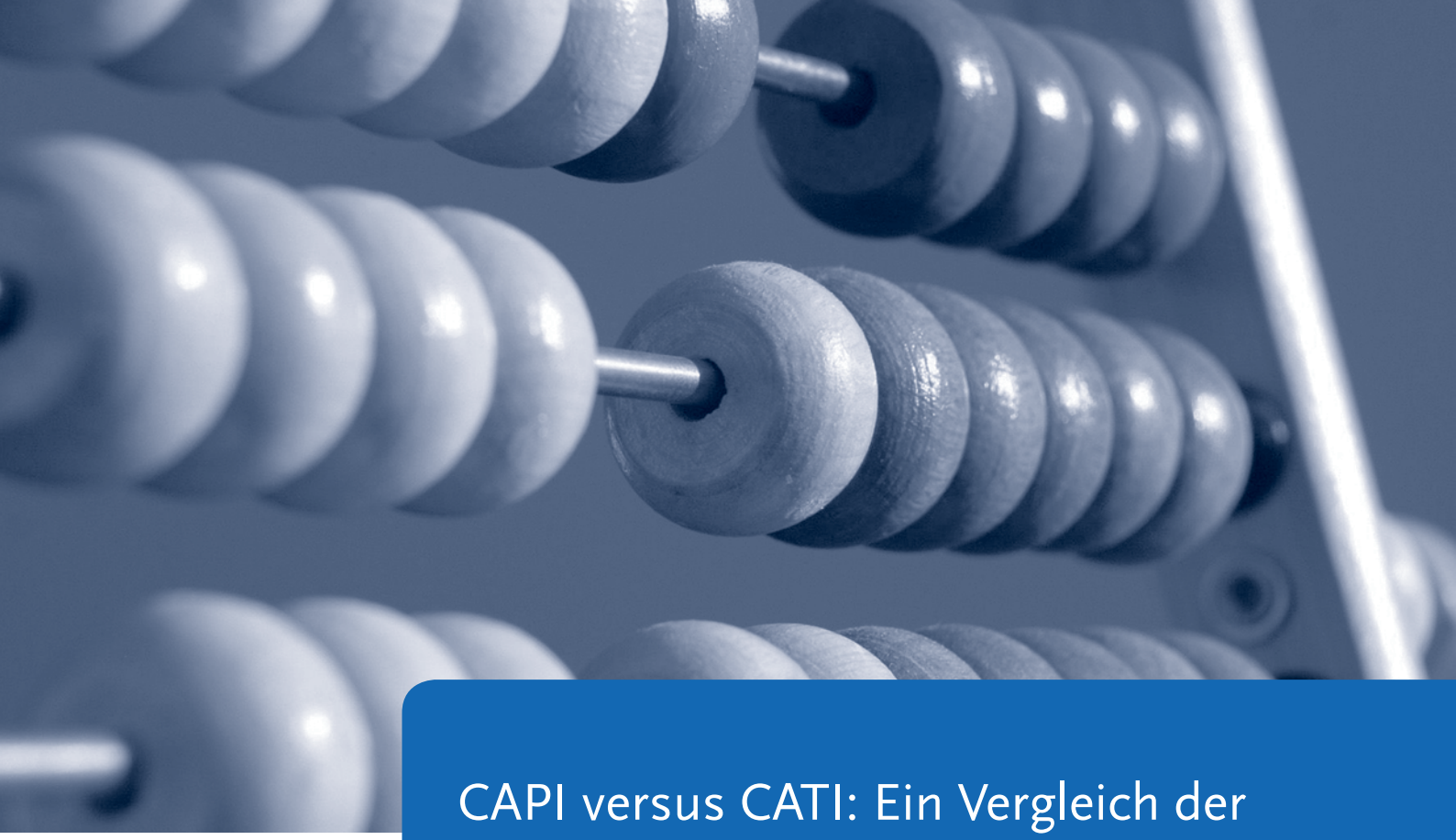


CAPI versus CATI: Ein Vergleich der BIBB/IAB-Erhebung 1998 mit der BIBB/BAuA-Erwerbstätigenbefragung 2005

M. Holler, N. Tschersich, M. Gensicke

baua:
Bundesanstalt für Arbeitsschutz
und Arbeitsmedizin

Forschung
Projekt F 2240

M. Holler
N. Tschersich
M. Gensicke

CAPI versus CATI: Ein Vergleich der BIBB/IAB-Erhebung 1998 mit der BIBB/BAuA-Erwerbstätigenbefragung 2005

Dortmund/Berlin/Dresden 2012

Diese Veröffentlichung ist die Kurzfassung des Abschlussberichtes zum Projekt „Methodenvergleich: Analyse der methodenimmanenten Implikationen verschiedener Befragungsinstrumente zur Beschreibung der Situation Erwerbstätiger“ – Projekt F 2240 – im Auftrag der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.
Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Autoren: Markus Holler
Internationales Institut für Empirische Sozialökonomie gGmbH, INIFES
Haldenweg 23, 86391 Stadtbergen

Nikolai Tschersich
Miriam Gensicke
TNS Infratest Sozialforschung GmbH
Landsberger Straße 284, 80687 München

Titelfoto: Uwe Völkner, Fotoagentur FOX, Lindlar/Köln

Umschlaggestaltung: Rainer Klemm
Bundesanstalt für Arbeitsschutz und Arbeitsmedizin

Herausgeber: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin
Friedrich-Henkel-Weg 1 − 25, 44149 Dortmund
Telefon 0231 9071-0
Fax 0231 9071-2454
poststelle@baua.bund.de
www.baua.de

Berlin:
Nöldnerstr. 40 − 42, 10317 Berlin
Telefon 030 51548-0
Fax 030 51548-4170

Dresden:
Fabricestr. 8, 01099 Dresden
Telefon 0351 5639-50
Fax 0351 5639-5210

Alle Rechte einschließlich der fotomechanischen Wiedergabe und des auszugsweisen Nachdrucks vorbehalten.

Inhaltsverzeichnis

		Seite
1	Forschungszusammenhang	4
2	Methodeneffekte: Literaturanalyse und empirische Hinweise	5
2.1	Vergleichbarkeit von Antwortalternativen und Skalierungen in den Erwerbstätigenbefragungen	5
2.2	Antwortverhalten von Befragten bei persönlichen und telefonischen Befragungen	6
3	Reliabilität und Validität	7
3.1	Strukturgleichungsmodelle	9
3.2	Erfassung berufsgruppenspezifischer Arbeitsbedingungen	13
4	Explorative Faktorenanalysen und Indexkonstruktionen: BIBB/BAuA 2005 und andere Datenquellen im Vergleich	14
5	Zusammenfassende Bewertung und Ableitung von Designempfehlungen	16
Literaturverzeichnis		18

1 Forschungszusammenhang

Der Erhalt und die Förderung von Arbeitsfähigkeit sind angesichts des Wandels der Arbeitswelt, der demographischen Entwicklung und im Zusammenhang mit sozialpolitischen Herausforderungen wichtige Aufgaben unserer Gesellschaft. Notwendig sind dazu empirisch gesicherte Erkenntnisse, die ihrerseits belastbare Daten zur gegenwärtigen Erwerbssituation und Arbeitsbelastung sowie zur Gesundheit der Erwerbsbevölkerung voraussetzen. Eine der wenigen für diesen Zweck geeigneten Datenquellen sind die Erwerbstätigenbefragungen des Bundesinstituts für Berufsbildung (BIBB) und des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) bzw. der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA). 2005 wurde die Erhebungsmethodik von bis dato computerunterstützten persönlichen Interviews (CAPI) auf computerunterstützte telefonische Interviews (CATI) umgestellt. In diesem Zusammenhang wurden einige Frageformulierungen und Antwortkategorien gegenüber den vorherigen Erhebungswellen verändert. Ein Vergleich der Ergebnisse der Erwerbstätigenbefragungen von 1998 und 2005 ist aus diesem Grund lediglich mit Einschränkungen möglich. Wie ein umfangreicher Literaturbestand dokumentiert, erzeugen verschiedene Erhebungsmethoden sowie Änderungen in Frage- oder Antwortformulierungen häufig unterschiedliche Ergebnisse, deren Ausmaß nicht immer vernachlässigbar ist (vgl. z. B. SCHUMANN und PRESSER, 1996; FAULBAUM et al., 2009).

