Konstruktion eines änderungssensitiven State-Trait-Instruments zur Messung der Befindlichkeit - Eine explorative Feldstudie

DIPLOMARBEIT

im Rahmen der Diplomprüfung im Studiengang Psychologie
des Fachbereichs Psychologie der Universität Hamburg

Erster Prüfer	:	Prof. Dr. Lothar Buse
Zweiter Prüfer	:	Prof. Dr. Burger Heinze
FB Psychologie Klassifikation	:	731 Testkonstruktion
vorgelegt von	:	Carsten Riepe Schulterblatt 29 20357 Hamburg

Hamburg, im Januar 1994

Am Zustandekommen dieser Arbeit haben sehr viele Menschen auf ganz unterschiedliche Art und Weise mitgewirkt.

Sie sollen nicht unerwähnt bleiben:

Herr Prof. Dr. Lothar Buse regte das Thema an und betreute die Arbeit in allen Phasen ihrer mehr als zweieinhalbjährigen Entstehung mit großem Engagement und noch größerer Geduld,

Dutzende von Kommilitoninnen und Kommilitonen erklärten sich bereit, als Versuchspersonen an der Felduntersuchung teilzunehmen oder mir bei der Datenaufbereitung bzw. -erfassung zu helfen,

die Geschäftsleitung der Firma Partner Research Marktforschungs GmbH in Hamburg ermöglichte mir als ihrem Mitarbeiter das Anfertigen von mehreren tausend Photokopien,

Ricarda übernahm das Korrekturlesen des Manuskripts.

Ihnen allen sei an dieser Stelle herzlich gedankt!

Ein ganz besonderes Wort des Dankes geht an die 31 Versuchspersonen aus der Stichprobe der Nicht-Psychologiestudenten. Sie alle stammen aus meinem Freundes-, Bekannten- oder Kollegenkreis und erklärten sich bereit, an der einwöchigen Felduntersuchung ohne jede wirkliche Gegenleistung teilzunehmen!

Inhalt

"Ursprünglich verhielt sich der Mensch natürlich. Als Kind spielte er mit anderen Kindern, ging zur Schule und lernte fürs Leben, war fröhlich und traurig. Als Erwachsener ging er wichtigen Geschäften nach, hatte Umgang mit seinesgleichen, erzog seine Kinder. Er tat, was er als notwendig befand, wie er es als richtig erachtete, wenn es ihm sinnvoll erschien, und kein Fremder hatte ihm dreinzureden.

Dann entdeckte ihn die Wissenschaft. Jetzt lernte er ab und zu sinnlose Silben, unterschied Helligkeitsstufen, beantwortete Fragen, die er sich selber nie gestellt hätte, interagierte mit Menschen, die er nicht kannte und die ihn nicht interessierten. Das hatte fast nichts damit zu tun, was er sonst tat.

Dann entdeckte die Wissenschaft, daß man das natürliche Verhalten auch untersuchen kann, und sie nannte es Feldforschung. Und man entwickelte Methoden und Modelle, um dieses Verhalten und dessen Bedingungen zu studieren. Man verließ das Labor und begann, das Leben dort zu beobachten, wo es stattfindet, wie es stattfindet."

(PATRY, 1982, S. 5)

1. Überblick

In der vorliegenden Arbeit geht es um die Entwicklung eines Meßinstruments für psychologische Merkmale. Damit soll der nahezu unüberschaubaren Vielfalt solcher Datenerhebungsverfahren (DORSCH, 1987, verzeichnet alleine ca. 700 Testnachweise) ein weiteres hinzugefügt werden. Dies bedarf einer Erläuterung.

1.1. Explikation des Themas

Die Themenstellung sieht die Entwicklung eines Verfahrens zur Messung der Befindlichkeit, d.h. der Stimmung, vor. Dieses Meßinstrument soll sich sowohl zur Erfassung des momentanen Zustands (State) eines Individuums als auch zur Ermittlung zeit- und situationsinvarianter interindividueller Differenzen in der Stimmungsdisposition (Eigenschaften oder Traits) eignen. Es soll sensitiv sein für Zustandsänderungen von einem Zeitpunkt zum anderen, d.h. Veränderungen der momentanen Stimmung sollen mit hoher Genauigkeit erfaßbar gemacht werden. Das Attribut "änderungssensitiv" bezieht sich somit nicht auf die Messung von Veränderungen auf der Eigenschaftsebene.

Stimmungen als psychologische Konstrukte eignen sich ganz besonders für die Entwicklung eines so skizzierten Instruments, denn sie "werden im allgemeinen als vorübergehende Zustände der subjektiven Befindlichkeit angesehen, die starken Veränderungen über die Zeit hinweg unterworfen sind" (BOHNER, HORMUTH & SCHWARZ, 1991, S. 135).

Das Thema verlangt weiterhin die Durchführung einer Feldstudie, die explorativen Charakter haben soll. Der empirische Teil dieser Arbeit wurde daher im alltäglichen Lebensraum der beteiligten Versuchspersonen (Vpn) durchgeführt und nicht unter Laborbedingungen. Die Untersuchung hat Erkundungscharakter und ist damit explorativ in dem Sinne, daß in ihr erstmals ein Meßverfahren angewandt wird, das auf einem testtheoretischen Modell beruht, das kürzlich von BUSE & PAWLIK (1991) als Erweiterung der Klassischen Testtheorie vorgeschlagen wurde. Außerdem erfolgt die Generierung der Daten mit Hilfe eines randomisierten Zeitstichprobenplans unter Einsatz eines elektronischen Signalgebers, einer Datenerhebungstechnologie also, deren Einsatz in der sozialwissenschaftlichen Forschung keineswegs alltägliche Routine ist.

1.2. Zielsetzung der Arbeit

Da unter den Bedingungen der Anfertigung einer Diplomarbeit selbstverständlich keine komplette Testkonstruktion bis hin zu einer normierten, verkaufsfähigen Endversion geleistet werden kann, wird die Bearbeitung des Themas auf die Durchführung eines Pretests und die Diskussion weiterer notwendiger Konstruktionsschritte eingeschränkt. Konkret sollen dabei drei Ziele erreicht werden:

1. Dokumentation empirisch ermittelter Parameter für das Modell von BUSE & PAWLIK (1991),

2. ausgehend von diesen Parametern eine Itemanalyse zum Zwecke der Skalenkonstruktion unter

zusätzlicher Verwendung faktorenanalytisch gewonnener Informationen bei gleichzeitiger erster Abschätzung der Testgütekriterien sowie

3. eine Analyse der Untersuchungsmethode und der Einsatzmöglichkeiten eines solchen Meßinstruments.

Es gibt bereits eine Reihe von Erhebungsverfahren, die Befindlichkeitsdimensionen bzw. Stimmungen erfassen sollen. Beispielhaft seien genannt:

1. das State-Trait-Angstinventar (STAI) (LAUX, GLANZMANN, SCHAFFNER & SPIELBERGER, 1981), das mit zwei unterschiedlichen Teilinventaren sowohl State- als auch Trait-Angstwerte liefert und das sowohl bei gesunden Vpn als auch bei Patienten im klinischen Bereich eingesetzt werden kann; es ist außerdem gut geeignet für Längsschnittuntersuchungen und für den Einsatz unter feldähnlichen Bedingungen,

2. die Eigenschaftswörterliste (EWL) (JANKE & DEBUS, 1978) und ebenso

3. die Befindlichkeits-Skala (v. ZERSSEN, 1976), die beide in gesunden und klinischen Populationen zur Beschreibung des momentanen Befindens (State) sowie zum Erfassen von zeitlichen Verläufen eingesetzt werden können,

4. die Kieler Änderungssensitive Symptomliste (KASSL) (ZIELKE, 1979), die zur therapiebegleitenden Diagnostik entwickelt wurde,

5. das Emotionalitätsinventar (EMI) (ULLRICH DE MUYNCK & ULLRICH, 1981), das im klinischen Bereich auch zur Veränderungsmessung auf der Traitebene eingesetzt werden kann, und schließlich

6. im psychiatrischen Bereich die Eppendorfer Stimmungs-Antriebs-Skala (ESTA) (SUPPRIAN, 1976) zur Abbildung periodischer Prozesse bei manisch-depressiven Psychosen.

Innovativ an dem in dieser Arbeit vorgestellten Ansatz ist also weder die Messung von Stimmungen an sich, noch die in Aussicht gestellte Änderungssensitivität der Items und auch nicht der Umstand, daß hier ein Instrument zur Feldpsychodiagnostik entwickelt werden soll. Neu ist vielmehr das zugrundegelegte testtheoretische Modell und als Konsequenz daraus die Notwendigkeit, dieselben Items zu ein und demselben Meßzeitpunkt doppelt vorzugeben (vgl. Kap. 2.1.3.), eine Vorgehensweise, die angesichts der Konzeption des zu entwickelnden Meßinstruments als eines unter Feldbedingungen einsetzbaren Papier- und Bleistiftverfahrens größtmögliche Sorgfalt bei der Planung und der Durchführung erfordert. Außerdem sollen aus denselben Daten sowohl State- als auch Traitwerte abgeleitet werden, was einhergeht mit dem Anspruch, aus den individuellen Biotopen der Vpn ökologisch repräsentative Stichproben von Erhebungsbedingungen zu gewinnen. Beides ist im Rahmen der Konstruktion von psychologischen Meßinstrumenten alles andere als trivial (vgl. PETERMANN, 1992).

Die Arbeit untergliedert sich im wesentlichen in drei Abschnitte: in einen theoretischen Teil, einen empirischen Teil und in einen Diskussionsteil. Im theoretischen Teil werden die testtheoretischen sowie die öko- und die persönlichkeitspsychologischen Grundlagen erläutert, die zur Konstruktion des angestrebten Erhebungsinstruments nötig sind. Die Darstellung der durchgeführten Untersuchung einschließlich der Ergebnisse erfolgt im empirischen Teil. Abschließend erfolgt noch eine Diskussion, in der die Ergebnisse aus dem empirischen Teil kritisch bewertet und analysiert werden.

2. Theoretischer Teil

2.1. Testtheoretische Grundlagen

Das zu entwickelnde Meßinstrument kann in seiner angestrebten Endform als ein "wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung" (LIENERT, 1989, S. 7) aufgefaßt werden und soll somit eine "Messung der diagnostisch relevanten Sachverhalte ermöglichen" (MICHEL & CONRAD, 1982, S. 1). Es erfüllt damit gängige Anforderungen, die an (psychometrische) Tests gestellt werden. Inhaltlich fällt es als Befindlichkeitsskala in den Bereich der Persönlichkeitstests (in Abgrenzung von den Fähigkeitstests) und dort in die Untergruppe der Persönlichkeitsfragebogen (in Abgrenzung von Interessentests, verbalen Ergänzungsverfahren, Formdeuteverfahren u.a.) (RAUCH-FLEISCH, 1989).

Die Items als "kleinste Informationseinheiten" (MICHEL & CONRAD, 1982, S. 2) dieses Meßinstruments werden von Adjektiven zur Kennzeichnung von Befindlichkeitszuständen gebildet, die in die Form einfacher Hauptsätze gekleidet sind (Beispiel: Ich bin vergnügt). Ihnen muß die Vp auf einer Ratingskala entsprechend ihrer momentanen Befindlichkeit mehr oder weniger zustimmen. Da der Vp also gar keine Fragen gestellt werden, handelt es sich streng genommen nicht wirklich um einen Fragebogen, sondern eher um eine Stimmungsliste.

2.1.1. Klassische Testtheorie

Aus den Informationen, die aus dem Ankreuzverhalten der Vp auf Itemebene gewonnen werden, werden nach bestimmten Gesichtspunkten Testwerte gebildet, aus denen dann wiederum auf die Ausprägung eines Persönlichkeitsmerkmals bei der untersuchten Vp geschlossen wird.

2.1.1.1. Grundlagen

Die Klassische Testtheorie (KT) (AMELANG & BARTUSSEK, 1981, FISCHER, 1974, GULLIKSEN, 1950, LIENERT, 1989, MICHEL & CONRAD, 1982, MOOSBRUGGER, 1992, MURPHY & DAVIDSHOFER, 1991, RAUCHFLEISCH, 1989, SCHELTEN, 1980, SINGH, 1986) ist eine Meßfehlertheorie. Sie unterscheidet zwischen dem empirisch ermittelten Testwert x im Test t bei einer Vp v und dem wahren Wert w im selben Test bei derselben Vp, den diese erhalten würde, wenn es möglich wäre, den Testwert bei einmaliger Testung ohne jeglichen Meßfehler zu erheben. Der Fehlerwert e im Test t bei einer Vp v ergibt sich aus der Differenz zwischen Testwert und wahrem Wert:

$$(1)\quad e_{tv} = x_{tv} - w_{tv}$$

oder anders:

$$(2)\quad x_{tv} = w_{tv} + e_{tv}$$

Der Meßfehler soll eine Zufallsvariable sein, die bei jeder konkreten Realisation des Testwertes bei einer beliebigen Vp in einer beliebigen Testsituation in völlig unsystematischer Weise zu einer Abweichung des gemessenen Testwertes vom wahren Wert der Vp führt und sich dabei mal in Richtung einer Erhöhung, mal in Richtung einer Verringerung des Testwertes auswirkt. Er soll für jede Vp die gleiche Varianz s^2_e und den Erwartungswert null haben. Das bedeutet, daß die Summe aller Fehlerwerte bei unendlich vielen Testungen mit demselben Test unter gleichen Bedingungen an derselben Vp gleich null ist:

$$(3)\quad \sum_{t=1}^{\infty} e_{tv} = 0$$

Gleiches gilt bei einmaliger Testung an unendlich vielen Vpn:

$$(4) \quad \sum_{v=1}^{\infty} e_{tv} = 0$$

Dies bezieht sich nur auf zufällig variierende Fehlerwerte, nicht auf systematische, d.h. durch überzufällig in eine Richtung wirkende Fehleranteile, wie sie z.B. durch Störeinflüsse verschiedenster Art zustandekommen können. Diese würden sich nicht zu null herausmitteln, sondern zu einer systematischen Meßwerterhöhung oder -verringerung führen, die die KT fälschlicherweise als Teil des wahren Wertes identifizieren würde.

In testdiagnostischen Anwendungen wird ein Herausmitteln des Meßfehlers durch Vorgabe von verschiedenen, aber vergleichbaren Items, die alle dasselbe Konstrukt messen sollen, angestrebt, so daß der Testwert einer Vp, der sich durch Aufsummierung der Werte der einzelnen Items errechnet, eine fehlerreduzierte Annäherung an den wahren Wert dieser Vp darstellt.

Der meßfehlerfreie sogenannte wahre Wert einer Vp darf nicht mißverstanden werden als ihre wirkliche, echte Ausprägung in dem jeweiligen psychologischen Merkmal, denn die hängt auch von der der Messung zugrundeliegenden Persönlichkeitstheorie und ihrer Operationalisierung in dem Test bzw. von der Güte der Testkonstruktion ab. Ein völlig invalider Test kann auch "wahre" Werte produzieren, sie haben dann nur nichts mit dem angeblich zu messenden Persönlichkeitskonstrukt zu tun.

Aus der unsystematischen Variabilität des Meßfehlers e_t folgt, daß er weder mit den wahren Werten im Test t noch mit den wahren Werten in einem anderen Test u noch mit den Fehlerwerten in diesem Test u korreliert ist:

$$(5) \quad r_{wtet} = 0$$

$$(6) \quad r_{wuet} = 0$$

$$(7) \quad r_{eteu} = 0$$

Da der Mittelwert einer Summe gleich der Summe der Mittelwerte seiner Summanden ist, muß gelten:

$$(8) \quad \bar{x}_{tv} = \bar{w}_{tv} + \bar{e}_{tv}$$

Aus (3) und (4) folgt, daß

$$(9) \quad \bar{e}_{tv} = 0$$

und damit gilt:

$$(10) \quad \bar{x}_{tv} = \bar{w}_{tv}$$

Der mittlere Testwert einer (unendlich großen) Stichprobe von Vpn entspricht also dem mittleren wahren Wert der Vpn und ist damit meßfehlerfrei (bzw. bei in praxi gezogenen Stichproben mit endlich großer Zahl der Vpn: meßfehlerreduziert).