Vereinfacht kann davon ausgegangen werden, dass sich Ergebnisunterschiede beim Vergleich einer persönlichen und einer telefonischen Befragung aus einem Coverage- bzw. Abdeckungsfehler, einem Nonresponse-Fehler und einem Messfehler aufgrund der Erhebungsmethode zusammensetzen (vgl. LASS et al., 1997).[1] Eine Quantifizierung dieser Anteile ist beim Vergleich der beiden Erwerbstätigenbefragungen von 1998 und 2005 allerdings nicht möglich, da es sich um zwei getrennte Erhebungen mit großem zeitlichem Abstand handelt. Hinzu kommen im Fall der Erwerbstätigenbefragungen zudem Ergebnisunterschiede aufgrund veränderter Antwortmöglichkeiten. Zum Vorliegen von bestimmten Methodeneffekten können auf dieser Datenbasis nur grobe Hinweise gesammelt werden. Hauptsächliches Forschungsziel des Projektes waren daher Analysen zur Reliabilität und Validität der Daten, die vergleichend gegenübergestellt wurden.

Der vorliegende Artikel stellt eine Zusammenfassung ausgewählter Ergebnisse des Forschungsprojektes dar. Beispielsweise erfolgte neben den hier zusammengefassten uni- und multivariaten Analysen ein Vergleich der beiden Netto-Stichproben, der in diesem Rahmen nicht dargestellt ist.[2] Interessierte Leser seien für die Beschreibung der beiden Stichproben auf die Berichte von HARTMANN (2006a, 2006b), HALL (2009), HARTMANN et al. (1999) und BIBB/IAB (2003) verwiesen.

---

[1] Genauer gesagt sind bei Umfragen Stichprobenfehler und Nicht-Stichprobenfehler zu unterscheiden. Letztere lassen sich einteilen in Nicht-Beobachtungsfehler (Coverage- bzw. Abdeckungsfehler und Nonresponse-Fehler) und Beobachtungsfehler, welche – abgesehen von Verarbeitungsfehlern oder technischen Fehlern – Messfehler der Befragungsmethoden sind (vgl. FAULBAUM et al., 2009).

[2] Es wird davon ausgegangen, dass gewisse bestehende Unterschiede hinsichtlich der Bildung der Befragten (Bildungsbias) durch die Gewichtung ausgeglichen werden. Andere Unterschiede (z. B. Anteil von befristet Beschäftigten etc.) lassen sich durch den zeitlichen Abstand der Erhebungen erklären.

2 Methodeneffekte: Literaturanalyse und empirische Hinweise

2.1 Vergleichbarkeit von Antwortalternativen und Skalierungen in den Erwerbstätigenbefragungen

In der empirischen Sozial- und Methodenforschung gibt es zahlreiche Beispiele für scheinbar äquivalente Fragen, die aber aufgrund geringer Änderungen in der Verbalisierung der Frage und der Antwortmöglichkeiten selbst bei einer identischen Erhebungsmethodik zu unterschiedlichen Ergebnissen führen. Neben einer Sichtung von möglichen Effekten durch die Untersuchungsmethode („mode effects“) in der Literatur (und soweit möglich in den Daten) ist daher die Vergleichbarkeit der Fragestellungen von Bedeutung. In der Untersuchung wurden daher nur ausgewählte, für den Auftraggeber inhaltlich relevante Variablen untersucht, die zudem in beiden Erhebungen in vergleichbarer Fragestellung vorliegen. Ausgewählt wurden Variablen aus den Bereichen körperliche Arbeitsbedingungen, psychische Arbeitsbedingungen, Gesundheit bzw. gesundheitliche Beschwerden und Zufriedenheit mit der Erwerbssituation.

Bei den Fragen zu den körperlichen und psychischen Arbeitsbedingungen wurde die fünfstufige unipolare Häufigkeitsskala der Erhebung von 1998 („Praktisch immer“, „Häufig“, „Immer mal wieder“, „Selten“, „Praktisch nie“) in 2005 auf eine vierstufige Skala reduziert („Häufig“, „Manchmal“, „Selten“, „Nie“). Aufgrund der telefonischen Befragung war die Antwortskala für den Befragten optisch nicht verfügbar, sondern wurde vom Interviewer genannt. In der CAPI-Erhebung von 1998 wurde die Skala mit Hilfe einer Listenvorlage visualisiert. Die Bedeutung der beiden Skalen ist inhaltlich nahezu äquivalent, wenn die Kategorien „Praktisch immer“ und „Häufig“ aus der Befragung von 1998 zur Kategorie „Häufig“ zusammengefasst werden, wie sie auch in der Erwerbstätigenbefragung von 2005 enthalten war.

Bei den in den Erwerbstätigenbefragungen von 1998 und 2005 enthaltenen Zufriedenheitsabfragen wurde nicht die Länge der Antwortskala, sondern ihre Verbalisierung geändert. Die Zufriedenheitsabfragen 1998 waren aufgrund ihrer Abstufung zwischen den beiden Polen „zufrieden“ und „unzufrieden“ bipolar („Sehr zufrieden“, „Im großen und ganzen zufrieden“, „Eher unzufrieden“, „Sehr unzufrieden“) und wurden mit Hilfe einer Listenvorlage erhoben. Die neuere Version der Zufriedenheitsabfrage beinhaltet mit „zufrieden“ nur eine Polarität („Sehr zufrieden“, „Zufrieden“, „Weniger zufrieden“, „Nicht zufrieden“).

Die möglichen Effekte von veränderten Antwortmöglichkeiten können anhand der Erwerbstätigenbefragungen von 1998 und 2005 aufgrund des zeitlichen Abstands nicht eindeutig nachgewiesen werden. Es kann lediglich der „total survey error“ einschließlich der Veränderung der Ergebnisse durch den zeitlichen Abstand der Erhebungen ermittelt werden. Um die Plausibilität von möglichen Methodeneffekten zu prüfen, wurden daher auch vergleichende Analysen mit anderen Datensätzen[3] ange-

[3] Hauptsächlich wurden Plausibilitätsprüfungen mit den European Working Conditions Surveys 2000 und 2005 sowie mit der INQA-Erhebung „Was ist gute Arbeit?“ angestellt. Die Untersuchungspopulationen der verwendeten Erhebungen wurden aneinander angeglichen.

stellt. Ergebnisse, die in die vermutete Richtung weisen, sind dennoch nur als ein grober Hinweis zu verstehen. Um eindeutigere Aussagen zu treffen, müsste ein Methodenexperiment durchgeführt werden.

Die Ergebnisse deuten darauf hin, dass die Antworten der Probanden durch den Wechsel von der bipolaren zu einer unipolaren Zufriedenheitsskala beeinflusst werden. Tendenziell zeigt sich in der Erhebung von 2005 im Vergleich zur Erhebung von 1998 im Durchschnitt der Fragen eine leichte Zunahme des Anteils der negativsten Kategorie und eine Abnahme der zweitpositivsten Kategorie. Die Vermutung, dass die Mittelkategorie der 5er-Skala in der Erhebung von 1998 („Immer mal wieder)“ häufiger genannt wurde als die inhaltlich ähnliche Kategorie „Manchmal“ in 2005 (4er-Skala), bestätigte sich dagegen nicht. Gleiches gilt für die Vermutung, dass die zwei Kategorien „Praktisch immer“ und „Häufig“ in der Erhebung von 1998 einen größeren Anteil der Antworten auf sich vereinen als die Kategorie „Häufig“ in der Erhebung von 2005.