Zerlegt man die Varianz einer Summe (hier die Testwertevarianz) in einzelne Summanden, so muß die gesamte Varianz-Kovarianzmatrix zwischen den einzelnen Summanden aufaddiert werden. Das bedeutet für die Varianz der gemessenen Testwerte s^2_{xt}:

$$(11) \quad s^2_{xt} = s^2_{wt} + s^2_{et} + 2\, cov_{wtet}$$

Weil

$$(12)\quad r_{wtet} = \frac{cov_{wtet}}{s_{wt} * s_{et}} \qquad \text{(BORTZ, 1989, S. 251)}$$

gilt

$$(13)\quad cov_{wtet} = r_{wtet} * s_{wt} * s_{et} \,.$$

Aus (5) folgt, daß die zweifache Summe der Kovarianzen null sein muß und es bleibt:

$$(14)\quad s^2_{xt} = s^2_{wt} + s^2_{et}$$

Letzteres gilt ganz allgemein für alle voneinander unabhängigen Variablen (vgl. AMELANG & BARTUSSEK, 1981). Bei der Varianzzerlegung bleibt die Fehlervarianz also erhalten und trägt zu einer Erhöhung der Testwertevarianz gegenüber der Varianz der wahren Werte bei.

Für die Kovarianzen zwischen den Testwerten in zwei verschiedenen Tests t und u gilt, daß die Kovarianz einer Summe gleich der Summe der Kovarianzen ihrer Summanden ist, also:

$$(15)\quad cov_{xtxu} = cov_{wtwu} + cov_{wteu} + cov_{wuet} + cov_{eteu}$$

Da nach (6) und (7) die wahren Werte mit den Fehlerwerten und die Fehlerwerte untereinander zu null korrelieren, müssen entsprechend (13) die letzten drei Summanden alle gleich null werden, d.h. die Kovarianz der Testwerte aus zwei verschiedenen Tests ist gleich der Kovarianz der wahren Werte und damit wie die mittleren Testwerte meßfehlerfrei:

(16) $cov_{xtxu} = cov_{wtwu}$

Entsprechend (12) läßt sich die Korrelation der Werte aus zwei Tests schreiben als:

$$(17) \quad r_{xtxu} = \frac{cov_{xtxu}}{s_{xt} * s_{xu}}$$

Aus (16) folgt:

$$(18) \quad r_{xtxu} = \frac{cov_{wtwu}}{s_{xt} * s_{xu}}$$

Andererseits ist

$$(19) \quad r_{wtwu} = \frac{cov_{wtwu}}{s_{wt} * s_{wu}} \quad .$$

Aus (14) folgt, daß

(20) $s_{xt} * s_{xu}$

stets größer als

(21) $s_{wt} * s_{wu}$

sein muß, was bedeutet, daß

$$(22)\quad \frac{cov_{wtwu}}{s_{xt} * s_{xu}} < \frac{cov_{wtwu}}{s_{wt} * s_{wu}} .$$

Die empirische Korrelation zwischen zwei Testwertereihen unterschätzt also den statistischen Zusammenhang zwischen den wahren Werten dieser Tests.

2.1.1.2. Testgütekriterien

Ein nach den Gesichtspunkten der KT konstruiertes Meßinstrument muß bestimmten Anforderungen (Gütekriterien) genügen. Es muß objektiv, reliabel und valide sein. Daneben sollten die Gesichtspunkte der Normierung, der Vergleichbarkeit, der Ökonomie und der Nützlichkeit hinlänglich berücksichtigt werden (LIENERT, 1989).

Ein Test ist nach LIENERT (1989, S. 13f) in dem Maße objektiv, "in dem die Ergebnisse eines Testes unabhängig vom Untersucher sind" (ebd.). Diese Unabhängigkeit kann in drei verschiedenen Stadien einer Testanwendung verletzt werden: während der Durchführung, während der Auswertung oder bei der Interpretation der Ergebnisse. Vielfältige Störquellen während der Phase der Testvorgabe (Durchführung) können zu einer Verzerrung der Testwerte führen. Dies betrifft insbesondere ungenaue Instruktionen für den Versuchsleiter, die zu unerwünschten Variationen in der Anleitung der Vpn oder allgemein zu unkontrollierten sozialen Interaktionen zwischen Versuchsleiter und den Vpn führen können, durch die wiederum die Testergebnisse der Vpn beeinflußt werden können. Aber auch Unterschiede in der sonstigen Gestaltung der Testsituation können einen Einfluß auf die Testwerte ausüben (z.B. Lärm, Beleuchtung, Raumtemperatur, Tageszeit, Müdigkeit der Vpn usw.). Die Auswertungsobjektivität bezieht sich auf die Eindeutigkeit der Regeln, nach denen das registrierte Testverhalten der Vpn vom Auswerter zu einem wie auch immer gearteten, weiterverwendbaren Testwert aufbereitet wird, d.h. auf die Unabhängigkeit des Auswertungsergebnisses von der Person des Auswerters. Am höchsten ist die Auswertungsobjektivität in den Fällen, in denen das Testverhalten der Vpn darin besteht, Kreuze oder Kringel auf dem Papier zu machen, die eindeutig ausgezählt oder bestimmten Antwortkategorien zugeordnet werden können. Die Auswertungsobjektivität ist geringer bei der freien Beantwortung von Items und sehr niedrig bei projektiven Tests. Als Interpretationsobjektivität wird das Ausmaß der Unabhängigkeit der Interpretation der Testergebnisse (Schlußfolgerungen) von der interpretierenden Person bezeichnet. Sie ist am höchsten, wenn die Testauswertung Rohwerte ergibt, die aufgrund bereits existierender (z.B. alters- und geschlechtsspezifischer) Normierung anhand von Tabellen umgerechnet werden können in einen Normwert, der die Stellung der Vpn im Vergleich zu einer definierten Normstichprobe eindeutig bestimmt. Am niedrigsten ist sie in der Regel bei projektiven Tests wie etwa beim Rorschach-Formdeuteverfahren, bei dem sich "eine eigentliche Anleitung zur Verrechnung der Resultate, oder etwa gar eine Art Bestimmungstafel" (RORSCHACH, 1972, S.

117) nicht geben läßt, denn "Erlebnistypus, Merkmale von Verstimmungen, Intelligenzkomponenten, Zahl der Antworten, vorhandene oder mangelnde Willfährigkeit der Versuchsperson, ungefähre Reaktionszeiten usw. - alles ist insgesamt zu überblicken, und es muß bald vom einen, bald vom andern Faktor ausgegangen werden, bis sich ein Gesamtbild ergibt" (ebd.). Um eine hohe Objektivität zu gewährleisten, wird versucht, den gesamten Prozeß der Gewinnung von Testdaten umfassend zu standardisieren, d.h. alle unerwünschten varianzgenerierenden Einflüsse möglichst auszuschalten oder zumindest zu kontrollieren.

Die Reliabilität eines Tests kennzeichnet "den Grad, in dem die beobachteten interindividuellen Unterschiede der Testergebnisse durch tatsächliche psychische Merkmalsunterschiede erklärbar sind" (MICHEL & CONRAD, 1982, S. 37). Sie bezeichnet also den Grad der Genauigkeit, mit dem ein Merkmal gemessen wird, unabhängig davon, ob dieses Merkmal tatsächlich von dem Test erfaßt wird oder nicht. Ein vollkommen reliabler Test wäre so beschaffen, daß er eine bestimmte Vp ohne Meßfehler beschreiben bzw. sie auf einer Meßskala exakt lokalisieren könnte (LIENERT, 1989, S. 14f). Die interindividuelle Testwertevarianz kann sich aus unterschiedlichen Quellen speisen. Nach THORNDIKE (1949, zit. bei MURPHY & DAVIDSHOFER, 1991, S. 74) gibt es sechs Kategorien solcher Varianzquellen: lasting and general, lasting but specific, temporary but general und temporary and specific characteristics of the individual sowie systematic or change factors, die Aspekte der Testsituation betreffen, und schließlich Zufallsfaktoren. Welche dieser Einflüsse als erwünscht bzw. als unerwünscht zu gelten haben, hängt von der Zielrichtung des Tests und von der Fragestellung der Untersuchung ab. Entscheidend ist die Kontrolle der als unerwünscht geltenden Einflüsse im Rahmen der Testobjektivität (s.o.), deren Güte durch Reduktion des Fehleranteils am Testwert unmittelbar zur Reliabilität der Testwerte beiträgt. Eine hohe Objektivität ist also Voraussetzung für eine hohe Reliabilität. Der Reliabilitätskoeffizient R ist definiert als der Anteil der Varianz der wahren Werte s^2_{wt} an der Varianz der Testwerte s^2_{xt}:

$$(23) \quad R = \frac{s^2_{wt}}{s^2_{xt}}$$

Wegen (14) gilt auch

$$(24) \quad R = \frac{s^2_{wt}}{s^2_{wt} + s^2_{et}} .$$

Die Höhe von R hängt also entscheidend von der Varianz der Fehlerwerte s^2_{et} ab: je höher s^2_{et}, desto niedriger R und umgekehrt. Der Reliabilitätskoeffizient kann Werte zwischen eins und null

annehmen. Er wird gleich eins, wenn die Fehlervarianz null ist; er wird gleich null, wenn die Testwertevarianz ausschließlich aus Fehlervarianz besteht. Mit hundert multipliziert gibt er den Prozentsatz an, zu dem die Varianz der Testwerte aus der Varianz der wahren Werte besteht. Der Reliabilitätskoeffizient ist damit ein zentraler Parameter zur Beschreibung der psychometrischen Qualität eines Tests. Die Schwierigkeiten seiner empirischen Bestimmung resultieren aus dem Wesen der Formel (14), die als Zerlegung einer gemessenen Größe in zwei abstrakte Summanden eine Gleichung mit zwei Unbekannten ist und somit keine direkte Bestimmung der Varianz der wahren Werte ermöglicht. Es gibt im wesentlichen vier Methoden, R empirisch zu ermitteln: die Retestmethode, die Paralleltestmethode, die Split-Half-Methode als Spezialfall der Paralleltestmethode und die Bestimmung der inneren Konsistenz als Verallgemeinerung der Split-Half-Methode.

Alle Methoden gehen von einer mehrfachen (mindestens doppelten) Messung desselben psychologischen Merkmals bei denselben Vpn aus. Zentrale Voraussetzung dabei ist die Konstanz des zu messenden Merkmals, d.h. seine Invarianz von einer Messung zur nächsten. Diese in der ursprünglichen Axiomatik der KT nicht enthaltene Annahme wird hier als eine notwendige Zusatzannahme eingeführt. Bei der Retestmethode wird derselbe Test nach Verstreichen eines sinnvoll erscheinenden Zeitintervalls ein zweites Mal vorgegeben. Die Länge dieses Intervalls bestimmt sich aus Vermutungen über die Dauer der Wirksamkeit von Lern-, Erinnerungs-, Übungseffekten u.ä., die aus der Ersttestung resultieren können. Annahme bei der Retestung ist, daß solche Effekte nicht mehr wirksam sind. Die Paralleltestmethode setzt das Vorliegen mindestens zweier äquivalenter Formen desselben Tests voraus. Zwei Testformen sind äquivalent, wenn sie gleiche Validität, gleiche Reliabilität, gleiche Verteilungskennwerte und gleiche Häufigkeitsverteilungen haben. Bei der Split-Half-Methode wird der Test nur ein einziges Mal vorgegeben und anschließend in zwei quasi-parallele (äquivalente) Hälften zerlegt; der ermittelte Reliabilitätskoeffizient wird nach der Formel von SPEARMAN-BROWN (LIENERT, 1989, S. 221) aufgewertet:

$$(24a) \qquad R = \frac{2 * r_{12}}{1 + r_{12}}$$

(Hierbei bezeichnet r_{12} die Korrelation zwischen den beiden Testhälften.) Bei der Bestimmung der inneren Konsistenz wird der Test nicht in zwei Hälften, sondern in so viele Teile zerlegt, wie er Items hat. Die Konsistenzschätzung erfolgt mit Hilfe verschiedener Ableitungsverfahren (LIENERT, 1989, S. 226ff., SINGH, 1986, S. 70ff.). Der später in dieser Arbeit verwendete Konsistenzkoeffizient α geht von folgender Beziehung zwischen den mittleren Kovarianzen und den mittleren Varianzen der k Items (1 ... i j ... k) eines Tests aus:

$$(24b)\quad \alpha = \frac{k * \overline{cov}_{ij} / \bar{s}^2_i}{1 + (k - 1) * \overline{cov}_{ij} / \bar{s}^2_i} \qquad \text{(BROSIUS, 1989, S. 267)}$$

Zur Errechnung des Reliabilitätskoeffizienten R und damit zur Bestimmung der Varianz der wahren Werte wird bei den drei erstgenannten Methoden die Korrelation $r_{xtxt'}$ zwischen den Meßwerten aus den beiden Testvorgaben (erste / zweite Messung, parallele Testformen bzw. beide Hälften) herangezogen (vgl. (17)):

$$(25)\quad r_{xtxt'} = \frac{cov_{xtxt'}}{s_{xt} * s_{xt'}}$$

Aus (16) folgt

$$(26)\quad r_{xtxt'} = \frac{cov_{wtwt'}}{s_{xt} * s_{xt'}} \; .$$

Da zugleich angenommen wird, daß beide Testwertereihen äquivalent sind bzw. daß das Merkmal konstant bleibt (s.o.), was zu gleichen mittleren Testwerten, gleichen mittleren wahren Werten und gleichen Standardabweichungen in beiden Verteilungen führt, und weil die Kovarianz einer Wertereihe mit sich selbst gleich ihrer Varianz ist, gilt:

$$(27)\quad \frac{cov_{wtwt'}}{s_{xt} * s_{xt'}} = \frac{s^2_{wt}}{s^2_{xt}} = R$$

Folglich

$$(28)\quad r_{xtxt'} = R \; .$$

Die Varianz der wahren Werte ergibt sich aus

$$(29) \quad s^2_{wt} = s^2_{xt} * R .$$

Formel (23) läßt sich auch schreiben als

$$(30) \quad R = \frac{s^2_{xt} - s^2_{et}}{s^2_{xt}} = 1 - \frac{s^2_{et}}{s^2_{xt}} .$$

Die Fehlervarianz, d.h. der Anteil an der Testwertevarianz, der zu Lasten der Unreliabilität des Tests geht, errechnet sich aus

$$(31) \quad s^2_{et} = s^2_{xt} * (1 - R) .$$

Wird Gleichung (31) radiziert, ergibt sich der Standardmeßfehler SMF

$$(32) \quad SMF = s_{et} = s_{xt} * \sqrt{1 - R} .$$

Der Standardmeßfehler wird benötigt, um die Grenzen CL (confidential limits) des Vertrauensintervalls zu errechnen. Das Vertrauensintervall ist der Bereich der Testwerteskala, innerhalb dessen der wahre Wert einer getesteten Vp mit einer bestimmten Wahrscheinlichkeit (z.B. 95%) liegt:

$$(33) \quad CL = x_{tv} \pm 1.96 * SMF$$

(In dieser Formel gibt der Faktor 1.96 denjenigen z-Wert einer Standardnormalverteilung an, jenseits dessen bei zweiseitiger Fragestellung 5% aller Meßwerte liegen. Bei Verringerung der Irrtumswahrscheinlichkeit müßten entsprechend andere Werte eingesetzt werden.) Wäre die Reliabilität gleich eins, würde der SMF null sein und der Testwert einer Vp wäre zugleich ihr wahrer Wert.

Die empirisch ermittelte Korrelation r_{xtxu} zwischen zwei beliebigen Tests ist kleiner als die Korrelation ihrer wahren Werte r_{wtwu} (vgl. (18) bis (22)). Soll r_{wtwu} bestimmt werden, so muß eine Korrektur der Korrelation der Testwerte um den Betrag vorgenommen werden, um den sie gegenüber der Korrelation der wahren Werte gemindert ist. Aus (16) und (19) folgt, daß

$$(34) \qquad r_{wtwu} = \frac{cov_{xtxu}}{s_{wt} * s_{wu}} \quad .$$

Die Standardabweichung s_{wt} der wahren Werte im Test t ergibt sich durch Radizieren von Gleichung (29):

$$(35) \qquad s_{wt} = s_{xt} * \sqrt{R_t}$$

Gleiches gilt für den Test u, so daß

$$(36) \qquad r_{wtwu} = \frac{cov_{xtxu}}{s_{xt} * s_{xu} * \sqrt{R_t} * \sqrt{R_u}} \quad .$$

Vereinfachung entsprechend (17) ergibt:

$$(37) \qquad r_{wtwu} = \frac{r_{xtxu}}{\sqrt{(R_t * R_u)}}$$

Die Korrelation der wahren Werte erlangt man also durch Division der Korrelation der Testwerte durch das geometrische Mittel aus den Reliabilitäten beider Verteilungen.