Theoretisch ist zu erwarten, dass die Validität bei mehr Skalenpunkten ansteigt und die Reliabilität bei 5 bis 7 Skalenpunkten am höchsten ist (vgl. FAULBAUM et al. 2009). Im Zuge der Analysen zu Konstrukten und Strukturgleichungsmodellen wurde daher ein Vergleich der Messung mit der 1998 verwendeten 5er-Skala sowie der 2005 verwendeten 4er-Skala vorgenommen (vgl. Kapitel 3).

2.2 Antwortverhalten von Befragten bei persönlichen und telefonischen Befragungen

Der Methodenliteratur kann eine Vielzahl von Effekten entnommen werden, die bei einer Umstellung von einer persönlichen auf eine telefonische Befragung zu erwarten sind. Dazu gehören u. a.: Effekte sozialer Erwünschtheit[4], Zustimmungstendenzen (vgl. KROSNICK, 1999), nichtsubstantielle Antworten (vgl. SCHUMANN und PRESSER, 1996), die häufigere Nennung positiver Extremkategorien bzw. generell extremere Antworten bei Ratingskalen (vgl. CHRISTIAN et al., 2008), Recency-Effekte insbesondere bei längeren Listen (vgl. KROSNICK, 1999), Nondifferentiation bei Ratingskalen (vgl. YAN, 2008). Die meisten dieser Effekte lassen sich in Einklang mit der Theorie des „satisficing“ von Krosnick bringen. Aufgrund der schnelleren Gesprächssituation (vgl. HÄDER et al., 2009; CHRISTIAN et al., 2008), der geringeren Kontrolle über die Interviewsituation (vgl. ANAND, 2008) und des aufgrund der lediglich akustischen Information höheren Schwierigkeitsgrades der Fragen neigen Probanden in telefonischen Interviews eher dazu, Fragen nicht optimal zu beantworten, sondern den Aufwand für die Antwortgenerierung gering zu halten (vgl. KROSNICK, 1999).

---

[4] Das telefonische Interview ist im Vergleich zur persönlichen Befragung zwar anonymer und ermöglicht es, ungehemmter zu antworten, was insbesondere bei heiklen Fragen von Bedeutung ist (vgl. SCHOLL, 2009). Umgekehrt wird aber auch davon berichtet, dass in der Face-to-face-Situation ehrlichere Antworten gegeben werden, da eine größere Verbindlichkeit bzw. Vertrautheit zwischen Interviewer und Befragtem besteht (vgl. LAMNEK und SCHÄFER, 1998; vgl. BRICK und LEPOWSKI, 2008).

Als zentrales Ergebnis der durchgeführten Analysen lässt sich festhalten, dass sich die in der Fachliteratur identifizierten Methodeneffekte mit den Daten der Erwerbstätigenbefragungen 1998 und 2005 überwiegend nicht bestätigen lassen. Bei den Zufriedenheitsfragen in der telefonischen Erhebung ist eine häufigere Nennung der (an erster Stelle abgefragten) positiven Extremkategorie zu vermuten, belastbare Schlussfolgerungen lassen sich jedoch nur schwer ziehen, da gleichzeitig mehrere methodische Änderungen vorliegen und ein langer Zeitraum zwischen beiden Erhebungen liegt. Somit ist es nicht möglich zu beurteilen, ob die laut Literatur erwartbaren Methodeneffekte aufgrund dieser Unterschiede nicht nachzuweisen sind oder ob empirisch keine Messunterschiede vorliegen. Zudem ist anzumerken, dass auch ohne diese Probleme die Ergebnisse von Studien über bestimmte Methodeneffekte weitaus weniger übereinstimmen, als theoretisch zu erwarten ist. Methodeneffekte sind vermutlich durch die Interaktion des Erhebungsmodus, des Interviewers, des Befragten, der Erhebung und ihrem Inhalt sowie der Art ihrer Präsentation bedingt und in ihren Ursachen somit äußerst komplex (vgl. BRICK und LEPOWSKI, 2008).

Um Aussagen über die Datenqualität treffen zu können, wurden daher „relationale“ Vergleiche bzw. Zusammenhangsanalysen innerhalb der beiden Erhebungen angestellt. Zum einen betrifft dies die Bildung von Indizes und die Durchführung von Reliabilitäts- und Validitätsanalysen, zum anderen wurden Zusammenhangsanalysen bzw. Strukturgleichungsmodelle berechnet.

3 Reliabilität und Validität

Analysen zur Reliabilität und Validität, sollten zeigen, ob die Messungen in beiden Erhebungen vergleichbar sind. Diesem Vorgehen liegt die Annahme zugrunde, dass Validität dann vorliegt, wenn zwischen einem Konstrukt und mindestens einem geeigneten Außenkriterium (zumeist auch in Form eines Konstrukts) eine hohe Übereinstimmung besteht (vgl. WEIBER und MÜHLHAUS, 2010). Aus diesem Grund wurden sachlogisch bzw. theoretisch begründete Hypothesen über Zusammenhänge zwischen zwei oder mehreren Konstrukten gebildet. Bei einer hohen Modellgüte kann davon ausgegangen werden, dass es sich um valide Messungen handelt.

Zur Überprüfung der Konstruktvalidität[5] wurden mehrere Strukturgleichungsmodelle entwickelt, die mit den Daten der BIBB/BAuA-Erwerbstätigenbefragungen 1998 und 2005 gerechnet wurden. Das Ziel war es zu überprüfen, ob in beiden Erhebungen mit unterschiedlichen Modi dasselbe gemessen wurde bzw. dies klarer zu identifizieren als dies mit deskriptiven Analysen möglich ist. Zudem weisen Zusammenhänge und Indizes zeitlich eine größere Stabilität auf als vergleichende Analysen einzelner Items.

---

[5] Konstruktvalidität in einem umfassenden Verständnis schließt den Aspekt der Kriteriumsvalidität mit ein (vgl. KNAPP, 2008). Dennoch wird zum Teil noch zwischen Konstruktvalidität und Kriteriumsvalidität unterschieden. Im vorliegenden Artikel wird z. B. im Rahmen der Strukturgleichungsmodelle Konstruktvalidität in einem umfassenden Sinn behandelt, wohingegen die gruppenspezifischen Analysen in Kapitel 3.2 als Test der Kriteriumsvalidität zu sehen sind.

Wenn sich die vermuteten Zusammenhänge in beiden Erhebungen gleichermaßen zeigen, kann dies dahingehend interpretiert werden, dass die Konstrukte in beiden Erhebungen vergleichbar gemessen wurden. Dass dies nicht selbstverständlich ist, zeigen beispielsweise die Analysen von DE LEEUW et al. (1996), die für die Hypothese sprechen, dass Methodeneffekte in multivariaten Analysen eine Rolle spielen bzw. sich in mutlivariaten Analysen sogar vergrößern. Diesem Ergebnis steht die „form-resistant correlation hypothesis“ gegenüber, derzufolge Korrelationen stabil gegenüber eventuellen Methodeneffekten sind.

Zur Bewertung von Reliabilität und Validität wurden im Forschungsprojekt Gütekriterien und Schwellenwerte definiert, die in der folgenden Tabelle im Überblick dargestellt sind.