Die Validität eines Tests beschreibt den "Grad der Genauigkeit, mit dem ein Test das mißt, was er messen soll" (MICHEL & CONRAD, 1982, S. 52). Ausgehend von CRONBACH & MEEHL (1955) lassen sich vier Validitätsarten unterscheiden: Vorhersage-, Übereinstimmungs-, Inhalts- und Konstruktvalidität. Die ersten drei dieser Validierungsstrategien versuchen von den Testwerten der Vpn auf das Verhalten dieser Vpn außerhalb der Testsituation zu schließen. Im Falle der Konstruktvalidität wird demgegenüber versucht, auf Eigenschaften, Fähigkeiten usw., mithin auf psychologische Konstrukte, zu schließen.

Bei der Vorhersage- und bei der Übereinstimmungsvalidierung wird mit Hilfe eines Korrelationsschlusses versucht, einen Zusammenhang zwischen dem Testverhalten der Vpn und einem aus dem Alltagsverhalten gewonnenen Kriterium herzustellen. Wird bei dieser kriterienorientierten Validierungsstrategie das Kriterium gleichzeitig mit dem Testverhalten erhoben, so wird dies Übereinstimmungsvalidität genannt. Die Vorhersagevalidität benötigt dagegen ein zukünftiges Kriterium, das durch die Testwerte prognostiziert werden soll. Im einfachsten Falle wird als Validitätskoeffizient ein Korrelationskoeffizient r_{tc} zwischen den Testwerten t und den Kriteriumswerten c errechnet. Der quadrierte Korrelationskoeffizient (Determinationskoeffizient) wird mit hundert multipliziert und gibt dann den Prozentsatz der Varianz der Kriteriumswerte an, der durch die Varianz der Testwerte aufgeklärt wird. Ein solcher Validitätskoeffizient hängt sowohl von der Reliabilität der Testwerte als auch von der der Kriteriumswerte ab. Je höher die Reliabilitäten, desto höher kann der Validitätskoeffizient werden. Sind die Reliabilitäten bekannt, so können Validitätskoeffizienten durch Korrelation der Testwerte mit den Kriteriumswerten (Vorhersagekraft), durch Korrelation der Testwerte mit den wahren Kriteriumswerten (Validität im engeren Sinne) oder durch Korrelation zwischen den wahren Werten im Test und den wahren Kriteriumswerten (Relevanz) bestimmt werden. Später in dieser Arbeit wird eine Validität im engeren Sinne als minderungskorrigierte Korrelation $r_{tc'}$ zwischen den Testwerten und dem Kriterium verwendet, die die Reliabilität R_c des Kriteriums mit einbezieht:

$$(38) \quad r_{tc'} = \frac{r_{tc}}{\sqrt{R_c}} \qquad \text{(SINGH, 1986, S. 92)}$$

Mit Hilfe eines Repräsentationsschlusses wird demgegenüber versucht, ein Testverfahren inhaltsvalide zu machen, indem der Nachweis erbracht wird, daß das im Test geforderte Verhalten repräsentativ ist für das Verhalten, auf das geschlossen werden soll. Wird dieser Nachweis nicht erbracht und statt dessen lediglich eine Strukturgleichheit zwischen beiden Verhaltensbereichen angenommen, so wird die Inhaltsvalidität zur Augenscheinvalidität (face-validity). Glaubt der Testkonstrukteur darüber hinaus, auch auf Verhalten außerhalb des im Test gezeigten Verhaltens

schließen zu können, wird die face-validity zu einer Glaubensvalidität (faith-validity). Unter den pragmatischen Fragestellungen angewandter Diagnostik reichen die bisher angesprochenen Validierungsarten aus. Erhebt man jedoch den Anspruch, ein psychologisches Konstrukt messen zu wollen, bedarf es einer Konstruktvalidierung (vgl. CRONBACH & MEEHL, 1955). Methodisch beinhaltet sie alle nur denkbaren Verfahren zur Überprüfung des Zusammenhangs zwischen den Testwerten und allen möglichen Operationalisierungen dieses Konstrukts. Wichtiger Gesichtspunkt dabei ist eine hohe konvergente und diskriminative Validität, d.h. der Test muß hoch korrelieren mit Verfahren, die das gleiche oder ähnliche Konstrukte messen sollen, und er muß niedrig korrelieren mit solchen Verfahren, die unabhängige Konstrukte messen sollen. Wo aufgrund theoretischer Konzeptionen keine Zusammenhänge auftreten sollten, dürfen diese empirisch auch nicht festgestellt werden und umgekehrt (RETTLER, 1992).

Ein Test sollte nach LIENERT (1989) , wie erwähnt, auch normiert sein, d.h. der Testwert einer Vp sollte so interpretierbar sein, daß Aussagen über ihre Position im Vergleich zu einer Referenzpopulation gemacht werden können. Ohne eine solche Normierung kann der Test nicht für Zwecke der Routinediagnostik, sondern nur für Forschungszwecke eingesetzt werden. Der Test sollte weiterhin vergleichbar sein mit anderen Tests, die das gleiche Merkmal messen. Er soll außerdem ökonomisch, also mit geringem organisatorischem und materiellem Aufwand einsetzbar sein. Schließlich soll er ein Merkmal messen, für dessen Untersuchung auch ein praktisches Bedürfnis besteht. Dadurch wäre das Gütekriterium der Nützlichkeit erfüllt.

2.1.1.3. Kritik der Klassischen Testtheorie

Kein Zweifel besteht unter pragmatischen Gesichtspunkten an der Brauchbarkeit von Datenerhebungsverfahren, die nach den Prinzipien der KT konstruiert worden sind. Fast alle gängigen Verfahren sind so erstellt worden und liefern z.T. gute Prognosen von Kriteriumsvariablen. Von den Einwänden, die dennoch gegen die KT erhoben werden, sollen hier nur einige genannt werden (vgl. MICHEL & CONRAD, 1982, MOOSBRUGGER, 1992, RAUCHFLEISCH, 1989, SCHELTEN, 1980). Die Autoren so konstruierter Tests behaupten nämlich oft, daß sich ihre Verfahren nicht nur zur Vorhersage von Kriteriumswerten eignen würden, sondern daß in ihnen auch psychologische Merkmale operationalisiert und damit meßbar gemacht werden. Das Zustandekommen der empirisch erhobenen Testwertevariablen wird dabei zwar durch die Axiome der KT grundsätzlich erklärt, ohne daß aber auf einer theoretischen Ebene der Bezug zwischen den Testwerten und den zu messenden Konstrukten hinlänglich deutlich gemacht wird. Auf das Spekulative des Zusammenhangs zwischen Test und zu messendem Konstrukt hat REXILIUS (1978) hingewiesen: Beweise für den Zusammenhang zwischen dem Testinhalt und einer zugrundeliegenden Eigenschaft seien "zirkulär, indem von der angenommenen Eigenschaft angenommen wird, daß sie in bestimmten Testaufgaben sich wiederfindet, von denen angenommen wird, daß sie alle die gleiche angenommene Eigenschaft repräsentieren" (ebd., S. 118). Das leuchtet ein, denn "zweifellos dürfte bislang nur ein sehr kleiner Teil der Tests, aus deren Testwerten auf 'Fähigkeiten', 'Eigenschaften' etc. geschlossen wird, als hinreichend validiert im Sinne einer strengen Konstruktvalidität gelten, zumal der Prozeß der Konstruktvalidierung eines Tests ohnehin niemals als abgeschlossen angesehen werden kann"

(MICHEL & CONRAD, 1982, S. 71).

Die KT geht davon aus, mit Variablen zu arbeiten, die intervallskaliert und normalverteilt sind. Die angenommene Normalverteilung der Testdaten ist aber oft nicht gegeben und "im allgemeinen gibt es innerhalb der heutigen Testpsychologie keine Verfahren, die Intervallskaleneigenschaft der Testpunktzuordnung bezüglich des hypothetischen psychologischen Kontinuums zu sichern" (KRISTOF, 1968, zit. bei MICHEL & CONRAD, 1982, S. 23). Dennoch wird sehr oft (auch in der vorliegenden Arbeit) unterstellt, daß normalverteilte Intervallskalendaten vorliegen, allein schon um bestimmte Datenanalyseverfahren einsetzen zu können.

Es wird auch die Unabhängigkeit der Fehlerwerte von den wahren Werten sowie die Zufälligkeit der Fehlervarianz in Frage gestellt und auf die unvermeidbare Abhängigkeit der Testparameter von der oft willkürlich zusammengestellten Vpn-Stichprobe hingewiesen. Probabilistische Meßmodelle versuchen die Schwächen der KT zu überwinden, doch ist die Überlegenheit dieser Modelle keineswegs erwiesen (vgl. WALTER, 1978).

Kritik an der Testpsychologie im allgemeinen übt GRUBITZSCH (1978a, b), indem er psychologische Tests als Ausdruck der bestehenden gesellschaftlichen Verhältnisse ansieht, die vor allem "der Informationsbeschaffung über Personen zum Zwecke der Auslese, der Klassifikation oder der Zuweisung zu bestimmten Institutionen" dienen (GRUBITZSCH, 1978a, S. 40). In ihnen werden "gesellschaftliche Anforderungen zu subjektiven Fähigkeiten verkehrt" (GRUBITZSCH, 1978b, S. 78), indem sie nicht die Frage beantworten sollen, welches Verhalten das Individuum äußert, sondern ob es das geforderte Verhalten zeigt oder nicht. Menschen als Besitzer von Arbeitskraft sollen bewertet werden; damit wird ihre unmittelbare Vergleichbarkeit vorausgesetzt, die aber nur durch Abstraktion von ihrer wirklichen Ungleichheit festgestellt werden kann. Dies wird erreicht, indem ihr Verhalten und Erleben als quantifizierbar und folglich als meßbar unterstellt wird. Sie werden somit von Subjekten zu kalkulierbaren Testpersonen gemacht, deren Verhalten vorhersagbar und planbar gemacht wird.

Kritik entzündet sich auch an der Annahme, daß ein und dasselbe Merkmal bei derselben Vp mehrfach gemessen werden könne, ohne daß es dabei zu Wiederholungseinflüssen auf das Merkmal kommen würde. Die "Fiktion eines einzigen, bedingungsinvarianten 'wahren' Testwertes einer Person" (MICHEL & CONRAD, 1982, S. 25) führt zu besonderen Problemen bei Testwiederholungen und aus ihnen gebildeten Differenzwerten.

2.1.2. Veränderungsmessung

Die Kritik an der angenommenen Existenz eines einzigen wahren Wertes, der resistent gegen solche Einflüsse sein soll, die sich aus der Testwiederholung selbst ergeben, zielt auf eine "allenfalls bedingt auf die Realität psychometrischer Testuntersuchungen" (MICHEL & CONRAD, 1982, S. 25) zugeschnittene KT ab und damit auf die Unmöglichkeit der empirischen Bestimmung der Reliabilität oder der Validität. Zugleich beraubt sich die KT selbst eben durch die Zusatzannahme zeit- und situationsstabiler wahrer Werte der Möglichkeit zur Diagnostik von Veränderungen in denselben

Merkmalen von einem Meßzeitpunkt zum nächsten, erhoben mit demselben Meßinstrument. Sie ist gezwungen, intraindividuelle Veränderungen der Testwerte als Fehlerwerte zu interpretieren (CONRAD, 1992, PETERMANN, 1978). Meßinstrumente, die nach den Prinzipien der KT konstruiert worden sind, ermöglichen daher lediglich im Rahmen einer Statusdiagnostik die Feststellung eines Ist-Zustands, der eine hohe Stabilität besitzen und auf andere Zeitpunkte und Situationen verallgemeinerbar sein soll (PAWLIK, 1982). Abgesehen von der bereits angesprochenen Problematik des unklaren Skalenniveaus, das bei Vorliegen von ordinalskalierten Daten ohnehin nur sinnfreie Differenzenbildungen zulassen würde, stoßen Versuche, intraindividuelle Veränderungen trotz aller Vorbehalte im Rahmen der KT meßbar machen zu wollen, auf mindestens vier gravierende Probleme: auf das Reliabilitäts-Validitäts-Dilemma, auf den Regressionseffekt, auf den niveaukorrelierten Meßfehler und auf das Meßwertbedeutungsproblem (BEREITER, 1967, BORTZ, 1984, JÄGER & SCHEURER, 1992, LANGER, 1969, LORD, 1967, PETERMANN, 1978).

Wenn im Rahmen der KT eine Reliabilitätsbestimmung erhobener Veränderungswerte versucht wird, so zeigt sich, daß bei wachsender Korrelation der Werte aus der ersten und der zweiten Messung, mithin bei wachsender empirisch bestimmbarer Retestreliabilität des verwendeten Tests, die Reliabilität der Differenzwerte (Zweitmessung minus Erstmessung) sinkt. Umgekehrt steigt die Reliabilität der Differenzwerte bei sinkender Korrelation der Ausgangswerte an, was zu der Frage führt, welche inhaltliche Interpretation sich unter der KT für Differenzwerte angeben läßt, "wenn die in die Differenz gesetzten Tests unkorreliert sind und ... offenbar inhaltlich Verschiedenes messen" (PAWLIK, 1982, S. 26), also zu den verschiedenen Meßzeitpunkten nicht für dasselbe Merkmal valide sind. Belege für diese veränderte Validität finden sich gelegentlich in veränderten Strukturen der Faktorräume der Merkmale vor und nach einem Treatment (ROLLETT, 1982). Immerhin steigt die Reliabilität der Differenzwerte aber mit den Reliabilitäten der Ausgangswerte an. Für die Differenzwertreliabilität $R_{(xt'-xt)}$ und die durchschnittliche Reliabilität der Ausgangswerte $\bar{R}$ gilt:

$$(39) \qquad R_{(xt'-xt)} = \frac{\bar{R} - r_{xtxt'}}{1 - r_{xtxt'}}$$

(Zur Ableitung dieser Beziehung vgl. GULLIKSEN, 1950, S. 352f., TACK, 1986a, S. 51ff.) Der Grund hierfür liegt darin, daß bei der Differenzwertbildung die wahren Werte voneinander subtrahiert, die unkorrelierten Fehlerwerte aber addiert werden.

Der Regressionseffekt führt dazu, daß hohe oder niedrige Testwerte aus einer Erstmessung in einer Zweittestung dahin tendieren, sich in Richtung des Mittelwertes der Verteilung zu verändern. Dies kann dazu führen, daß die Wirkung eines möglicherweise erfolgten Treatments unter- bzw. überschätzt wird. So ein Regressionseffekt kann natürliche Ursachen haben, wenn er eine Veränderung der wahren Werte widerspiegelt; kommt er aber aufgrund mangelnder Reliabilität des Meßinstruments zustande, ist er ein statistisches Artefakt. Letzteres läßt sich veranschaulichen an einer Anzahl von Vpn, die alle denselben extremen Testwert in der Erstmessung erhalten. Durch die

mangelnde Reliabilität des Meßinstruments erhalten sie in der Zweitmessung unterschiedliche Werte. Der Mittelwert dieser Testwerte aus der Zweitmessung der Vpn liegt näher am Mittelwert der Gesamtverteilung als der Ausgangswert. Ist die Reliabilität des Tests gleich null, erhalten diese Vpn in der Zweittestung beliebige Werte, die im Durchschnitt gleich dem Verteilungsmittelwert sind.

Der Regressionseffekt kann konfundiert sein mit einem mit dem Ausgangsniveau der Testwerte korrelierten Meßfehler der Veränderung, der besonders in Fragebögen durch die Begrenzung der Itemanzahl auftritt. Das bedeutet, daß die Richtung der durch den Fehlerwert bedingten Zufallsänderung der Testwerte in dem Maße stärker determiniert ist, in dem der Testwert näher am theoretisch von der Vp erreichbaren Minimal- bzw. Maximalwert liegt. Dieser liegt durch die Art der Skalierung des Merkmals fest. Wenn der Testwert einer Vp bei der Erstmessung am Rand der Skala liegt, kann sich der Meßfehler in der Zweitmessung nur noch in Richtung einer Veränderung des Testwertes zur Skalenmitte hin auswirken.

Das Meßwertbedeutungsproblem schließlich, auch als Physikalismus-Subjektivismus-Dilemma bezeichnet, beschreibt die Schwierigkeit der inhaltlichen Interpretation von numerisch gleichen (wahren) Veränderungswerten, die durch unterschiedlich hohe Ausgangswerte zustandegekommen sind. Eine Veränderung im Mittelbereich einer Verteilung ist anders zu bewerten als eine vom Betrag her gleich große Veränderung am Rande der Verteilung. Letztere ist möglicherweise durch ein erheblich schwierigeres oder aufwendigeres Treatment erreicht worden als die erstgenannte.

Es sind vielfältige Anstrengungen unternommen worden, um trotz aller Schwierigkeiten brauchbare, d.h. vor allem reliable, Indices für Veränderungen in psychologischen Merkmalen erhalten zu können. Es sind Differenz-, Korrelations-, Regressions- und Residualmaße entwickelt worden, von denen jedes einzelne zwar die eine oder andere Verbesserung erbringt, von denen aber keines in der Lage ist, alle Probleme der Veränderungsmessung wirklich zu beseitigen. Zum Teil werden auch neue Probleme eingeführt (PAWLIK, 1982, PETERMANN, 1978, S. 36ff.). JÄGER & SCHEURER (1992) und auch PETERMANN (1986) führen noch eine Reihe anderer Verfahren und Modelle an, die der Optimierung der Veränderungsmessung dienen sollen. Dazu gehören u.a. die Bestimmung von Trend- oder Transferparametern im Rahmen probabilistischer Meßmodelle, die Anwendung von Wahrscheinlichkeitsmodellen z.B. im Rahmen der Bayes'schen Statistik oder auch varianz-, faktoren- oder zeitreihenanalytische Methoden. Alle diese Ansätze haben zumindest den Nachteil, daß sie einen hohen methodischen und rechentechnischen Aufwand erfordern, der ihren möglichen Einsatz an der Front alltäglicher Routinediagnostik nicht erleichtert. Durch die Administration von Paralleltests wurde außerdem versucht, das Auftreten korrelierter Fehlerwerte zu vermeiden, wie sie als Folge der wiederholten Vorgabe desselben Tests z.B. aus Erinnerungseffekten u.ä. resultieren können. Doch die Konstruktion hinreichend vieler Paralleltests für eine Meßserie von mehr als zwei Meßzeitpunkten stößt auf viele praktische Schwierigkeiten.

Ein besonderer Aspekt der Optimierung von Veränderungsmessungen besteht darin, änderungssensitive Items zu entwickeln. BEREITER (1967) war einer der ersten, der ein Konzept zur Entwicklung spezieller "change items" vorstellte und diese nach Gesichtspunkten der inneren Konsistenz auf der Basis der KT zu Skalenwerten aggregierte. Dabei wird die Vp selbst oder eine

dritte Person gebeten, das Ausmaß der Veränderung zwischen zwei Zeitpunkten direkt, also durch nur einmalige Testvorgabe, einzuschätzen. RENN (1973) schlägt dagegen auf der Suche nach Attitüdenindikatoren, die "das Spezifische der Dimension der Attitüdenvariabilität" (ebd., S. 61) erfassen und mithin "änderungssensitive Indikatoren" (ebd., S. 62) sein sollen, ein statistisches Kriterium zur Bestimmung der Änderungssensitivität vor. Er geht dabei nicht von den Testwerten selbst, sondern von den Abweichungswerten der Testwerte von den jeweiligen Verteilungsmittelwerten aus. Ein Item gilt demnach als änderungssensitiv, wenn die Varianz der aus den Abweichungswerten gebildeten Differenzwerte zwischen zwei Meßzeitpunkten größer als null ist. Zum Zwecke einer besseren Vergleichbarkeit sowie einer späteren Interkorrelation der Differenzwerte werden die Testwerte zuvor noch z-transformiert. Als Maß der Änderungssensitivität führt RENN (1973, S. 68f.) den Parameter C ein, der gleich der Varianz $s^2_{(zt'-zt)}$ der aus den z-Werten ermittelten Differenzwerte ist und eine Funktion der Korrelation zwischen den Testwerten aus der ersten und zweiten Messung darstellt:

$$(40) \qquad C = s^2_{(zt'-zt)} = 2(1 - r_{ztzt'})$$

Wird $r_{ztzt'}$ als Retestreliabilität aufgefaßt, so zeigt sich allerdings, daß dieser Ausdruck umso größer ist, je unreliabler der Test ist. Um den Parameter C also zum Zwecke einer Itemselektion einsetzen zu können, wären meßfehlerfreie, d.h. wahre Werte, nötig, denn sonst lassen sich wahre Merkmalsveränderungen nicht von Zufallsfluktuationen unterscheiden. RENN (1973) zeigt zwar Ansätze zur Meßfehlerbereinigung der Testwerte auf, doch kann er die Bedeutsamkeit seiner Vorschläge im Rahmen seiner eigenen Untersuchungen nicht eindeutig nachweisen (PETERMANN, 1978). TACK (1986a) bemängelt ebenso wie PETERMANN (1986) nicht nur die semantischen und konzeptuellen Unklarheiten im Umgang mit den Begriffen Änderungssensitivität bzw. Veränderung, sondern auch, daß "Vorschläge für numerische Kennwerte der Änderungssensitivität ... oft ebenfalls mit Mängeln und Ungereimtheiten belastet" (TACK, 1986a, S. 50) seien. Dennoch sind die Vorschläge von RENN (1973) z.B. von ZIELKE (1979) zur Testkonstruktion mit aufgegriffen worden und PETERMANN (1978) befindet abschließend, daß "die Überlegung, Items aufgrund ihrer Änderungssensitivität für die Veränderungsmessung heranzuziehen, ein neuer Weg" (ebd., S. 42) sei.

Die Probleme der Veränderungsmessung können jedoch im Rahmen der KT nur dann einer Lösung näher gebracht werden, wenn die Fiktion eines einzigen, zeit- und situationsstabilen wahren Wertes einer Vp zugunsten einer dynamischen, prozeßorientierten Meßkonzeption fallengelassen wird. Denn daß Menschen reifen und altern, lernen und vergessen oder sich veränderten Lebensumständen anpassen, sind ebenso alltägliche und selbstverständliche Dinge wie die Tatsache, daß Eigenschaften sich verändern können oder daß Aussagen über Personen zu einem Zeitpunkt wahr, zu einem anderen aber falsch sein können; "erst eine auf Eigenschafts-Konstanz bedachte differentielle Psychologie [läßt]

die Erfassung von Veränderungen zum Problem werden" (TACK, 1986b, S. 1). In Anbetracht des Umstands, daß nicht alles, was sich von einem Meßzeitpunkt zum nächsten im Testwert einer Vp ändert, notwendigerweise Fehleranteil sein muß, sondern auch auf wahre Merkmalsveränderung (z.B. aufgrund situativer Einflüsse) zurückzuführen sein kann, und daß empirisch der fehlerbedingte Veränderungsanteil mit einer Veränderung des wahren Wertes konfundiert sein kann, erscheint es gerechtfertigt, im Rahmen einer solchen prozeßorientierten Meßkonzeption den Ansatz der KT zu erweitern und den Testwert einer Vp zu zerlegen in einen wahren intraindividuell stabilen Wert, in einen wahren intraindividuell variablen Wert und in einen unsystematisch variierenden Fehlerwert (PAWLIK, 1982). Wendet man den Begriff der Veränderung nicht nur, wie oben nahegelegt, auf sich langfristig verändernde Merkmale an, sondern auch auf solche, die in kurzen Zeitintervallen oszillieren, z.B. Stimmungen, so gewinnt die Kritik von PETERMANN (1986, S. 13f.) an den gängigen Veränderungsstudien an Gewicht, wenn er für den Forschungsalltag u.a. fordert, nicht nur Vorher-Nachher-Vergleiche anzustellen, sondern die Strukturen von Veränderungsverläufen abbildbar zu machen, differenzierte Veränderungsbegriffe zu operationalisieren, ein Maximum an Meßwiederholungen anzustreben und Unklarheiten über Meßfehlerkonzepte zu beseitigen.

Genau in diesem Sinne der Erprobung einer prozeßorientierten Meßkonzeption unter gleichzeitiger Berücksichtigung des Forschungsbedarfs im Bereich der Veränderungsmessung versucht die vorliegende Arbeit einen Beitrag zu leisten.

2.1.3. Das State-Trait-Modell von BUSE & PAWLIK (1991)

Der Ansatz von BUSE & PAWLIK (1991) setzt die mehrfache Vorgabe desselben Items bzw. die mehrfache Messung desselben Merkmals zu mehreren aufeinanderfolgenden Zeitpunkten an mehreren Vpn voraus. Dabei wird jedes Item außerdem pro Zeitpunkt und Vp nicht nur einmal administriert, sondern zweimal. Für die Datenanalyse liegen also pro Item zwei Testwertematrizen vor mit den Testwerten von N Vpn zu n Zeitpunkten, eine mit den Testwerten aus der ersten Messung je Zeitpunkt, eine mit den Testwerten aus der zweiten Messung. Die Vpn-Vektoren sind dabei als Zeilenvektoren, die Zeitpunkte als Spaltenvektoren angeordnet. Der Testwert x in einem Test (oder Item) t bei einer Vp v wird entsprechend (2) zerlegt in einen wahren Eigenschafts- oder Traitwert t, in einen wahren Zustands- oder Statewert s und in einen Fehlerwert e:

$$(41) \qquad x_{tv} = t_{tv} + s_{tv} + e_{tv}$$

Der Traitwert ist derjenige wahre Teil des Testwerts einer Vp, der über alle Testwerte dieser Vp invariant bleibt. Der Statewert ist derjenige wahre Teil des Testwerts einer Vp, der über die beiden Messungen zu einem Zeitpunkt invariant bleibt, aber durchaus von einem Zeitpunkt zum nächsten Zeitpunkt variieren kann. Der Fehlerwert ist derjenige Teil des Testwerts, der innerhalb eines Zeitpunkts über die beiden Messungen unsystematisch variiert. Zentrale Annahme dieses Modells ist

also, daß sich die wahre Merkmalsausprägung je Zeitpunkt von der ersten zur zweiten Messung innerhalb eines Zeitpunkts nicht ändert!

Unter der weiteren Annahme, daß die Fehlerwerte weder mit den Trait- noch mit den Statewerten korreliert sind und daß auch die Statewerte (als Abweichungswerte) unabhängig von den Traitwerten sind, läßt sich die interindividuelle Varianz der Testwerte s^2_x in Anlehnung an (14) pro Zeitpunkt zerlegen in eine Summe aus der Varianz der Trait-, der State- und der Fehlerwerte:

(42) $$s^2_x = s^2_t + s^2_s + s^2_e$$

Die intraindividuelle Varianz der Testwerte s^2_i pro Vp setzt sich analog zusammen aus

(43) $$s^2_i = s^2_s + s^2_e \ .$$

Die Varianz der Traitwerte fließt in diese Beziehung nicht ein, da es sich um Testwerte ein und derselben Vp handelt. Angenommen wird ferner, daß die über die n Zeitpunkte gemittelte interindividuelle Varianz der Statewerte aus (42) gleich der über die N Vpn gemittelten intraindividuellen Varianz der Statewerte aus (43) ist; entsprechendes gilt für die Varianzen der Fehlerwerte.

Zur Schätzung der drei Modellparameter aus (42) werden drei empirisch bestimmbare Größen benötigt: die interindividuelle Varianz der pro Vp über die n Zeitpunkte gemittelten Testwerte $s^2_{\bar{x}}$, die über die N Vpn gemittelte intraindividuelle Varianz der Testwerte zwischen den n Zeitpunkten $\bar{s}^2_i$ sowie die über die N Vpn gemittelte intraindividuelle Korrelation zwischen den beiden Testwerten zu den n Zeitpunkten $\bar{r}_{ii'}$.

Ausgehend von der Varianz der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ läßt sich die Varianz der Eigenschaftswerte s^2_t errechnen. Dazu wird die Varianz der Mittelwerte (als Summenwerte) zerlegt in die Varianzen der Summanden (die Varianzen der Testwerte zu den n Zeitpunkten) dividiert durch die n Zeitpunkte. Außerdem wird die Varianz der Testwerte je Zeitpunkt noch einmal aufgespalten in die Varianzen der Trait-, der State- und der Fehlerwerte. Es wird angenommen, daß die Korrelationen zwischen den Trait- und den Statewerten, zwischen den Trait- und den Fehlerwerten, zwischen den Statewerten zum Zeitpunkt i und den Fehlerwerten zum selben Zeitpunkt

i, zwischen den Statewerten zum Zeitpunkt i und den Statewerten zum Zeitpunkt j sowie zwischen den Fehlerwerten zum Zeitpunkt i und den Fehlerwerten zum Zeitpunkt j gleich null sind. Außerdem soll gelten, daß sowohl die Varianz der Statewerte zum Zeitpunkt i als auch die Varianz der Fehlerwerte zum Zeitpunkt i gleich denen zum Zeitpunkt j sind. Die Kovarianz der Traitwerte soll desweiteren gleich der Varianz der Traitwerte sein. Unter diesen Voraussetzungen läßt sich die Varianz der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ darstellen als

$$(44) \qquad s^2_{\bar{x}} = s^2_t + \frac{(\bar{s}^2_s + \bar{s}^2_e)}{n}$$

oder anders

$$(45) \qquad s^2_{\bar{x}} = s^2_t + \frac{\bar{s}^2_i}{n} \; .$$

Die gesuchte Varianz der Traitwerte s^2_t ist dann

$$(46) \qquad s^2_t = s^2_{\bar{x}} - \frac{\bar{s}^2_i}{n} \; .$$

Aus (46) wird deutlich, daß sich die Varianz der Traitwerte mit zunehmender Anzahl der n Zeitpunkte immer stärker der Varianz der intraindividuellen Mittelwerte annähert. (Zur exakten Ableitung dieser Beziehung vgl. BUSE & PAWLIK, 1991, S. 524f.)

Die intraindividuelle Korrelation $r_{ii'}$ zwischen den Testwerten der ersten Messungen i und den Testwerten der zweiten Messungen i' einer Vp läßt sich entsprechend (17) auch darstellen als

$$(47) \qquad r_{ii'} = \frac{cov_{ii'}}{s_i * s_{i'}} \; .$$

Jeder Testwert einer Person pro Zeile der Matrix wird zerlegt in einen wahren Statewert und in einen unsystematisch variierenden Fehlerwert. (Der Traitwert kann unberücksichtigt bleiben, da er je Vp konstant bleibt und die Statewerte als Abweichungswerte vom Traitwert aufgefaßt werden. Der mittlere intraindividuelle Testwert einer Vp stellt also eine Schätzung ihres wahren Traitwertes dar.) Daher kann die Kovarianz aus (47) auch analog zu den Beziehungen in (15) und (16) geschrieben werden als

$$(48) \qquad r_{ii'} = \frac{cov_{ss'}}{s_i * s_{i'}} .$$

Es wird angenommen, daß die Varianz der Statewerte in der ersten Messung gleich derjenigen der zweiten Messung ist. Unter dieser Bedingung kann die Kovarianz der Statewerte aus der ersten und der zweiten Messung $cov_{ss'}$ gleich der Varianz der Statewerte s^2_s gesetzt werden. Auch die Fehlervarianzen zum ersten und zweiten Meßzeitpunkt sollen gleich sein. Folglich gilt:

$$(49) \qquad r_{ii'} = \frac{s^2_s}{s^2_i}$$

Wegen (43) muß gelten:

$$(50) \qquad r_{ii'} = \frac{s^2_s}{s^2_s + s^2_e}$$

Damit läßt sich die intraindividuelle Korrelation zwischen den ersten und zweiten Testwerten auffassen als das Verhältnis von der Varianz der wahren Statewerte zur Varianz der Testwerte, mithin als intraindividuelle State-Retest-Reliabilität. Aus (29) und (31) folgt damit unmittelbar:

$$(51) \qquad s^2_s = \bar{s}^2_i * \bar{r}_{ii'} \qquad \text{und}$$

$$(52) \qquad s^2_e = \bar{s}^2_i * (1 - \bar{r}_{ii'})$$

Es ist zu beachten, daß die Parameter s^2_s und s^2_e ebenso wie s^2_t sich auf die gesamte N x n Matrix beziehen und nicht für eine einzelne Vp gelten, wenngleich eine Berechnung der beiden erstgenannten Parameter selbstverständlich auch für jede Vp gesondert erfolgen könnte.