Tab. 3.1 Überblick über verwendete Gütekriterien und definierte Schwellenwerte

	Verwendete Maße	**Schwellenwerte**[6]
Signifikanz	Critical Ratio	C.R. > |1,96|
	P-Werte	P < 0,001
Reliabilität	Faktorreliabilität	Rel (ξ_j) 0,6
	Durchschnittlich extrahierte Varianz (DEV)	DEV (ξj) 0,5
	Cronbachs Alpha	α 0,70
Modellgüte	Standardized Root Mean Square (SRMR)	SRMR 0,10
	Comparative Fit Index (CFI)	CFI 0,90
	Root-Mean-Square-Error of Approximation (RMSEA)	≤0.05: guter Modell-Fit ≤ 0.08: akzeptabler Modell-Fit ≥ 0.10: inakzeptabler Modell-Fit

Insgesamt wurden fünf Konstrukte gebildet und bewertet. Tabelle 3.1 gibt einen Überblick über die Bewertungsergebnisse der fünf Konstrukte. Bei vier der fünf Konstrukte waren die Gütemaße für die Modellgüte in einem guten oder akzeptablen Bereich. Lediglich bei Konstrukt 5 war die Modellgüte in beiden Datensätzen nicht ausreichend.

Der Mehrgruppenvergleich zeigt für vier der fünf Konstrukte, dass in beiden Erhebungen relativ gut dasselbe gemessen wurde. Eine Ausnahme stellt hier das Konstrukt “Psychovegetative Beschwerden” dar, wobei die etwas besseren Werte der Erhebung 2005 auch in einer realen Zunahme dieser Beschwerden im Verlauf der sieben Jahre, die zwischen den beiden Erhebungen liegen, begründet sein kann.

---

[6] Die Wertebereiche dieser Gütemaße liegen zwischen Null und Eins – einzige Ausnahme ist die Critical Ratio, deren Werte von minus Unendlich bis plus Unendlich gehen können.

Tab. 3.2 Überblick zu den Konstruktbewertungen[7]

Konstrukt		Reliabilitäten		Modellgüte		Mehrgruppen-vergleich
		2005	1998	2005	1998	
1	Psychische Arbeitsbedingungen	o / -	+ / o	++	+	+
2	Physische Arbeitsbedingungen	++	++	+	+	++
3	Psychovegetative Beschwerden	+ / o	-	++	++	-
4	Muskel-/ Skelettbeschwerden	-	-	++	++	+
5	Arbeitszufriedenheit	o / -	o	-	-	+

Legende: + : gute Bewertung; – : negative Bewertung; o : neutral/nicht eindeutig.

3.1 Strukturgleichungsmodelle

Basis der Strukturgleichungsmodelle waren die fünf Konstrukte zu Arbeitsbedingungen und zur Lebens-/Gesundheitssituation. Für diese Konstrukte wurden auf der Basis von sachlogisch begründeten Hypothesen Zusammenhänge formuliert und im Rahmen von Strukturgleichungsmodellen überprüft.

Es wurden vier Hypothesen zu Zusammenhängen von Arbeitsbedingungen und Lebenssituation formuliert. Um einen Vergleich der beiden Erhebungen zu ermöglichen, war die Voraussetzung, dass in die Konstrukte nur Items einbezogen wurden, die in beiden Erhebungen möglichst identisch erhoben wurden. Mit Hilfe der Strukturgleichungsmodelle wurde anschließend überprüft, inwieweit die Ergebnisse mit den Hypothesen übereinstimmen und ob sich in beiden Erhebungen ähnliche Ergebnisse zeigen. Ein guter Modell-Fit in Kombination mit einem bedeutungsvollen Zusammenhang lässt auf eine hohe Übereinstimmung zwischen Konstrukten schließen – was letztendlich auf Konstruktvalidität hindeutet.

Hypothese 1: Je belastender die psychischen Arbeitsbedingungen, desto wahrscheinlicher sind psychovegetative Beschwerden.

Hypothese 2: Je belastender die physischen Arbeitsbedingungen, desto wahrscheinlicher sind Muskel-/Skelettbeschwerden. Je belastender die psychischen Arbeitsbedingungen, desto wahrscheinlicher sind psychovegetative Beschwerden sowie Muskel-/Skelettbeschwerden.

---

[7] Die Konstrukte setzen sich aus folgenden Items zusammen:
Konstrukt 1 und 3: siehe Abbildung 3.1.
Konstrukt 2: Arbeiten im Stehen, Lasten heben und tragen, in gebückter/kniender/liegender Stellung arbeiten, Erschütterungen/Stöße.
Konstrukt 4: Kreuzschmerzen, Schmerzen in den Armen/Händen, Schmerzen in der Hüfte, Schmerzen in den Knien, Schmerzen in den Beinen/Füßen.
Konstrukt 5: Zufriedenheit mit Einkommen, Aufstiegsmöglichkeiten, Arbeitszeit, Betriebsklima, Vorgesetzten, Art und Inhalt der Tätigkeit, räumlichen Gegebenheiten, Möglichkeiten, eigene Fähigkeiten anzuwenden, Weiterbildungsmöglichkeiten.

Hypothese 3: Je belastender die zeitlichen Arbeitsbedingungen sind, desto wahrscheinlicher sind psychovegetative sowie Muskel-/Skelettbeschwerden.

Hypothese 4: Betriebliche Umorganisation erhöht das Risiko von Stellenabbau. Stellenabbau erhöht die Wahrscheinlichkeit, dass Arbeitnehmer weniger zufrieden mit dem Vorgesetzten und dem Betriebsklima sind.

Die vier beschriebenen Hypothesen wurden basierend auf den Konstrukten in Strukturgleichungsmodelle umgesetzt. Dabei wurde jeweils ein Modell mit den Daten der Erwerbstätigenerhebung 1998 und den Daten der Erhebung 2005 gerechnet. Die Ergebnisse zeigen generell Folgendes auf:

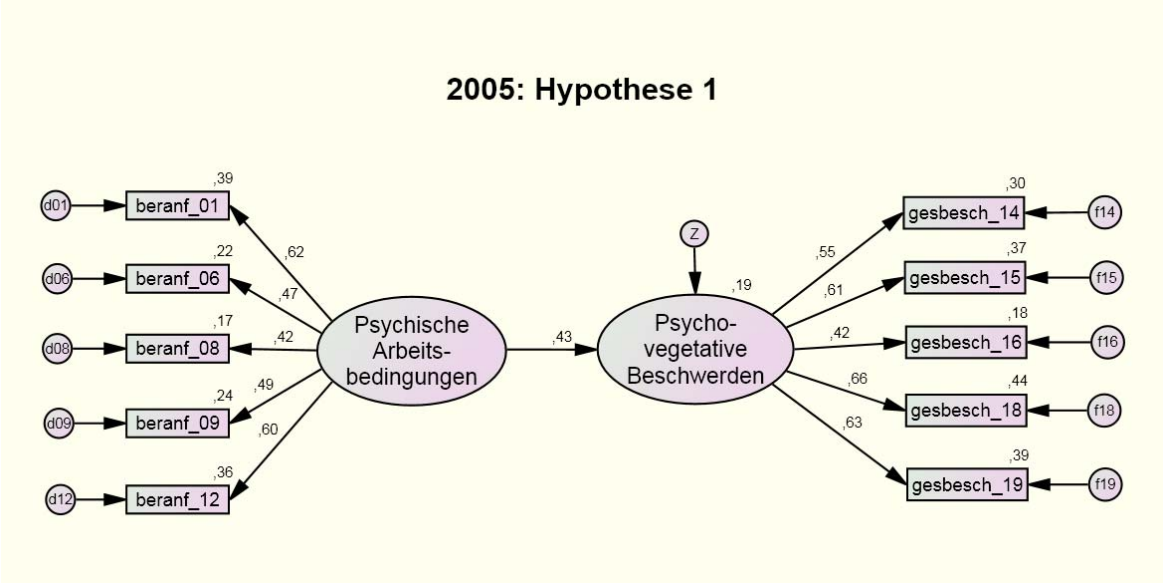
Zusammenhang: Greifen die theoretischen Überlegungen, d. h. lassen sich Zusammenhänge nachweisen? Lassen sich auf Basis der beiden Erhebungen grundsätzlich gleichgerichtete Zusammenhänge nachweisen? Sind die festgestellten Zusammenhänge jeweils etwa gleich stark?