Als statistisches Maß der Änderungssensitivität einer Variablen schlagen BUSE & PAWLIK (1991) ein Varianzenverhältnis von Statevarianz zur Summe aus State- und Traitvarianz, also zur gesamten wahren Merkmalsvarianz, vor. Die Formel für diese State-Charakteristik S lautet:

$$(53) \qquad S = \frac{s^2_s}{s^2_s + s^2_t}$$

Mit hundert multipliziert gibt der Parameter S an, wieviel Prozent der Varianz der wahren Werte auf die Varianz der Statewerte zurückzuführen sind. Eine hohe State-Charakteristik eines Items besagt, daß es besonders auf Veränderungen im Zustand der Vpn anspricht und somit änderungssensitiv ist. Solche Items können als State-Indikatoren bezeichnet werden. Umgekehrt sind Trait-Indikatoren solche Items, die weniger auf intraindividuelle Merkmalsvariation ansprechen als vielmehr auf interindividuelle, habituelle Unterschiede zwischen den Vpn. Bei ihnen trägt die Trait-Varianz mehr zur Gesamtvarianz der wahren Werte bei als die State-Varianz. Die Formel für die Trait-Charakteristik T lautet:

$$(54) \qquad T = \frac{s^2_t}{s^2_s + s^2_t}$$

Beide Parameter addieren sich zu eins. Ein Item kann als State-Indikator angesehen werden, wenn deutlich mehr als die Hälfte der Gesamtvarianz der wahren Werte zulasten der Varianz der Statewerte geht. Umgekehrtes gilt für die Trait-Charakteristik.

Bei den üblichen Verfahren der Reliabilitätsbestimmung innerhalb der KT ist die State-Varianz mit der Varianz der wahren Werte konfundiert, wenn R aufgrund einer einzigen Testvorgabe zu einem bestimmten Zeitpunkt und in einer bestimmten Situation ermittelt wird. Hierbei ist jede Vp (möglicherweise) in einem jeweils anderen Zustand, der während der Testvorgabe (möglicherweise) nicht

variiert. Erfolgt die Bestimmung von R durch Vorgabe zweier Tests zu zwei aufeinanderfolgenden Zeitpunkten, so ist die State-Varianz mit der Fehlervarianz konfundiert, wenn die Vpn bei der zweiten Vorgabe in einem anderen Zustand sind als bei der ersten. Im Rahmen des hier vorgestellten State-Trait-Modells läßt sich eine statefreie Reliabilitätsbestimmung der Traitwerte vornehmen:

$$(55) \qquad R = \frac{s^2_t}{s^2_t + s^2_e}$$

Die in dieser Arbeit beabsichtigte Konstruktion eines änderungssensitiven State-Trait-Instruments setzt auf Item- wie auf Skalenebene hohe State-Charakteristiken S, zugleich von null verschiedene Trait-Charakteristiken T im Sinne von (54), hohe intraindividuelle State-Retest-Reliabilitäten im Sinne von (47) bzw. (50) und damit niedrige Fehlervarianzen im Sinne von (52) und außerdem kleine Vertrauensintervalle für die State- bzw. Traitwerte (analog zu (33)) voraus. Zusätzlich müssen die Testgütekriterien der Objektivität, auf Skalenebene die der Reliabilität im herkömmlichen Sinne (vgl. Kap. 2.1.1.2.) sowie die der (im Rahmen dieser Arbeit zumindest angenäherten) Validität erfüllt werden. (Zu weiteren ökospychologischen Gütekriterien vgl. Kap. 2.2.3.) Obwohl dieses Meßinstrument auch zur Ermittlung von Traitwerten tauglich sein soll, ist eine spezielle Berücksichtigung solcher Items, die über eine hohe Trait-Charakteristik verfügen, nicht erforderlich, da eine hohe Reliabilität der Traitwerte durch Aggregation der Testwerte über viele Zeitpunkte bzw. Situationen sichergestellt ist (vgl. Kap. 2.3.1.).

2.2. Ökopsychologische Grundlagen

Unter Ökopsychologie soll "die ... Erforschung der Interaktion ... zwischen Erleben und Verhalten auf der einen und den natürlichen ... Umweltbedingungen auf der anderen Seite verstanden werden" (PAWLIK, 1978, S. 112f.). Sie muß dabei von der Umweltpsychologie abgegrenzt werden, die sich eher mit konkreten, oft aus außerpsychologischen Gründen für forschungsbedürftig erachteten und abgegrenzten Umweltfragen auseinandersetzt und die dabei ein recht buntes Bild hinsichtlich ihres Themenkreises abgibt, zu dem auch Fragen des Umweltschutzes und des Umweltbewußtseins gehören (ebd., vgl. auch JOERGES, 1990, S. 449). Beide Begriffe werden sowohl von PAWLIK (1978) als auch von KAMINSKI & BELLOWS (1982) unter dem der "Ökologischen Psychologie" subsumiert, während JOERGES (1990) alle drei Begriffe als Synonyme betrachtet. So befinden denn KAMINSKI & BELLOWS (1982) auch, daß es sich bei diesen Teilgebieten um "bislang kaum mehr als eine sehr heterogene Ansammlung von Perspektiven, Ansätzen, Bestrebungen" (ebd., S. 87) handele, über die nur unter Vorbehalten einige verbindende Charakteristika aufgelistet werden können. Einer der Interessenschwerpunkte innerhalb der Psychologie, die das Attribut "ökologisch" für sich in Anspruch nehmen, liegt demnach dort, wo unter methodologischen Gesichtspunkten eine Verlagerung des Forschungsansatzes weg von künstlich geschaffenen Untersuchungsbedingungen hin zu einer naturalistischen Herangehensweise stattfindet, die menschliches (Alltags-) Verhalten in situ, d.h. unter natürlichen Lebensbedingungen, erforschbar machen will. Dadurch wird zugleich eine

Abkehr von Erklärungsansätzen des Verhaltens oder Erlebens auf einer molekularen Reiz-Reaktionsebene hin zu einer komplexen, molaren und zugleich auch differenzierten Ebene alltäglicher Person-Umwelt-Interaktionen ermöglicht (JOERGES, 1990, KAMINSKI & BELLOWS, 1982, WICKER, 1979). In praxi verlangt dies vom Psychologen nicht weniger als das wirkliche Leben der Menschen dort zu untersuchen, wo es stattfindet, und es erfordert nicht weniger Aufwand, als wenn ein Ethologe den Zoo verlassen und in den Urwald fahren würde, um das Verhalten der Tiere in freier Wildbahn zu beobachten. Entsprechend meint Ökopsychologie auch "kein besonderes Teilgebiet der Psychologie, sondern ... eine besondere Perspektive in Fragestellung, Methode und Theorie" (PAWLIK, 1978, S. 113). So grotesk es erscheinen mag, daß ein Ethologe sein verhaltensbiologisches Wissen fast ausschließlich aus der Beobachtung von solchen Tieren erlangen könnte, die in Zoogehegen gehalten werden, so bedrückend erscheint der Gedanke, daß genau dieser Vergleich den gegenwärtigen Zustand der Psychologie als der Wissenschaft vom Verhalten und Erleben des Menschen treffend beschreibt. Die vorliegende Arbeit ist als ein Beitrag zur Entwicklung des für den Aufbruch in den Lebensraum der Menschen notwendigen methodischen Instrumentariums konzipiert.

2.2.1. Labor- versus Feldforschung

Zwar liegt auch der Ursprung der Psychologie, wie der jeder Wissenschaft, in dem gelebten und erlebten Alltag denkender, fühlender und handelnder Menschen, doch der Alltag psychologischer Forschung spielte sich in der kurzen Geschichte dieser Wissenschaft zum größten Teil eher in einem Raum ab, der vom gelebten Alltag weit entfernt war, nämlich im Testlabor "mit seinen eigens zum Zweck der Untersuchung geschaffenen Bedingungen" (BUSE & PAWLIK, 1990, S. 213). Da es vermutlich kaum einen komplexer beschaffenen wissenschaftlichen Gegenstandsbereich gibt als den der Humanwissenschaften, verwundert es nicht, wenn zunächst mit einfachen Methoden und Theorien unter restringierten Bedingungen versucht wurde, ein wenig Licht in das Dunkel zu tragen: Man lud die Menschen ins psychologische Labor ein, nannte sie Versuchspersonen und bat sie, sinnlose Silben zu lernen, Helligkeitsstufen zu unterscheiden oder Fragen zu beantworten, die sie sich selber nie gestellt hätten (PATRY, 1982). Gelegentlich griff und greift man auch auf die Verhaltensweisen spezieller Versuchstiere zurück, die in Labyrinth-Käfigen oder in Skinner-Boxen beobachtet werden, ähnlich wie Zootiere. Daneben gibt es allerdings schon seit Jahrzehnten Versuche, "eine 'Psychologie der Umwelt' zu begründen und psychologische Fragestellungen nach den Beziehungen zu großräumigen Umwelten zu systematisieren" (JOERGES, 1990, S. 448), d.h. die Fragestellungen und die Anliegen, die hinter ihnen stehen, dort zu untersuchen, wo sie ihrem Wesen nach entstammen, im wirklichen Lebensraum lebendiger Menschen.

Die hier angedeuteten gegensätzlichen Vorgehensweisen der Laborforschung auf der einen und der Feldforschung auf der anderen Seite "markieren die Extreme eines Kontinuums unterschiedlich 'lebensnaher' ... Untersuchungen" (BORTZ, 1984, S. 33). Eine genaue begriffliche Unterscheidung zwischen beiden Forschungsvarianten läßt sich schon deshalb nicht vornehmen, weil es zum einen im Sinne eines Kontinuums zwischen beiden Polen fließende Übergänge gibt und weil es zum anderen sehr unterschiedliche Definitionen besonders für den Begriff "Feld" bzw. "Feldforschung" gibt. So fordern einige Autoren, von Feldforschung könne erst dann gesprochen werden, wenn der Feldstatus

der hinter der Untersuchung stehenden Theorie, d.h. das Ausmaß der in ihr enthaltenen Feldbeziehungen, geklärt sei, andere Autoren sprechen von Feldforschung und meinen eine ganz bestimmte Datenerhebungsmethode während wieder andere den Ort der Datenerhebung für das entscheidende Charakteristikum halten (GACHOWETZ, 1987). PAWLIK (1988) will von Laborforschung dann sprechen, wenn die psychologischen Daten "unter Bedingungen erhoben werden, die eigens zum Zweck solcher Datenerhebung vom Untersuchungsleiter eingerichtet sind" (ebd, S. 169), und von Feldforschung dann, wenn die Daten unter Bedingungen generiert werden, "die unabhängig von solcher Datenerhebung für sich als Lebenswirklichkeit bestehen" (ebd.). Eine solche eher pragmatisch-technologische Begriffsbestimmung nimmt auch PATRY (1979, 1982, zit. bei GACHOWETZ, 1987, S. 258) vor, wenn er die Feldcharakteristik einer Untersuchung durch deren Ausprägung in fünf Merkmalsdimensionen beschreibt. Zunächst kann eine Untersuchung nach der Natürlichkeit bzw. Künstlichkeit des Zustandekommens der unabhängigen Variablen (Treatment), der abhängigen Variablen (Vpn-Verhalten) und des Settings klassifiziert werden. Darüber hinaus kann unterschieden werden zwischen informierten Vpn, die um die Durchführung der Untersuchung wissen, und uninformierten Vpn, die dies nicht wissen, sowie bei den informierten zwischen naiven und eingeweihten Vpn, je nachdem, ob ihnen zusätzlich noch das Ziel der Untersuchung bekannt ist. Die reinste Form einer Felduntersuchung liegt vor, wenn uninformierte, naive Vpn unter (im genannten Sinne) natürlichen Bedingungen untersucht werden. Die Unterscheidung zwischen natürlich (synonym: naturalistisch) und künstlich ist daher nicht einfach deckungsgleich mit der zwischen Labor und Feld, vielmehr gilt: je natürlicher, desto mehr Feldcharakter und umgekehrt. Je mehr der Untersucher in die Bedingungsvariationen der Datenerhebung eingreift, je weniger diese von sich aus passieren können, desto künstlicher, laborähnlicher der Untersuchungscharakter (vgl. PAWLIK, 1988). Eine unter vollkommen natürlichen Bedingungen ablaufende Felduntersuchung mit uninformierten Vpn muß dennoch nicht automatisch zu unverfälschten Daten führen, vielmehr kann der Prozeß der Datengewinnung selbst zu nicht kontrollierten, unbeabsichtigten Veränderungen des Feldes führen. Solche Reaktivitätseffekte sind immer dann nicht auszuschließen, wenn den Vpn bewußt ist, daß eine Untersuchung durchgeführt wird (GACHOWETZ, 1987, S. 259). Dies kann auch bei uninformierten Vpn schon durch auffälliges Verhalten eines schlecht geschulten Beobachters geschehen. ORLIK (1979) untergliedert den Bereich der Feldforschung in Feldstudien, Feldexperimente und Ansätze der Aktionsforschung. Während die letzten beiden Kategorien durch mehr oder minder starke Eingriffe des Forschers in das Feld gekennzeichnet sind, zeichnen sich die Feldstudien durch eine eher passiv registrierende Haltung des Untersuchers aus. Sie werden eher in solchen Bereichen des Alltagslebens angewandt, über die nur unzureichende Erkenntnisse und Erklärungsmodelle vorliegen; sie haben somit eher phänomenologisch-deskriptiven Charakter. In diesem Sinne handelt es sich auch bei der vorliegenden Arbeit um eine Feldstudie, wenngleich sie der Erkundung und Beschreibung der Untersuchungsmethode und nicht des Untersuchungsinhaltes dient.

Die bisher geschilderten Unterschiede zwischen Labor- und Feldforschungsansätzen legen die Vermutung nahe, daß die Feldforschung in jedem Falle die erstrebenswertere (weil realitätsnähere) Vorgehensweise sei. Dies ist jedoch nur bedingt richtig. Der Grund für das Vorherrschen von Laboruntersuchungen in der Wissenschaftstradition der Psychologie liegt vermutlich weniger in dem geringeren methodischen Aufwand, den sie im Gegensatz zu Felduntersuchungen erfordern, sondern

(neben einer Favorisierung einfacher Reiz-Reaktionsmuster als theoretischer Grundlage zur Erklärung menschlichen Verhaltens) gerade in ihrer Beschränkung auf wenige, leicht überschaubare Variable. Sie ermöglicht es, Störeinflüsse, die die Ausprägung der abhängigen Variablen beeinflussen könnten, auszuschalten, konstant zu halten oder zu kontrollieren. Durch diese Beschränkung nimmt die Wahrscheinlichkeit zu, Veränderungen in den abhängigen Variablen ursächlich auf Einflüsse der unabhängigen Variablen zurückführen zu können. Damit sinkt gleichzeitig die Anzahl plausibler Alternativerklärungen für die Ergebnisse. Genau das Gegenteil gilt für Felduntersuchungen, in denen vielfältige Einflußgrößen häufig mehrere gleichwertige Erklärungsalternativen ermöglichen. Laboruntersuchungen stehen also in dem Ruf, im Gegensatz zu Felduntersuchungen eine hohe interne Validität zu besitzen (BORTZ, 1984). Andererseits häufen sich in der Literatur die Belege dafür, daß dies eine Fiktion ist. Zwar können die Untersuchungsbedingungen im Labor besser kontrolliert werden, doch welche Wirkungen dies im einzelnen auf die Vpn hat, bleibt oft unklar (GACHOWETZ, 1987).

Felduntersuchungen stehen ihrerseits in dem Ruf, durch ihre Realitätsnähe und damit durch ihre leichte Generalisierbarkeit über die untersuchten Personen und die spezielle Untersuchungssituation hinaus eine hohe externe Validität zu besitzen. Laboruntersuchungen sollen demgegenüber wegen der Unnatürlichkeit der Untersuchungsumgebungen nur eine geringe externe Validität besitzen (BORTZ, 1984). Von den gegen diese Ansicht vorgebrachten Einwänden soll nur der von SCHULER (1980, zit. bei GACHOWETZ, 1987, S. 264) angeführt werden: Gerade wenn die Situation in einer Felduntersuchung als Folge ihrer Realitätsnähe durch einen hohen Komplexitätsgrad gekennzeichnet ist, so ist eine Generalisierung auf andere Situationen bzw. Personen nur möglich, wenn diese komplexen Strukturen in ihren relevanten Zusammenhängen auf einer theoretischen Ebene hinlänglich verstanden sind. Ansonsten können spezielle Nebenbedingungen bei der Durchführung einer konkreten Felduntersuchung der Generalisierbarkeit ihrer Ergebnisse enge Grenzen setzen.

Die Frage, welcher Forschungsansatz generell der bessere sei, läßt sich gegenwärtig nicht entscheiden. Im Spannungsfeld von interner und externer Validität geht es vor allem um zwei extreme Polarisierungen wissenschaftlicher Fragestellungen: um die Erlangung von gesichertem Wissen einerseits und um die Relevanz der Ergebnisse andererseits. Ökologisch orientierte Psychologen beklagen denn am methodischen Vorgehen vieler ihrer laborientiert arbeitenden Kollegen auch, "that generalization to the authentic significance of the person in the real environment has been sacrificed to the quest for certainty in our knowledge" (GIBBS, 1979, S. 127). Felduntersuchungen sind besser als Laboruntersuchungen geeignet, den Forschungsprozeß in Bereichen mit geringem theoretischem Kenntnisstand durch Generierung von Hypothesen und Aufdeckung von Strukturen voranzutreiben. Laboruntersuchungen "fordern nämlich eine Kenntnis der Struktur der zu untersuchenden Probleme, um eine geeignete Transposition ... dieser Strukturen in Laborsituationen durchführen zu können" (GACHOWETZ, 1987, S. 257f.).

Liegen allerdings genügend Erkenntnisse über einen bestimmenten realitätsnahen (z.B. angewandt-psychologischen) Untersuchungsbereich vor, so ist es durchaus möglich, ökologisch orientierte Forschung auch im Labor vorzunehmen. Es ist nämlich nicht so, daß Feldforschung der einzige methodische Zugang für einen ökopsychologischen Untersuchungsansatz ist. Wenn, wie bereits

erwähnt, unter "Ökopsychologie" im Sinne von PAWLIK (1978, S. 113) "eine besondere Perspektive in Fragestellung, Methode und Theorie" verstanden wird, so bedeutet dies auf der Methodenseite zunächst einmal, sich von der herkömmlichen, auf die Untersuchung isolierter Phänomene beschränkter Laborforschung zu verabschieden. "Die Sozialwissenschaften [gehören] ins Feld geworfen" (GACHOWETZ, 1987, S. 276), um sich dort zu bewähren. Sind aber die Strukturen und Bedingungen, die Interaktionen zwischen Menschen und ihren Lebensräumen einigermaßen geklärt, so kann eine Rückkehr ins Labor unter Bewahrung des Anspruches auf externe Validität erwogen werden, um dort gezielt die für das Verhalten und Erleben im Feld wirklich relevanten Bedingungen, zumindest in einigen Dimensionen, systematisch zu variieren. Dann sind die Voraussetzungen für eine Transposition der Lebensstrukturen in Laborsituationen tatsächlich gegeben und es kann auch dort authentische, wirklichkeitsbezogene Forschung vorgenommen werden. In einigen Bereichen wird dies bereits versucht, indem z.B. Realitätsausschnitte (Museum, Arbeitsplatz) im Labor simuliert werden und der Laborforschung dadurch zumindest eine Ergänzungsfunktion zur naturalistischen Methodik zugedacht wird (KAMINSKI & BELLOWS, 1982).

2.2.2. Methoden ökopsychologisch orientierter Feldforschung

Begibt sich der Psychologe aber zunächst einmal ins Feld, d.h. an die Front, so begegnet ihm ein Gegenstandsbereich von unüberschaubarer Weite und Komplexität, denn mit der Absicht, naturalistische Methoden zur Erkundung des wirklichen Lebens einzusetzen, könnte er an jedem Alltagsverhalten eines jeden Menschen in jeder Lebenssituation ansetzen. Erschwerend kommt der Anspruch hinzu, die vorgefundenen Zusammenhänge nicht unterhalb eines gewissen Differenziertheitsgrades analysieren zu wollen (KAMINSKI & BELLOWS, 1982). Entsprechend vielfältig ist das methodische Instrumentarium, das bisher verwendet wurde. Es wurden Selbst- oder Fremdprotokollierungen von Verhaltens-, Erlebnis- und Settingvariablen, von kritischen Lebensereignissen oder Alltagsereignissen vorgenommen, es wurden Zeit- oder Ereignisstichproben gezogen und Fragestellungen unterschiedlichster Art bearbeitet, um nur einige Unterscheidungsdimensionen von Feldforschungsaktivitäten zu nennen. Es wurde exzessive Datensammlung betrieben, indem etwa das Verhalten eines ganz gewöhnlichen sieben Jahre alten Jungen während seiner gesamten Wachphase an einem ganz gewöhnlichen Alltag in einer nordamerikanischen Kleinstadt von 7:00h morgens bis 20:33h abends minutiös fremdprotokolliert wurde (BARKER & WRIGHT, 1951), oder es wurden Selbstprotokollierungen exzessiven Verhaltens angeleitet, wie die eines 54jährigen Amerikaners, der alleine, ohne Zwischenstop und ohne Landsicht innerhalb von 150 Tagen die Welt umsegelte und dabei täglich im Wechsel verschiedene psychologische Tests bearbeitete (PALUS, NASBY & EASTON, 1990).

Im Rahmen der vorliegenden Arbeit sind (Selbst-) Protokollierungen von alltäglichem Verhalten bzw. Erleben mit Hilfe von Mehrzeitpunkterhebungen von besonderem Interesse, die daher im folgenden auch vorwiegend betrachtet werden sollen. Obwohl es sie schon mindestens seit der Arbeit von BARKER & WRIGHT (1951) gibt, sind sie in den fünfziger und sechziger Jahren nur verhalten vorangetrieben worden. Zu dem dann allmählich wachsenden Interesse an solchen Untersuchungsmethoden trugen Fragestellungen aus dem Bereich des Behaviorismus, aus dem Bereich der Verhaltensmedizin und besonders solche aus dem industrie- bzw.

organisationspsychologischen Bereich bei (WHEELER & REIS, 1991). Der Sprung kam Mitte der siebziger Jahre, als die Zahl der Untersuchungen bzw. Veröffentlichungen zu diesem Thema merklich zunahm (PAWLIK, 1988, TENNEN, SULS & AFFLECK, 1991). Neben der Untersuchung von Bedingungszusammenhängen bzw. Mensch-Umwelt-Wechselwirkungen (z.B. im Rahmen differentiell-psychologischer Fragestellungen) steht die Entwicklung einer auf naturalistischen Daten aufbauenden Feldpsychodiagnostik bzw. Feldpsychometrie im Zentrum des Forschungsinteresses. Bedarf nach Methoden zur Registrierung von Verhalten im Feld besteht allerdings auch im therapeutischen und pädagogisch-psychologischen Bereich (PAWLIK, 1988, PAWLIK & BUSE, 1982, 1992). Inzwischen bedient sich die psychophysiologische Forschung unter Einsatz entsprechender Aufzeichnungsgeräte ebenfalls solcher Untersuchungsansätze (FAHRENBERG & HEGER, 1991, PAWLIK & BUSE, 1992).

Die hier zur Diskussion stehenden Datenerhebungsverfahren fordern die Psychologie im konzeptionellen, methodischen und datenanalytischen Bereich heraus (TENNEN, SULS & AFFLECK, 1991). Es geht dabei um Klärungen der Interaktionen zwischen Persönlichkeit und Umwelt, insbesondere um Fragen der intersituativen Verhaltenskonsistenz (BUSE & PAWLIK, 1984, PAWLIK & BUSE, 1992); es geht um erhöhte Anforderungen an die Vpn und an die Versuchsleiter, um die Berücksichtigung möglicher Erinnerungs- und Reaktivitätseffekte; schließlich müssen die gewaltigen Datenmengen, die bei der Erhebung von vielen Merkmalen an vielen Vpn über viele Zeitpunkte anfallen, sinnvoll aufbereitet, erfaßt, strukturiert und analysiert werden. Warnend bemerken TENNEN, SULS & AFFLECK (1991, S. 323) dazu: "The field has not matured sufficiently to establish accepted procedures." Im übrigen gibt es kaum Gründe anzunehmen, daß die herkömmlichen Störeffekte, die empirische Forschungsarbeiten beeinträchtigen können (vgl. AMELANG & BARTUSSEK, 1981, GNIECH, 1976, LÖSEL, 1992, SCHULZ, MUTHIG & KOEPPLER, 1981), bei der Protokollierung von Alltagsverhalten nicht auftreten könnten.

PAWLIK (1988, S. 176ff.) empfiehlt anhand von mehreren Kriterien eine optimale Gestaltung von Verfahren der Feldpsychodiagnostik, die auf im Alltag gewonnene Verhaltens- oder Erlebnisdaten zurückgreifen. Die Basis der Beobachtungen sollten demzufolge Protokollierungen in situ, d.h. unmittelbar im Feld und gleichzeitig mit dem Verhalten, bilden. Alternativ dazu können u.U. Audio- oder Videoaufzeichnungen erfolgen. Retrospektive Protokollierungen sollten wegen möglicher Gedächtniseffekte vermieden werden (vgl. dazu HEDGES, JANDORF & STONE, 1985). Der Beobachter des Verhaltens sollte zugleich auch die zu beobachtende Person sein, da nur so ethisch einwandfreie Beobachtungen auch des nicht-öffentlichen Verhaltens der Vpn durchführbar sind. Neben der möglichen Unzuverlässigkeit der Vpn als Beobachter kann dabei allerdings das Unterbrechen des Verhaltensflusses zum Zwecke seiner Protokollierung eine Schwierigkeit darstellen. Das Beobachtungsschema sollte eine Inventarliste für Setting- und Verhaltensvariable mit vorgegebenen Kategorien sein, die ein geschlossenes Antwortformat haben sollten. Als Protokollform sollten für Setting- und Verhaltensvariable dichotome Antwortformate (z.B. "trifft zu" - "trifft nicht zu"), für Stimmungsvariable Ratingskalen verwendet werden. Als Registriermethode befürwortet PAWLIK (1988) den Einsatz von Verhaltensdatenrecordern, wie sie von ihm selbst zusammen mit BUSE (BUSE & PAWLIK, 1984, PAWLIK & BUSE, 1982) entwickelt worden sind. Ein ähnliches Verfahren haben später auch PERREZ & REICHERTS (1989) in Form eines selbstprogrammierten

Minicomputers angewandt. Zweifellos hat dieser Ansatz erhebliche Vorteile gegenüber der von PAWLIK (1988) und auch von STONE, KESSLER & HAYTHORNTHWAITE (1991) nicht favorisierten herkömmlichen (Papier-Bleistift-) Protokollbogen-Methode, doch ist diese "very sophisticated technology" (HORMUTH, 1986, S. 275) auch nicht ohne Nachteile: sie ist zumindest teuer und kann ein intensives Training der Vpn erforderlich machen (PAWLIK & BUSE, 1982). Für einen diagnostischen Routineeinsatz erscheint die Protokollbogen-Methode daher mindestens ebenso geeignet.

Die Notwendigkeit der Erstellung eines solchen expliziten Anforderungsprofils für Datenerhebungsverfahren, die verwertbare Informationen auf der Basis von Alltagsverhalten liefern sollen, resultiert aus dem Wesen des zu beobachtenden Verhaltensstroms, d.h. der "zeitlich gereihten Abfolge der Erlebnis- und Verhaltensvariationen einer Person in ihrem natürlichen Lebensraum" (PAWLIK, 1988, S. 171), dessen psychodiagnostische Erfassung das Ziel der Feldpsychodiagnostik ist. Mit herkömmlichen Fragebogenmethoden ist dies nicht leistbar, denn die sind lediglich zur "Abbildung der Wahrnehmung, Vorstellung oder Gedächtnisreproduktion von eigenem oder Fremdverhalten" (PAWLIK & BUSE, 1982, S. 102) geeignet, nicht aber zur Registrierung des Verhaltens selbst. Die Verarmung psychologischer Diagnostik ist also in vielen Fällen eine zweifache: Zum einen werden die Vpn meist nur unter Laborbedingungen und nicht im Alltag untersucht und zum anderen wird dann im Labor oft noch nicht einmal deren richtiges Verhalten beobachtet, sondern die mentale Repräsentation ihres (angeblich) üblichen Verhaltens draußen im Alltag erfragt, d.h. verbal erhoben. Das wirkliche Verhalten der Vpn im Feld unterscheidet sich jedoch in mancher Hinsicht beträchtlich von dem im Labor erfaßbaren, was im Rahmen der Untersuchungsmethodik berücksichtigt werden muß (PAWLIK, 1988, S. 171ff.): So ist der Verhaltensstrom räumlich nicht stationär, d.h. Alltagsverhalten findet in wechselnden Umgebungen statt und verlangt damit nach mobilen Aufzeichnungsverfahren. Er ist außerdem kontinuierlich, d.h. nicht nur, daß er nicht einfach irgendwann aufhört, sondern daß die Ausprägungen des interessierenden Verhaltens oder Erlebens mit Phasenlängen im Sekundenbereich und darunter über die Zeit von Minuten bis Stunden variieren können, was die Erfassung durch Beobachter erheblich erschweren kann. Das Leben hört nicht auf, wenn der Labortermin vorbei oder die Kanalkapazität der Beobachter erschöpft ist. Zudem ist der Verhaltensstrom zeitlich nicht fertig vorgegliedert. Die zu untersuchenden Einheiten müssen entsprechend der Fragestellung erst festgelegt werden. Das gleiche gilt für die Auswahl der zu untersuchenden Merkmale aus der Menge aller, z.T. kovariierender Merkmale, von denen nur ein kleiner (im Sinne der Fragestellung) relevanter Teil erfaßt werden kann. Dabei ist zu berücksichtigen, daß Handlungen als sinnvolle Verhaltens- bzw. Merkmalsstrukturen, die z.B. durch einen Ausgangs- und einen Endzustand gekennzeichnet sein können, einander zeitlich überlagert sein können oder ineinander verschachtelt auftreten können. Der Verhaltensstrom findet weiterhin auch nicht losgelöst von der Umwelt statt. Zu diagnostischen Zwecken müssen daher solche Umgebungsvariablen miterhoben werden, von denen anzunehmen ist, daß sie den Verhaltensstrom beeinflussen; ansonsten bleibt die Feststellung von Verhaltensänderungen mehrdeutig. Letztlich treten in der Feldpsychodiagnostik besondere Probleme auf, die daraus resultieren, daß der interessierende Verhaltensstrom zumindest teilweise nicht-öffentlich ist oder aber zwar öffentlich, jedoch nicht für eine Weiterverwertung durch Dritte freigegeben ist. Andererseits ist das Interessante am Alltagsverhalten gerade, daß es "natürlich" ist

und nicht nur zu Forschungszwecken stattfindet.

Soll das Alltagsverhalten zwar systematisch, aber nicht mit dem Anspruch quantitativer und qualitativer Vollständigkeit erfaßt werden (wie bei BARKER & WRIGHT, 1951), so ist es erforderlich, Stichproben aus diesem Verhaltensstrom zu ziehen, d.h. das Verhalten oder Erleben der Vpn nur zu bestimmten Zeitpunkten zu erheben. Die Güte einer psychodiagnostischen Untersuchung unter Feldbedingungen hängt entscheidend von der Wahl geeigneter Protokollzeitpunkte ab. Im wesentlichen lassen sich dabei zwei Arten von Stichprobenplänen unterscheiden: die Ereignis- und die Zeitstichprobenpläne. Ereignisstichproben gehen von einer kontinuierlichen Beobachtung des Verhaltensstroms aus. Tritt dabei ein (entsprechend der Fragestellung) genau festgelegtes, aber eher selten vorkommendes Verhalten auf, so wird es protokolliert. Hierbei wird der Protokolltermin also durch das Auftreten des interessierenden Verhaltens festgelegt. Der entscheidende Nachteil dieses Verfahrens bei Selbstprotokollierungen liegt in möglichen Reaktivitätseffekten, durch die die Auftretensfrequenz des Verhaltens infolge der ständigen Selbstbeobachtung verändert werden kann; ein Vorgang, der in der Verhaltenstherapie gezielt genutzt wird (BUSE & PAWLIK, 1990, PAWLIK, 1988). Zeitstichprobenpläne sehen dagegen die Protokollierung des interessierenden Verhaltens bzw. Erlebens zu Beobachtungsterminen vor, die so gewählt sind, daß "eine repräsentative Stichprobe der Inhalte und Bedingungen des Verhaltensstroms zu erwarten ist" (PAWLIK, 1988, S. 173). Wichtige Parameter eines Zeitstichprobenplans sind der Beobachtungszeitraum, d.h. die Zeitspanne von der ersten bis zur letzten Protokollierung, das Beobachtungszeitfenster, d.h. die Länge der Zeit, für die die Angaben gemacht werden sollen, sowie das Beobachtungsintervall, womit die Dauer zwischen zwei aufeinanderfolgenden Beobachtungsterminen gemeint ist. Durch diese wird die zeitliche Auflösungsfähigkeit der Untersuchung für das untersuchte Verhalten festgelegt (PAWLIK, 1988). Soll von der Untersuchung auf den Änderungsverlauf des untersuchten Verhaltens geschlossen werden, so setzt dies voraus, daß bei periodisch schwankenden Merkmalen entsprechend dem Abtast-Theorem der Zeitreihenanalyse die Abtastrate, d.h. die Häufigkeit der Beobachtungstermine, mindestens doppelt so hoch ist wie die höchste Änderungsfrequenz des untersuchten Merkmals. Die Nachteile dieses Ansatzes sind zum einen, daß Vorkenntnisse über den Änderungsverlauf des Merkmals zum Zwecke seiner Abbildung erforderlich sind, aus denen auf die maximale Länge des Beobachtungsintervalls geschlossen werden kann. Zum anderen sind Zeitstichprobenpläne nicht zur Erfassung seltener Ereignisse geeignet.