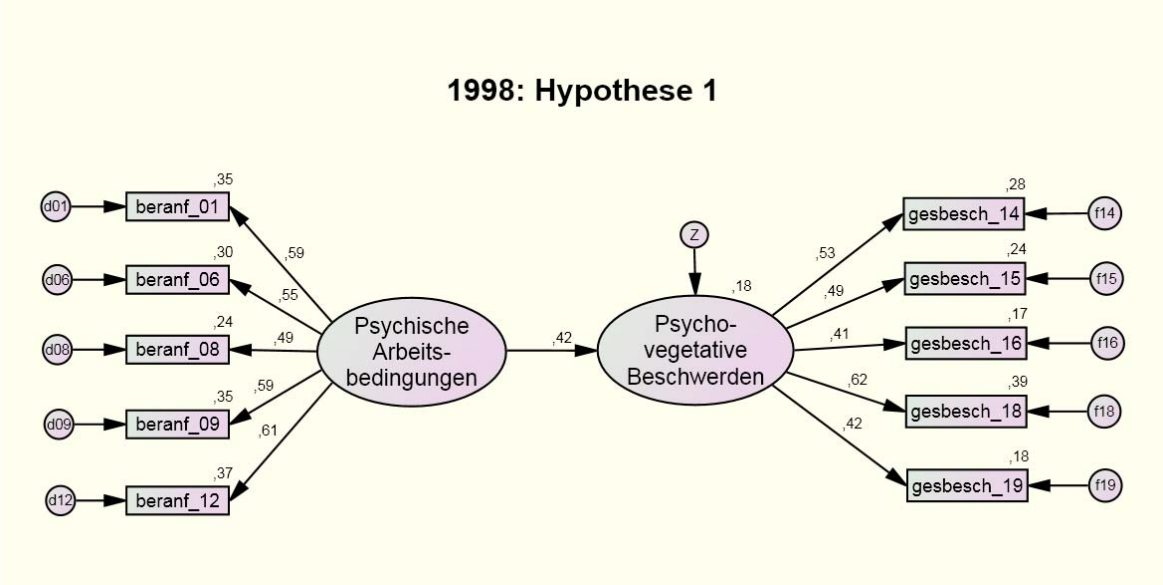
Ausreichende Modellgüte: Ist die jeweilige Modellgüte ausreichend gut, um die Ergebnisse als valide einschätzen zu können?

Vergleichbare Modellgüte: Ist die Modellgüte beider Modelle als etwa gleich gut einzuschätzen?

Wie bei der Bewertung der Konstrukte ist es auch bei der Beurteilung der Modelle möglich, dass nicht alle herangezogenen Maße ein gleichgerichtetes Ergebnis liefern. Dies bedeutet jedoch nicht, dass die Güte des Modells grundsätzlich nicht akzeptabel ist. Vielmehr geben mehrere Indikatoren bzw. Gütekriterien eine breitere Basis zur Beurteilung (vgl. dazu WEIBER und MÜHLHAUS, 2010). Zur Veranschaulichung sind das Modell und die Ergebnisse zu Hypothese 1 ausführlich dargestellt. Für die übrigen Modelle werden nur die Kernergebnisse dargestellt.

Die Signifikanz eines Zusammenhangs zwischen zwei Konstrukten kann analog zur Vorgehensweise bei den Konstrukten anhand der Critical Ratios sowie der P-Werte beurteilt werden. Wie in Abbildung 3.1 zu sehen ist, besteht – der Hypothese entsprechend – ein signifikanter Einfluss von dem Konstrukt „Psychische Arbeitsbedingungen" auf das Konstrukt „Psychovegetative Beschwerden", der für beide Erhebungen als relativ hoch einzustufen ist.





- beranf_01 'Arbeiten unter Termin-/Leistungsdruck'
- beranf_06 'Störungen/Unterbrechungen'
- beranf_08 'Dinge werden verlangt, die man nicht beherrscht'
- beranf_09 'verschiedene Vorgänge gleichzeitig im Auge behalten'
- beranf_12 'bis an die Grenzen seiner Leistungsfähigkeit gehen'
- gesbesch_14 'Schlafstörungen'
- gesbesch_15 'Müdigkeit'
- gesbesch_16 'Verdauungsbeschwerden'
- gesbesch_18 'Nervosität'
- gesbesch_19 'Niedergeschlagenheit'

Abb. 3.1 Modelle (2005/1998) für Hypothese 1

Die „Squared multiple Correlations“ (entsprechend dem R^2 von Regressionsmodellen) liegen bei 0,19 bzw. 0,18 – d. h., dass in beiden Erhebungen knapp 20 % der Varianz des Konstrukts „Psychovegetative Beschwerden” durch das Konstrukt „Psychische Arbeitsbedingungen” erklärt wird. Die Modellgüte ist anhand der definierten Gütemaße (siehe Tabelle 3.3) durchweg als sehr gut zu betrachten. Diese Analysen lassen somit auf Konstruktvalidität schließen.

Tab. 3.3 Zusammenhangs- und Gütemaße (2005/1998) für Hypothese 1

Hypothese 1: 2005						
Zusammenhangsmaße				Gütemaße		
Zusammenhang	Regression Weights	Critical Ratio	P-Wert	SRMR	RMSEA	CFI
Psych. Bed. --> Psych.-veg.-Beschw.	0,433	31,31	***	0,035	0,047	0,949

Hypothese 1: 1998						
Zusammenhangsmaße				Gütemaße		
Zusammenhang	Regression Weights	Critical Ratio	P-Wert	SRMR	RMSEA	CFI
Psych. Bed. --> Psych.-veg.-Beschw.	0,419	37,92	***	0,028	0,047	0,945

Die Strukturgleichungsmodelle unterscheiden sich für beide Erhebungen nur marginal und messen damit mehr oder weniger dieselben Sachverhalte. Die Werte des SRMR und des RMSEA bestätigen dies auch in der Mehrgruppenanalyse (siehe Tabelle 3.4), während die Werte des CFI mit zunehmender Restringierung der Modelle doch deutlich schlechter werden.[8]

Tab. 3.4 Ergebnisse der Mehrgruppenanalyse für Hypothese 1

Mehrgruppenalayse: Hypothese 1			
Modell	**SRMR**	**RMSEA**	**CFI**
Unconstrained	0,028	0,033	0,947
Measurement weights	0,033	0,036	0,932
Measurement intercepts	0,035	0,054	0,824
Structural weights	0,042	0,055	0,811
Structural covariances	0,040	0,056	0,806
Structural residuals	0,059	0,061	0,766
Measurement residuals	0,048	0,081	0,543

In allen vier Modellen wurden die formulierten Hypothesen nicht widerlegt, die Werte der Gütekriterien gaben keinen Anlass zur Falsifizierung der Hypothesen. Allerdings muss festgehalten werden, dass für Hypothese 3 nur für die Erhebung 2005 bedeutsame Effekte festgestellt werden konnten. Ebenso war der Einfluss von „Psychischen Arbeitsbedingungen“ auf „Muskel-/Skelettbeschwerden“ in Hypothese 2 eher unbedeutsam. Mit diesen Einschränkungen konnte Konstruktvalidität nachgewiesen werden.

Für alle Strukturgleichungsmodelle hat sich gezeigt, dass in den Erhebungen 2005 und 1998 gleiche bzw. sehr ähnliche Sachverhalte gemessen wurden.

---

[8] Der Wert des CFI ist bei den fixierten Faktorladungen noch sehr gut und fällt erst ab den restringierten Intercepts deutlich ab. Allerdings ist dies nicht sonderlich überraschend, da die Skala für die Erhebung 1998 zusammengefasst wurde. Dies kann durchaus dazu führen, dass sich die Höhe der Intercepts beider Studien voneinander unterscheiden.

Tab. 3.5 Übersicht zu den Modellbewertungen

Hypothese	Modellgüte		Einfluss der erklärenden Variable(n)		Mehrgruppen-vergleich	Hinweis auf Validität	
	2005	1998	2005	1998		2005	1998
1	++	++	+ / ++ *	+ / ++ *	+	+	+
2	+	+	+ / ++ * + / ++ * o / + *	+ / ++ * + / ++ * o / + *	+	+	+
3	+	+	+ * + *	o / + * o / + *	+	+	o / +
4	++	++	+ * + *	+ * + *	++	+	+

Legende: + : gute Bewertung; – :negative Bewertung; o :neutral/nicht eindeutig; * : Einfluss ist zum 1 %-Niveau signifikant.

Aufgrund der Forschungslage ist zu erwarten, dass die Validität bei mehr Skalenpunkten ansteigt und die Reliabilität bei 5 bis 7 Skalenpunkten am höchsten ist (vgl. FAULBAUM et al., 2009). Daher wurde ein weiterer Vergleich der Messung mit der 1998 verwendeten 5er-Skala mit einer 4er-Skala vorgenommen. Es sollte überprüft werden, ob die Verwendung einer 5er-Skala im Vergleich zu einer 4er-Skala Vorteile für die Modellgüte mit sich bringt. Aus diesem Grund wurden die Konstrukte „Psychische Arbeitsbedingungen“ und „Physische Arbeitsbedingungen“ mit den Daten aus der Erhebung 1998 einmal mit der ursprünglich verwendeten 5er-Skala berechnet und einmal mit der homogenisierten 4er-Skala, bei der die oberen beiden Kategorien “praktisch immer” und “häufig” zusammengefasst wurden. Anschließend wurden die Gütewerte der Modelle verglichen. Die Höhe der Faktorladungen war für beide Skalen ähnlich. Der Vergleich der verwendeten Gütekriterien SRMR, CFI und RMSEA zeigte sowohl für das Konstrukt „Psychische Arbeitsbedingungen“ als auch für das Konstrukt „Physische Arbeitsbedingungen“ beinahe identische Gütewerte auf. Diese Ergebnisse deuten darauf hin, dass die Konstrukte mit einer 4er-Skala und mit einer 5er-Skala praktisch identisch gemessen werden.

3.2 Erfassung berufsgruppenspezifischer Arbeitsbedingungen

Ergänzend zu den geschilderten Analysen zur Konstruktvalidität wurden im Sinne der Methode der bekannten Gruppen Analysen zur Kriteriumsvalidität angestellt. Gruppen, von denen erwartet wird bzw. von denen bekannt ist, dass sie Unterschiede bezüglich des untersuchten Merkmales aufweisen, stellen dabei das Kriterium dar (vgl. SCHNELL et al., 2005). In diesem Sinne ist davon auszugehen, dass Unterschiede in der beruflichen Tätigkeit bzw. verschiedene Arbeitsplatztypen ein geeignetes Kriterium darstellen, um Skalen zu Arbeitsbedingungen auf ihre Kriteriumsvalidität hin zu überprüfen. Um möglichst trennscharfe Arbeitsplatztypen zu bilden, wurden für die Analyse kleine Berufsgruppen (4-Steller der KldB 1992) in Kombination mit einer groben Gliederung der Wirtschaftszweige verwendet.[9] Basie-

[9] Bei Bildung der Gruppen wurde darauf geachtet, dass sich keine Zellenbesetzung mit N<10 ergibt. DRAGANO (2007) zufolge, der zu diesem Zweck eine Simulationsstudie durchgeführt hat, reicht eine Zellenbesetzung von mindestens 10 Fällen aus, um bei JEM (Job Exposure Matrizen) auf

rend auf explorativen Faktorenanalysen wurden durch Mittelwertbildung Indizes zu körperlich schwerer Arbeit, zu psychischen Arbeitsanforderungen, zur Autonomie bei der Tätigkeit sowie zu Schicht- und Nachtarbeit gebildet. Bei den physischen Belastungen wurden Umgebungsbelastungen ausgeklammert und es wurde nur körperlich schwere Arbeit als spezieller Aspekt behandelt.

Aus der Kreuzung der Arbeitsplatztypen mit ihrem Mittelwert auf diesen Indizes ergibt sich eine Tabelle mit der Arbeitsbelastung der Arbeitsplatztypen. Dabei ergeben sich zwischen den beiden Wellen durchaus Unterschiede in der mittleren Arbeitsbelastung, was aufgrund des zeitlichen Abstands der Messungen und der methodischen Unterschiede nicht verwunderlich ist. Die Korrelation der beiden Messwertreihen fällt bei psychischen Anforderungen aber dennoch hoch und bei körperlichen und Umgebungsbelastungen sowie bei Autonomie und bei Schicht- und Nachtarbeit sehr hoch aus.[10]

Auch die durch die Arbeitsplatztypen und weitere Kontrollvariablen[11] erklärte Varianz in den Arbeitsbelastungen fällt in beiden Erhebungen vergleichbar aus. Körperliche Anforderungen werden 1998 etwas besser durch die Arbeitsplatztypen und Kontrollvariablen erklärt als 2005. Bei psychischen Anforderungen sowie Schicht- und Nachtarbeit ist es umgekehrt. Hinsichtlich der Autonomie bei der Arbeitsdurchführung ist kaum ein Unterschied festzustellen. Somit lässt sich eine allgemeine Tendenz zu einer höheren oder niedrigeren erklärten Varianz in einer der beiden Erhebungen feststellen. Auf Basis der angestellten Analysen sind daher beide Messungen als vergleichbar anzusehen.

4 Explorative Faktorenanalysen und Indexkonstruktionen: BIBB/BAuA 2005 und andere Datenquellen im Vergleich

Da bei statistischen Analysen aufgrund der Vielzahl von relevanten Arbeitsbedingungen einzelne Items oft zu Indizes bzw. zu Dimensionen von Arbeitsbedingungen verdichtet werden, wurde geprüft, welche Items aus verschiedenen Erhebungen sich dafür eignen und welche Dimensionen von Arbeitsbedingungen sich dabei ergeben. Dabei wurden neben der BIBB/BAuA-Erhebung von 2005 vergleichend auch die Erhebung zum DGB-Index Gute Arbeit 2007 und der EWCS 2005 analysiert.[12] Die Analyse mit der Erwerbstätigenbefragung 2005 soll dabei auf Items aufmerksam machen, die für Indexkonstruktionen wenig geeignet sind, während der

---

Basis der BIBB/IAB-Surveys stabile Ergebnisse zu erzielen. Die Analyse wurde auf Männer beschränkt.

[10] Produkt-Moment-Korrelation nach Pearson: Psychische Anforderungen: r=0,81, Körperlich schwere Arbeit: r= 0,99, Geringe Autonomie: r= 0,91, Schicht- und Nachtarbeit: r= 0,96. Mit einer ähnlichen Vorgehensweise zeigt sich auch eine hohe Übereinstimmung der Erwerbstätigenbefragung 2005 mit der Erhebung zum DGB-Index 2007 (vgl. TRISCHLER und HOLLER, 2011).

[11] Als Kontrollvariablen wurden verwendet: Überstunden (ja/nein), Alter, Arbeitszeit, Betriebsgröße.