2.2.3. Experience Sampling Method

Zeitstichprobenpläne lassen sich anhand der Vorhersehbarkeit bzw. anhand der Beeinflußbarkeit eines konkreten Protokolltermins durch die Vpn noch weiter differenzieren in intervallabhängige und signalabhängige Versuchspläne (WHEELER & REIS, 1991). Bei den intervallabhängigen Versuchsplänen werden Protokollierungen zu theoretisch sinnvollen, aber von den Vpn vorhersehbaren Zeitpunkten vorgenommen, z.B. nach dem Aufstehen, nach dem Abendessen, vor dem Schlafengehen usw., zu denen die Vpn ihr Verhalten oder Erleben für den jeweiligen Zeitpunkt (CAMPBELL, CHEW & SCRATCHLEY, 1991) oder summarisch für die seit dem letzten Protokolltermin verstrichene Zeit (LARSEN & KASIMATIS, 1991) angeben müssen. Bei den signalabhängigen Versuchsplänen sind die Vpn aufgefordert, ihr Verhalten oder Erleben dann zu

protokollieren, wenn sie von dem Untersuchungsleiter eine nicht vorhersehbare Aufforderung, ein Signal dazu erhalten. Es gibt eine Reihe von Zwischenformen aus beiden Versuchsplänen, die so beschaffen sind, daß die Vpn zwar nicht mehr zu solchen Zeitpunkten ein Protokoll anfertigen müssen, die sich aus ihrem eigenen Tagesablauf ergeben (und die somit stark verhaltens- bzw. umgebungsabhängig sind), sondern zu Zeitpunkten, die vom Untersucher vorgegeben werden. Dabei haben sie aber entweder im voraus Kenntnis von diesen Zeitpunkten oder sie können den genauen Zeitpunkt in Grenzen selber festlegen. So forderten NOWLIS & COHEN (1968) ihre "free-ranging" (ebd., S. 551) Vpn auf, mindestens einmal stündlich eine Stimmungsliste zu bearbeiten, ZEVON & TELLEGEN (1982) gaben ihren Vpn täglich eines von drei Zeitintervallen, bestehend aus je fünf Stunden, vor, innerhalb dessen sie ihre Stimmungsliste bearbeiten sollten und BRANDSTÄTTER (1983) übergab seinen Vpn einen kompletten Plan mit allen Protokollterminen, an die sich die Vpn dann selbst erinnern mußten.

Reine signalabhängige Zeitstichproben aus dem alltäglichen Verhaltensstrom der Vpn unter nicht-restringierten Umweltbedingungen, bei denen die Vpn das Eintreten eines Protokolltermins weder vorhersehen noch beeinflussen können, lassen sich praktisch nur als Selbstbeobachtungen realisieren. Sie erfordern einen Signalgeber, der die Vpn beim Erreichen eines Protokolltermins unabhängig von der jeweiligen Umgebung, in der sie sich gerade befinden, und unabhängig von dem gerade gezeigten Verhalten darüber informiert, daß sie in dem jeweiligen Moment die geforderten Angaben zu ihrem Verhalten oder Erleben machen sollen. Ein solcher Untersuchungsansatz wird in der angelsächsischen Literatur als "experience sampling method" (ESM) (CSIKSZENTMIHALYI & LARSON, 1987, HORMUTH, 1986, STONE, KESSLER & HAYTHORNTHWAITE, 1991) bezeichnet und deutsch als "Erfahrungsstichprobe" (BOHNER, HORMUTH & SCHWARZ, 1991) übersetzt. Die der vorliegenden Arbeit zugrundeliegende Untersuchung ist als ESM-Studie konzipiert und durchgeführt worden.

Die erste Arbeit unter Einsatz eines Protokollterminplans und eines Taschenweckers, den die Vpn selbst nach jedem erfolgten Protokollieren für den nächsten Termin neu stellen mußten, wurde Mitte der sechziger Jahre von HINRICHS (1964, zit. bei WHEELER & REIS, 1991) durchgeführt. In neuerer Zeit ist dieses Vorgehen mittels Armbanduhren, die über eine Weckeinrichtung verfügten, von DIENER und Kollegen (DIENER, LARSEN & EMMONS, 1984, DIENER & LARSEN, 1984) praktiziert worden. Ab den siebziger Jahren wurden dann Pieper eingesetzt, wie sie etwa von Ärzten in Krankenhäusern verwendet werden, die die Vpn ständig mit sich führen mußten. Immer, wenn die Vpn vom Untersuchungsleiter "angepiept" wurden, mußten sie einen Protokollbogen ausfüllen (CSIKSZENTMIHALYI & FIGURSKI, 1982, McADAMS & CONSTANTIAN, 1983, WONG & CSIKSZENTMIHALYI, 1991). Später wurden dann auch, wie bereits erwähnt, computerartige Signalgeber verwendet, die die Vpn mit sich herumtrugen und die nicht nur die (eingespeicherten) Protokolltermine anzeigten, sondern in die auch die Antworten eingegeben werden konnten (BUSE & PAWLIK, 1984, PAWLIK & BUSE, 1982, PERREZ & REICHERTS, 1989). In der vorliegenden Arbeit wurde eine Verfahrensweise angewendet, wie sie bereits in ähnlicher Form bei HORMUTH (1986) und kürzlich auch bei CANTOR und Kollegen (CANTOR, NOREM, LANGSTON, ZIRKEL, FLEESON & COOK-FLANNAGAN, 1991) zum Einsatz gekommen ist: Die Vpn erhielten Armbanduhren, in die zufällig erzeugte Protokolltermine, die den Vpn nicht bekannt waren,

vor Untersuchungsbeginn eingespeichert worden waren. Das Erreichen eines solchen Termins wurde den Vpn durch einen Signalton angezeigt, worauf sie einen Protokollbogen ausfüllen mußten.

Bei diesem Ansatz ist sichergestellt, daß sich die Vpn nicht auf das Eintreten eines Protokolltermins einstellen können. Zudem ist gegenüber den als Funksignalempfängern arbeitenden Piepern gewährleistet, daß sich der Bewegungsspielraum der Vpn nicht auf den Funkempfangsbereich des Signalsenders beschränkt, sondern daß sie im Prinzip weltweit mobil sein können. Eine Armbanduhr ist außerdem weitaus handlicher, im Alltag unauffälliger und möglicherweise auch weniger störanfällig als ein Computer bzw. Verhaltensdatenrecorder. Dieser bietet allerdings den Vorteil, daß ein dynamischer Protokollterminplan (BUSE & PAWLIK, 1990, PAWLIK & BUSE, 1985) implementiert werden kann, der die Zeitdauer der Beobachtungsintervalle in Abhängigkeit von den Eingabeinformationen festlegen kann. Werden zusätzlich auch noch die Items im Anzeigenfeld des Recorders dargeboten (PAWLIK & BUSE, 1992), statt auf einer separaten Itemliste, so besteht ferner die Möglichkeit zu einer selektiven Itemvorgabe, d.h., daß bestimmte Items in Abhängigkeit von den Beantwortungen vorhergehender Items den Vpn gar nicht erst angeboten werden. Es können mit Datenrecordern auch Variablen wie Beginn und Dauer der Dateneingabe oder Latenzzeit zwischen Ertönen des Signals und Beginn der Dateneingabe exakt erhoben werden; Datentransfer per Modem an den Untersuchungsleiter ist ebenfalls möglich, um ein besseres Monitoring der Vpn und eine schnelle Auswertung der Daten zu erreichen (STONE, KESSLER & HAYTHORNTHWAITE, 1991). Überdies können so auch einfache psychophysiologische Variablen erhoben werden, z.B. Herzrate und Fingertemperatur (PAWLIK & BUSE, 1992). Außerdem sind, was das Hauptargument PAWLIKs (1988) gegen das Verwenden von Papier-Bleistift-Protokollbögen anbelangt, bei solchen Datenrecordern Fehlerquellen wie etwa das Abschreiben von bereits ausgefüllten Protokollbögen oder das Ausfüllen mehrerer Protokollbögen gleichzeitig nicht möglich, da solche Geräte Aufzeichnungen nur innerhalb eines bestimmten Zeitraums nach Ertönen des Signals gestatten und die bereits eingegebenen Informationen von den Vpn nicht wieder abrufbar sind. Daß solche Geräte jedoch teurer sind als Armbanduhren und sie außerdem eine intensivere Einweisung der Vpn erforderlich machen können, wurde bereits erwähnt. Das Instrument der vorliegenden Arbeit soll von der Grundidee her auch zu routinediagnostischen Zwecken eingesetzt werden. Für diesen Verwendungsbereich ist der Einsatz von Armbanduhren weitaus realistischer als der von Computern, da sie eben nicht nur billiger sind, sondern vom Untersuchungsleiter auch ein geringeres Know-how verlangen.

Es sei hier allerdings erwähnt, um keinen falschen Eindruck zu erwecken, daß für die vorliegende Untersuchung der Einsatz von Datenrecordern, die im Institut I durchaus vorhanden aber nicht einsatzbereit sind, gar nicht zur Diskussion stand. Und daß Signalgeber in Form von Armbanduhren zur Verfügung stehen würden, stellte sich erst bei Beginn der Vorarbeiten zur Datenerhebung heraus. Wäre für diese Anschaffung am Institut I kein Geld vorhanden gewesen, so wäre diese Untersuchung ähnlich wie die von BRANDSTÄTTER (1983) abgelaufen: die Vpn hätten einen Protokollterminplan mit zufällig ausgewählten Terminen erhalten und hätten für seine Einhaltung selber Sorge tragen müssen. Alternativ zu Datenrecordern könnte auch der Einsatz von Funktelefonen in Betracht kommen. HEDGES, JANDORF & STONE (1985) führten mit ihren Vpn mehrmals am Tag Telefoninterviews zur Erhebung der Stimmungslage. Solange dabei stationäre Telefone

verwendet werden (wie vermutlich bei HEDGES, JANDORF & STONE, 1985, geschehen), liegt keine ESM-Studie vor, weil die Datenerhebung an den Standort des Telefons und an die Situation des Telefongesprächs gekoppelt ist. Funktelefone sind vermutlich nicht viel teurer als Datenrecorder und erlauben neben allen Vorteilen, den diese bieten, auch noch den persönlichen Kontakt zwischen Vpn und Untersuchungsleiter, der für die Compliance der Vpn förderlich ist. Andererseits führen sie evt. einen Untersucher-Bias ein und sind personalaufwendiger in der Durchführung.

Um eine hohe psychometrische Qualität von ESM-Daten zu erreichen, muß sichergestellt sein, daß sie punktuell ein unverfälschtes Abbild des Verhaltensstroms unter Alltagsbedingungen liefern, wie ein Blitzlichtphoto. Dazu ist es notwendig, daß die Vpn möglichst sofort und möglichst auf alle angebotenen Protokolltermine mit wahrheitsgetreuen Antworten reagieren. Zugleich soll der Protokolltermin keine spürbare oder gar folgenschwere Unterbrechung bzw. Veränderung des Verhaltensstroms insgesamt oder des jeweiligen Umfeldes bedingen und es soll insbesondere auch keine systematische Veränderung des zu messenden Verhaltens oder Erlebens durch den Meßvorgang erfolgen, z.B. Absinken der Befindlichkeit während des Ausfüllens eines einzelnen Protokollbogens oder Veränderungen in Mittelwerten oder Varianzen der erhobenen Maße während des gesamten Beobachtungszeitraums infolge Gewöhnung an die Untersuchung. Die Fähigkeit der Vpn die gewünschten Angaben auch korrekt machen zu können, wird dabei vorausgesetzt (HORMUTH, 1986, STONE, KESSLER & HAYTHORNTHWAITE, 1991). Die Vpn sollten so instruiert werden, daß sie Angaben nur für den Zeitpunkt des Bearbeitens des Protokollbogens bzw. für die unmittelbar vorhergehende Zeit machen sollen und nicht für einen längeren Zeitraum oder für länger zurückliegende Ereignisse, um Gedächtniseffekte zu vermeiden. Um Informationen zur Sicherung der psychometrischen Güte der erzeugten Daten zu erhalten, empfiehlt sich der Einsatz eines Fragebogens, in dem die Vpn im nachhinein gebeten werden, Aussagen zum Ablauf der Untersuchung zu machen.

Die Durchführungsobjektivität von ESM-Studien dürfte unterschiedlich hoch sein, abhängig davon, wie gut die Vpn einerseits instruiert sind, d.h. vor allem, wie klar den Vpn die einzelnen Items sind und ob sie genügend Zeit hatten, sich mit dem Erhebungsinstrument vertraut zu machen, und andererseits wie genau sie sich an die Instruktionen halten. Letzteres ist bei Datenrecordererhebungen leichter zu kontrollieren als bei Protokollbögen, denn bei diesen bestehen deutlich mehr Möglichkeiten zur Fälschung der Angaben. Werden offene Fragen vermieden, dürfte die Auswertungsobjektivität sehr hoch sein. Die Interpretationsobjektivität hängt von Normen ab, die auf interindividueller Ebene wohl noch für kein ESM-Instrument vorliegen dürften. Im Falle der vorliegenden Untersuchung ist allerdings eine intraindivuelle Normierung auf Stateebene prinzipiell möglich, indem die intraindividuelle Testwerteverteilung je Vp errechnet werden kann und ein bestimmter Testwert aus dieser Verteilung zu einem bestimmten Protokolltermin unter Berücksichtigung des intraindividuellen Vertrauensintervalls für die Statewerte in Relation zu dieser Verteilung, z.B. in Form von Standardabweichungseinheiten, eingeordnet werden kann.

EPSTEIN (1977, 1979, 1980) konnte zeigen, wie die Reliabilität einzelner Verhaltens- und Befindlichkeitsitems bei Aggregation über mehrere Zeitpunkte mit steigender Zahl der Zeitpunkte zunimmt. Auf der Ebene einzelner Befindlichkeitsitems lassen sich so nach SPEARMAN-BROWN

(vgl. Kap. 2.1.1.2.) aufgewertete Split-Half-Reliabilitätskoeffizienten um R=.8 und höher mühelos erreichen (vgl. auch PAWLIK & BUSE, 1982, PERREZ & REICHERTS, 1989). In der vorliegenden Arbeit werden auf diese Weise Reliabilitätsbestimmungen der Traitwerte vorgenommen, die, ausgehend von den einzelnen Itemwerten, zusätzlich noch pro Zeitpunkt und Vp zu Skalen aggregiert werden. Außerdem werden die in Kap. 2.1.3. beschriebenen Reliabilitätskennwerte für die State- und die Traitwerte errechnet. Auf der Ebene von Skalenwerten werden dann noch pro Vp über alle wahrgenommenen Protokolltermine intraindividuelle (State-) Konsistenzkoeffizienten (vgl. Kap. 2.1.1.2.) berechnet.