[12] Die Untersuchungspopulationen der verwendeten Erhebungen wurden aneinander angeglichen. Die Analysen erfolgten ungewichtet.

Vergleich zu anderen Erhebungen Hinweise darauf ergeben kann, ob bestimmte Dimensionen oder Items optimiert werden können.

Neben inhaltlichen Überlegungen wurden für die Itemselektion und die Bestimmung der Faktorenanzahl in explorativen Faktorenanalysen gängige Kriterien verwendet.[13] Dennoch könnten die so ermittelten Skalen zum Teil noch weiter optimiert werden. Zur Beurteilung der Möglichkeit der Indexbildung mit den ermittelten Dimensionen wurden nach der Empfehlung von BÜHNER (2006) Cronbachs Alpha, die Präzision von Alpha sowie die mittlere Interitemkorrelation verwendet, deren Interpretation in Tabelle 4.1 dargestellt ist.

Tab. 4.1 Überblick über die Gütemaße zur Beurteilung der Reliabilität und Homogenität der Komponenten sowie die verwendeten Schwellenwerte

Maßzahl	Schwellenwerte	Symbol
Cronbachs Alpha	> 0,8: gut	++
	0,7-0,8: akzeptabel	+
	< 0,7: nicht ausreichend	-
	< 0,5	- -
Präzision von Alpha	< 0,01: gut	+
	> 0,01: nicht ausreichend	-
Mittlere Interitemkorrelation	> 0,6: zu hoch	-
	0,2 - 0,6: akzeptabel[14]	+
	<0,2: unakzeptabel	- -

Im Folgenden ist lediglich die Zusammenfassung der Ergebnisse dargestellt.[15] Wie sich zeigt, scheint bei der Präzision von Alpha die Forderung nach Werten unter 0,01 im vorliegenden Kontext kaum erfüllbar zu sein. Zudem hat die Anzahl der Items einen Einfluss auf die Präzision von Alpha (vgl. CORTINA, 1993). Insgesamt lässt sich festhalten, dass Indexkonstruktionen mit Daten der Erwerbstätigenbefragung von 2005 in der Regel vergleichbare Reliabilitäts- und Homogenitätskennwerte erzielen wie entsprechende Indizes aus anderen arbeitswissenschaftlichen Befragungen. Das etwas schlechtere Abschneiden des Index zur Autonomie stellt eine Ausnahme dar und kann inhaltlich begründet sein.

---

[13] Faktorladungen >0,3; MSA-Koeffizienten<0,5 (vgl. BÜHNER, 2006; vgl. BACKHAUS et al., 2000). Kommunalitäten unter 0,2 wurden als zu niedrig angesehen. Die Anzahl der Faktoren wurde nach dem MAP-Test von Velicer (vgl. O'CONNOR, 2000), dem Vergleich mit anderen Faktorlösungen und durch inhaltliche Überlegungen bestimmt.

[14] Prinzipiell sind im Sinne der konvergenten Validität zwar hohe Korrelationen zwischen den Items einer Skala erwünscht, BRIGGS und CHEEK (1986) weisen jedoch zu Recht darauf hin, dass Interitemkorrelationen nicht wie manchmal angenommen so hoch wie möglich sein sollten, da ein Index weder aus zu heterogenen noch aus zu ähnlichen Items bestehen sollte. Die genannte Obergrenze von 0,4 ist dabei nicht bindend, insbesondere bei Indizes mit wenigen Items (vgl. BÜHNER, 2006) oder einer geringen inhaltlichen Bandbreite. Entscheidender erscheint die genannte Untergrenze von 0,2.

[15] Die im Folgenden dargestellten Werte der Präzision von Alpha und der mittleren Interitemkorrelation beruhen auf nach Fisher's Z transformierten Korrelationskoeffizienten.

Tab. 4.2 Reliabilitäts- und Homogenitätswerte der vergleichbaren Indizes aus BIBB/BAuA 2005, DGB-Index 2007 und dem EWCS 2005

	BIBB/BAuA 2005	DGB-Index 2007	EWCS 2005
Körperliche und Umgebungsbelastungen			
Anzahl der Items	11	3	14
Mittlere Interitemkorrelation	+	+	+
Cronbachs Alpha	+ +	-	+ +
Präzision von Alpha	-	-	-
Psychische Arbeitsanforderungen			
Anzahl der Items	13	3	7
Mittlere Interitemkorrelation	+	+	+
Cronbachs Alpha	+ +	+	+
Präzision von Alpha	+	-	-
Autonomie/Einfluss- und Gestaltungsmöglichkeiten			
Anzahl der Items	5	4	5
Mittlere Interitemkorrelation	- -	+	+
Cronbachs Alpha	-	+	+
Präzision von Alpha	-	-	-
Wochenend-/Schichtarbeit			
Anzahl der Items	5		5
Mittlere Interitemkorrelation	+		+
Cronbachs Alpha	-		+
Präzision von Alpha	-		-
Arbeitsklima/Soziale Unterstützung bei der Arbeit			
Anzahl der Items	6	3	4
Mittlere Interitemkorrelation	+	+	+
Cronbachs Alpha	+	+	+
Präzision von Alpha	-	-	-

5 Zusammenfassende Bewertung und Ableitung von Designempfehlungen

Auf Basis der durchgeführten Analysen kann festgehalten werden, dass sowohl ein mündlich-persönlicher als auch ein telefonischer Erhebungsmodus für die Durchführung der Erwerbstätigenbefragung geeignet ist.

Zeitverlaufsanalysen sind auf der Ebene der durchgeführten Zusammenhangsanalysen über die beiden Modi hinweg zulässig. Schwieriger gestaltet sich aufgrund der Umstellung von einer persönlichen auf eine telefonische Befragung und der damit zusammenhängenden Änderung einiger Antwortmöglichkeiten bzw. Skalenformate die Bildung von Zeitreihen auf der Ebene von Kategorien einzelner Items. Zu Methodeneffekten lassen sich auf dieser Basis keine gesicherten Aussagen treffen. Die

Hinweise auf eine häufigere Nennung der positiven Extremkategorie bei Zufriedenheitsabfragen müssten in einem Methodenexperiment geprüft werden.

Aus den durchgeführten Analysen ergeben sich darüber hinaus Hinweise für die künftige Fragebogenentwicklung.

- 4er- versus 5er-Skala

Die vergleichenden Analysen, die mit dem Konstrukt ‚Psychische Arbeitsbedingungen' durchgeführt wurden, zeigen, dass sich bei der Verwendung einer 4er-Skala und einer 5er-Skala keinerlei nennenswerte Unterschiede hinsichtlich der Gütewerte ergeben. Die Umstellung von einer 5er- auf eine 4er-Skala, die für die Erhebung 2005 vorgenommen wurde, erfolgte aufgrund des telefonischen Erhebungsmodus und kann aus messtheoretischer Sicht ohne Bedenken wieder eingesetzt werden.

- Konstrukt[16] „Psychovegetative Beschwerden"

Ein Vergleich der Gütewerte der vier Varianten des Konstrukts ‚Psychovegetative Beschwerden' zeigt, dass die 2005 erstmals erhobenen Items „Burn-out" und „Depression" nicht zu einer Verbesserung dieses Konstrukts beitragen. Hintergrund dafür könnte die Tatsache sein, dass es sich bei den beiden neuen Items um Krankheitsbilder handelt, die sich u. a. in Krankheitssymptomen äußern, die unter dem Konstrukt ‚Psychovegetative Beschwerden' gefasst sind. Für die künftige Fragebogenentwicklung bedeutet dies, dass beide neuen Items für die Messung des Konstrukts ‚Psychovegetative Beschwerden' verzichtbar sind. Ob es unabhängig davon ein Forschungsinteresse an den beiden Items gibt, muss von den Auftrag gebenden Forschern geprüft werden.

- Index[17] zu körperlichen und Umgebungsbelastungen

Körperliche und Umgebungsbelastungen sind ausreichend erfasst und eine Indexkonstruktion ist möglich. Es könnten evtl. Items eingespart werden, die im Rahmen der Indexkonstruktion eher redundant sind wie z. B. die Items „Wie häufig arbeiten Sie an einem Platz, an dem geraucht wird?" und „Wie häufig arbeiten Sie mit mikrobiologischen Stoffen?". Die inhaltliche Wichtigkeit hängt dabei aber von den Erkenntnisinteressen der Auftraggeber der Erhebung ab und kann daher an dieser Stelle nicht entschieden werden.

- Index zu psychischen Anforderungen

Verzichtbar wäre aus der Perspektive von Indexkonstruktionen evtl. das aus der Analyse ausgeschlossene Item „Wie häufig können auch kleine Fehler/geringe Unaufmerksamkeit größere finanzielle Verluste verursachen?". Emotionale Belastungen sind nur durch ein Item erfasst, so dass bei Interesse an einem eigenen Konstrukt zu diesem Bereich evtl. Items ergänzt werden könnten.

---

[16] Die Bezeichnung „Konstrukt" weist darauf hin, dass es sich um einen im Rahmen der Strukturgleichungsmodelle bzw. konfirmatorischen Analysen in Kapitel 3 gebildeten Faktor handelt.

[17] Die Bezeichnung „Index" weist darauf hin, dass es sich um einen Index handelt, der auf Basis der explorativen Analysen in Kapitel 4 gebildet wurde.

- Index zur Autonomie

Obwohl die geringere Reliabilität des Autonomie-Index in BIBB/BAuA 2005 im Vergleich zu anderen Erhebungen nicht eindeutig durch andere Inhalte des Index zustande kommt, könnte es hilfreich sein, zusätzliche Items aufzunehmen, die weitere Facetten von Autonomie abdecken. Durch eigene Analysen mit anderen Datensätzen und die Hauptkomponentenanalysen von WIRTZ (2010) konnten einige dafür in Frage kommende Items identifiziert werden.

Literaturverzeichnis

Anand, S. (2008): Satisficing. In: Lavrakas, P. J. (Hrsg.): Encyclopedia of Survey Research Methods, Volume 2, Los Angeles, S. 797-799.

Backhaus, K., Erichson, B., Plinke, W. und Weiber, R. (2000): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Berlin/Heidelberg.

BIBB/IAB (2003): Abschlussbericht zum Forschungsprojekt 1.1.006. Erwerb und Verwertung beruflicher Qualifikationen - BIBB/IAB-Erhebung 1998/99 (3. Wiederholungsuntersuchung). http://www2.bibb.de/tools/fodb/pdf/eb_11006.pdf [letzter Zugriff 25.01.2010].

Brick, M. J. und Lepkowski, J. M. (2008): Multiple Mode and Frame Telephone Surveys. In: Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P., Link, M. und Sangster, R. (Hrsg.): Advances in Telephone Survey Methodology, New York, S. 149-169.

Briggs, S. R. und Cheek, J. M. (1986): The role of factor analysis in the development and evaluation of personality scales. Journal of Personality, 54 (1), S. 106-148.

Bühner, M. (2006): Einführung in die Test- und Fragebogenkonstruktion, München u. a. O.

Christian, L. M, Dillman, D. A. und Smyth, J. D. (2008): The effects of mode and format on answering to scalar questions in telephone and web surveys. In: Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P., Link, M. und Sangster, R. (Hrsg.): Advances in Telephone Survey Methodology, New York, S. 250-274.

Cortina, J. M. (1993): What is Coefficient Alpha? Examination of Theory and Applications. Journal of Applied Psychology, 78 (1), 98-104.

De Leeuw, E. D., Mellenbergh, G. J. und Hox, J. J. (1996): The Influence of Data Collection Method on Structural Models: A Comparison of a Mail, a Telephone, and a Face-to-Face Survey. In: Sociological Methods & Research 24, S. 443-472.

Dragano, N. (2007): Arbeit, Stress und krankheitsbedingte Frührenten. Zusammenhänge aus theoretischer und empirischer Sicht, Wiesbaden.

Faulbaum, F., Prüfer, P. und Rexroth, M. (2009): Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. VS Verlag für Sozialwissenschaften, Wiesbaden.

Häder, M., Kühne, M. und Schlinzig, T. (2009): Mode Effekte bei telefonischen Befragungen über Festnetz und Mobilfunk. In: Weichbold, M., Bacher, J., Wolf, C. (Hrsg.): Umfrageforschung – Herausforderungen und Grenzen. Sonderheft der Österreichischen Zeitschrift für Soziologie 9, S. 45-62.

Hall, A. (2009): Die BIBB/BAuA-Erwerbstätigenbefragung 2006: Methodik und Frageprogramm im Vergleich zur BIBB/IAB-Erhebung 1998. Bundesinstitut für Berufsbildung, Wissenschaftliches Diskussionspapier 107.

Hartmann, J., Bielenski, H. und von Rosenbladt, B. (1999): Erwerb und Verwertung beruflicher Qualifikationen von Erwerbstätigen (BiBB/IAB-Erhebung 1998/99). Gewichtung und Strukturkontrolle der Stichprobe. Infratest Burke Sozialforschung, München.

Hartmann, J. (2006a): BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 – Feldbericht. TNS Infratest Sozialforschung, München.

Hartmann, J. (2006b): BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 – Strukturkontrolle, Steuerung und Gewichtung der Stichprobe. TNS Infratest Sozialforschung, München.

Knapp, T. R. (2008): Validity. In: Lavrakas, P. J. (Hrsg.): Encyclopedia of Survey Research Methods, Volume 2, Los Angeles, S. 938-939.

Krosnick, J. A. (1999): Survey Research, Annual Review of Psychology 50, S. 537-567.

Lamnek, S. und Schäfer, W. J. (1998): Befragungsmethoden reviewed. Sozialwissenschaften und Berufspraxis 21 (2), S. 162-170.

Lass, J, Saris, W. E. und Kaase, M. (1997): Sizes of the different effects: Coverage, mode and nonresponse. In: Saris, W. E. und Kaase, M. (Hrsg.): Eurobarometer: Measurement Instruments for Opinions in Europe. Zuma Nachrichten Spezial Band 2, Mannheim, S. 64-74.

O'Connor, B. P. (2000): SPSS and SAS programs für determining the number of components using parallel analysis and Velicer's MAP test. Behavior Research Methods, Instruments & Computers, 32 (3), 396-402.

Schnell, R., Hill, P. B. und Esser, H. (2005): Methoden der empirischen Sozialforschung, München/Wien.

Scholl, A. (2009): Die Befragung, Konstanz.

Schumann, H. und Presser, S. (1996): Questions & Answers in Attitude Surveys. Experiments on Question Form, Wording, And Context, Thousand Oaks.

Trischler, F. und Holler, M. (2011): Die Befragung zum DGB-Index „Gute Arbeit“ als Datenbasis für die arbeitswissenschaftliche Forschung? Ein Vergleich mit der BIBB/BAuA-Erwerbstätigenbefragung 2005/06. Arbeit 2011 (2).

Weiber, R. und Mühlhaus, D. (2010): Strukturgleichungsmodellierung, Berlin/Heidelberg.

Wirtz, A. K. (2010): Lange Arbeitszeiten. Untersuchungen zu den gesundheitlichen und sozialen Auswirkungen langer Arbeitszeiten. Diss. Universität Oldenburg.

Yan, T. (2008): Nondifferentiation. In: Lavrakas, P. J. (Hrsg.): Encyclopedia of Survey Research Methods, Volume 2, Los Angeles, S. 520-521.