BOTTENBERG (1970) zufolge liefert "die verbale Ausformulierung von Stimmung ... den bislang sichersten Ansatzpunkt zu ihrer empirisch exakten Erfassung" (ebd., S. 20). So könnte denn auf die Frage nach der Validität von Befindlichkeitsskalen verzichtet werden, "weil der Zusammenhang zwischen Adjektivindex und Indiziertem evident" (HAMPEL, 1977, S. 54) ist und daher "durch die Verwendung von Stimmungswörtern als Testmaterial eine (psycho) logische Gültigkeit unmittelbar gegeben" (ebd.) ist. Allerdings: "Dermaßen 'triviale' Gültigkeitsbestimmungen genügen ... nicht heutigen Testansprüchen" (ebd.). So berichtet HAMPEL (1977) von Untersuchungen, in denen Befindlichkeitsveränderungen unter Psychopharmakaeinfluß von entsprechenden Stimmungslisten in erwarteter Richtung angezeigt wurden und CSIKSZENTMIHALYI & LARSON (1987) erwähnen ebenfalls Untersuchungen zur "situational validity" (ebd., S. 531), in denen Zusammenhänge in erwarteter Richtung z.B. zwischen ESM-Verhaltensdaten und synchron erhobenen psychophysiologischen Kennwerten gefunden wurden. Es wurden auch Beziehungen zwischen ESM-Daten und Variablen aus Persönlichkeitsfragebögen untersucht (vgl. CSIKSZENTMIHALYI & LARSON, 1987, PAWLIK & BUSE, 1992). Bei einigen Autoren wurden die ESM-Daten zur Validierung von Persönlichkeitsfragebögen benutzt und bei anderen war es genau umgekehrt (HORMUTH, 1986). Im vorliegenden Fall werden Daten aus Persönlichkeitsfragebögen in Zusammenhang gesetzt mit den ermittelten ESM-Traitwerten und es wird überdies eine Übereinstimmungsvalidierung der ESM-Statewerte mit einer vergleichbaren, etablierten State-Skala versucht.

Die Forderung, daß die Protokolltermine in ESM-Studien zufällig über den Beobachtungszeitraum verteilt sein und für die Vpn nicht vorhersehbar sein sollen und daß sie bei Eintreten eines solchen Termins ihre Angaben zur momentanen Situation unverzüglich vornehmen sollen, hat ihre Bedeutung in der Sicherung der beiden ökopsychologischen Gütekriterien einer Untersuchung, der ökologischen Validität und der ökologischen Repräsentativität. Befolgen die Vpn die Instruktionen und bildet der Beobachtungszeitraum tatsächlich einen unverfälschten Auszug aus dem Lebensraum und Verhaltensstrom der Vpn, dann ist der Blitzlichtcharakter der ESM-Protokolle gewährleistet und diese beiden ökopsychologischen Gütekriterien sind erfüllt. Eine Untersuchung ist ökologisch valide, wenn sie in Alltagssituationen gewonnene Daten liefert, und sie ist ökologisch repräsentativ, wenn die gezogene Situationsstichprobe für diese Alltagswelt repräsentativ ist (BUSE & PAWLIK, 1990). Nach PAWLIK (1978) besteht das verhaltenswirksame Biotop einer Vp aus der Menge aller verhaltenswirksamen Stimulusvariablen; diese variieren über die Zeit in Häufigkeit bzw. Intensität ihres Auftretens und bilden somit einen deskriptiven Rahmen zur Beschreibung unterschiedlicher Situationen in der Alltagswelt dieser Vp, die sich zeitlich nacheinander manifestieren. Für jede dieser Stimulusvariablen kann eine Häufigkeitsverteilung bzw. Dichteverteilung ermittelt werden, ihre

Kovarianzen mit anderen solchen Variablen können bestimmt werden. Für eine konkrete Vp ist eine konkrete Erhebungssituation "in dem Ausmaß ökologisch valide, in dem die in ihr enthaltenen Stimulusbedingungen eine unverzerrte Stichprobe der Grundgesamtheit der im Biotop ... repräsentierten Stimulusbedingungen sind" (ebd., S. 124). Daraus folgt, daß ESM-Untersuchungen notwendigerweise ökologisch valide sind (PAWLIK & BUSE, 1982). Die ökologische Repräsentativität bezieht sich demgegenüber auf die Fähigkeit einer Untersuchungsmethode, durch das Ziehen mehrerer ökologisch valider Stichproben aus der Alltagsumwelt einer Vp die tatsächlichen Varianzen und Kovarianzen dieser Stimulusvariablen, wie sie in der Alltagsumwelt über die verschiedenen Situationen hinweg auftreten, adäquat zu erfassen. Je größer die Zahl der Stimulusvariablen ist, je größer ihre Varianzen sind und je weniger sie kovariieren, desto mehr einzelne (ökologisch valide) Stichproben müssen gezogen werden, damit die Untersuchung die Stimulus-Varianzen und deren Kovarianzen unverzerrt ausschöpfen kann und damit sie dadurch ökologisch repräsentativ sein kann. Die ökologische Validität ist somit eine notwendige Voraussetzung für die ökologische Repräsentativität (PAWLIK, 1978, S. 124). Verletzungen der ökologischen Validität können bei ESM-Studien dadurch eintreten, daß während des Beobachtungszeitraums bisher ständig repräsentierte Stimulusvariablen plötzlich völlig unwirksam werden bzw. daß solche Umgebungsvariablen wirksam werden, die im bisherigen Leben der jeweiligen Vp noch niemals wirksam geworden sind und dies aller Voraussicht nach auch nicht wieder sein werden, z.B. Krankenhausaufenthalt nach Verkehrsunfall. Dies wäre kein für die jeweilige Vp typischer Beobachtungszeitraum mehr; er ähnelt dann schon mehr dem aus einer Laboruntersuchung. (Es wäre allerdings ein typischer Zeitraum, wenn genau nur solche Vpn untersucht werden sollen.) Ob der Beobachtungszeitraum in ESM-Studien für die Vpn typisch ist oder nicht, kann durch die Methode selber nicht nachgewiesen werden; es wird lediglich angenommen, daß es so ist. Die ökologische Repräsentativität wird in dem Maße verletzt, in dem die Latenzzeit (Zeit zwischen Ertönen des Signaltons und Beginn des Protokollierens) größer wird und in dem die Zahl der wahrgenommenen Protokolltermine sinkt. Durch beide Phänomene nimmt die Wahrscheinlichkeit der selektiven Protokollierung zu, d.h. daß nur in bestimmten Situationen protokolliert bzw. daß in bestimmten Situationen eben nicht protokolliert wird und daß somit bestimmtes zu protokollierendes Verhalten oder Erleben unter- bzw. überrepräsentiert wird. Zur Informationsgewinnung über mögliche Verletzungen dieser ökopsychologischen Gütekriterien empfiehlt sich ebenfalls der Einsatz eines Fragebogens, in dem die Vpn im nachhinein Aussagen zum Ablauf der Untersuchung machen können.

2.3. Persönlichkeitspsychologische Grundlagen

Obgleich es in dieser Arbeit um grundsätzliche Aspekte der Konstruktion psychologischer Meßinstrumente und um deren Einsatzmöglichkeiten geht, sollen inhaltliche Gesichtspunkte im Zusammenhang mit der Themenstellung nicht gänzlich außer acht gelassen werden, denn schließlich soll auch das hier zu entwickelnde Instrument Informationen über ganz bestimmte psychologische Gegenstandsbereiche liefern.

2.3.1. Persönlichkeit und Eigenschaften

Wie bereits eingangs angedeutet, ist dieses Instrument in seiner Funktion als Befindlichkeitsskala oder Stimmungsliste im Bereich der Persönlichkeitstests anzusiedeln. Mit seiner Hilfe sollen sowohl Statewerte, d.h. Ausprägungen des momentanen Zustands einer Vp, als auch Traitwerte, d.h. zeit- und situationsstabile Eigenschaften der Vp, diagnostiziert werden können (vgl. Kap. 2.1.3.).

"Persönlichkeit" und "Eigenschaft" sind Begriffe der Differentiellen Psychologie, die seit vielen Jahrzehnten z.T. heftig und äußerst kontrovers diskutiert worden sind (vgl. ALLPORT, 1966, AMELANG & BARTUSSEK, 1981, AMELANG & BORKENAU, 1986, BOWERS, 1973, BUSE & PAWLIK, 1984, EPSTEIN, 1977, 1979, EPSTEIN & O'BRIEN, 1985, EYSENCK, 1980, HJELLE & ZIEGLER, 1981, LÖSEL, 1992, MISCHEL, 1968, SCHWENKMEZGER, 1984). Es soll hier nicht der Versuch unternommen werden, diese Debatte umfassend oder gar detailliert im wissenschaftshistorischen Kontext wiederzugeben. Dennoch ist es unvermeidlich, zumindest einige zentrale Zusammenhänge und Standpunkte zu skizzieren.

Einigkeit hinsichtlich des Begriffs "Persönlichkeit" scheint dahingehend zu bestehen, daß es sich hierbei um ein psychologisches Konstrukt von eher hoher Komplexität handelt, das "die Summe der auf menschliches Erleben und Verhalten bezogenen Konstrukte, deren Wechselbeziehungen untereinander und Interaktionen mit organismischen, situativen und Außenvariablen" (AMELANG & BARTUSSEK, 1981, S. 50) darstellt. Es umfaßt also nahezu die gesamte Psychologie. Vor dem Hintergrund einer breiten Literaturübersicht betonen HJELLE & ZIEGLER (1981) das zeitlich Überdauernde dieses Konstrukts. Definitionen von Persönlichkeit haben außerdem noch weitere Merkmale gemeinsam (ebd., S. 7): Persönlichkeit ist etwas, das als ein Abstraktum aus Verhaltensbeobachtungen abgeleitet wird und in dem sich Individuen voneinander unterscheiden. In ihr wird die Einzigartigkeit eines jeden Menschen deutlich. Die meisten Definitionen betonen zudem auch den Prozeßcharakter von Persönlichkeit, also die Möglichkeit einer Veränderung über die Zeit und damit die Bedeutung einer entwicklungspsychologischen oder biographischen Perspektive. Persönlichkeitspsychologie ist ein Teilgebiet der Psychologie, das sich von anderen Fächern (z.B. Biopsychologie, Entwicklungspsychologie usw.) dadurch unterscheidet, daß in ihm versucht wird, die Annahmen, Theorien und Befunde dieser anderen Fächer einschließlich ihrer Wechselwirkungen zu einer Gesamtheit zu integrieren. Das bedeutet, "that no other area of psychology attempts to cover as much territory as the field of personality" (HJELLE & ZIEGLER, 1981, S. 8), denn "the focus of study has been nothing less than the 'total individual'" (ebd.). Daraus wird verständlich, daß es gegenwärtig "no general agreement within the field on a substantive definition of the term 'personality'" (ebd., S. 22) gibt. Üblicherweise wird der Bereich der Persönlichkeitspsychologie aufgesplittet in einen Merkmalsbereich mit Leistungscharakter und in einen ohne Leistungscharakter (AMELANG & BARTUSSEK, 1981). Entsprechendes gilt für die in diesem Zusammenhang verwendeten Meßinstrumente (RAUCHFLEISCH, 1989).

Ob das Verhalten und Erleben durch zeit- und situationsstabile Eigenschaften oder Dispositionen innerhalb des Individuums (die als gedankliche Konstrukte aus dem konkreten, beobachtbaren Verhalten erschlossen werden) bedingt ist oder ob es durch situationsspezifische Einflüsse gesteuert

wird bzw. in welchem Maße beide Faktoren wirksam werden, ist eine alte Streitfrage, an der sich die Ansichten der Persönlichkeitspsychologen von Anfang an geschieden haben (EPSTEIN & O'BRIEN, 1985). Dies verwundert nicht, lehrt doch die Alltagserfahrung, daß gute Gründe für die Stützung jeder dieser beiden Positionen sprechen. Persönlichkeitstheoretiker sind in erster Linie auch Menschen des Alltags und "like the rest of us, they hold divergent views about human nature" (HJELLE & ZIEGLER, 1981, S. 3). Und diese verschiedenen Ansichten spiegeln sich in ihren jeweiligen theoretischen Konzeptionen wider. Jedoch können diese Ansichten auch verzerrt sein, wie bei uns allen, denn die wissenschaftliche Untersuchung des Alltagsgeschehens lehrt, daß u.a. asymmetrische Attributionen bei der Zuschreibung von Verhaltensursachen zu Eigenschaften der Vp bzw. zu Situationsparametern vorkommen können, wie z.B. der fundamentale Attributionsfehler: "Zeigt jemand ein bestimmtes Verhalten, dann neigt er grundsätzlich dazu, es situativen Bedingungen zuzuschreiben. Ein fremder Beobachter desselben Verhaltens dagegen tendiert zu einer Attribution auf die Person des Handelnden" (UPMEYER, 1985, S. 169). Auch werden Persönlichkeitstheoretiker nicht völlig frei sein von aus dem Alltag übernommenen impliziten Persönlichkeitstheorien, die oft mehr Aufschluß geben über den Beurteiler als über die beurteilte Person (vgl. PASSINI & NORMAN, 1966). So liegt es nahe zu vermuten, daß der Streit zwischen Eigenschaftstheoretikern und Situationisten stark durch Ansichten bestimmt ist, die der Lebenserfahrung entstammen, somit eher geglaubt als gewußt und auch nicht immer explizit dargestellt werden.

Traits oder Dispositionseigenschaften sind "relativ breite und zeitlich stabile Dispositionen zu bestimmten Verhaltensweisen, die konsistent in verschiedenen Situationen auftreten" (AMELANG & BARTUSSEK, 1981). Das Problem der Traittheorien ist jedoch, daß eben diese postulierte zeitliche Stabilität und die transsituative Konsistenz empirisch nur schwer (wenn überhaupt) abgesichert werden können. Bei dem Versuch, Verhalten in bestimmten Situationen aufgrund von Verhalten in anderen Situationen oder aufgrund von Traits vorherzusagen, ist (zumindest im Bereich der Merkmale ohne Leistungscharakter) nur selten eine korrelative Beziehung von mehr als r=.3 zu erreichen (EPSTEIN, 1979, S. 1104: "the .30 personality barrier"). Diese Befunde und die mit dem sichtbaren Erfolg der Verhaltenstherapie einhergehende Begeisterung für den ihr zugrundeliegenden "faszinierend untheoretischen Behaviorismus" (HÖRMANN, 1982, S. 173) bestärkten die Kritiker der herkömmlichen Traittheorien darin, die situativen Determinanten des Verhaltens in den Vordergrund zu stellen (vgl. MISCHEL, 1968), d.h. "die verhaltensformende Kraft der Situation" (HÖRMANN, 1982, S. 174) zu berücksichtigen. Um diese zu untermauern, werden fast unausweichlich die klassischen Untersuchungen von HARTSHORNE und Kollegen (HARTSHORNE & MAY, 1928, 1929, HARTSHORNE, MAY & MALLER, 1929, HARTSHORNE, MAY & SHUTTLEWORTH, 1930, ausführlich zit. bei EPSTEIN & O'BRIEN, 1985, EYSENCK, 1980, MISCHEL, 1968, SCHWENKMEZGER, 1984) zur Ehrlichkeit an mehreren tausend Schulkindern angeführt. Für unterschiedliche Verhaltensitems (die unterschiedlichen Situationen entsprachen) ergab sich dort eine durchschnittliche Iteminterkorrelation von lediglich r=.23 (EPSTEIN & O'BRIEN, 1985, S. 522). Allerdings bleibt unklar, wie gering solche Koeffizienten sein müssen, damit der Traitansatz als widerlegt gelten kann, denn auch die Traittheoretiker verkennen nicht völlig, daß das Verhalten in Maßen auch situationsspezifisch sein kann, was BOWERS (1973) veranlaßt, gegen die Situationisten zu polemisieren: "Does trait theory become unraveled because the usually

honest man lies to a rapist concerning the whereabouts of his wife?" (ebd., S. 319) In manchen Fällen ist der Situationismus so weit gegangen, Eigenschaften auf dem Niveau von Adaptationsstufen als "the residual effect of previous stimulation" (HELSON, 1964, zit. bei ALLPORT, 1966, S. 2) zu betrachten oder so viel Verhaltensvarianz wie irgend möglich aus situativen bzw. soziodemographischen Variablen zu erklären, so daß "personality becomes an appendage to demography" (ALLPORT, 1966, S. 2). Es sind Versuche zur Präzisierung des Eigenschaftsbegriffs unternommen worden, indem versucht wurde, die Vorhersagegenauigkeit von Testwerten zu verbessern, und zwar durch Berücksichtigung von Moderatorvariablen, d.h. von solchen Einflußgrößen, die die prädiktive Validität der Testwerte für bestimmte Vpn-Gruppen erhöhen (CONRAD, 1992). BEM & ALLEN (1974) konnten zeigen, daß die Korrelationen zwischen verschiedenen Operationalisierungen (und damit auch verschiedenen Situationen) desselben Konstrukts "Freundlichkeit" höher waren bei solchen Vpn, die sich selbst als (für dieses Konstrukt) eher situationskonsistent eingeschätzt hatten, als bei denjenigen Vpn, die ihr Verhalten eher für inkonsistent gehalten hatten; ein Befund, der von AMELANG & BORKENAU (1986) jedoch nicht repliziert werden konnte. Eine Synthese zwischen Eigenschaftstheoretikern und Situationisten nehmen die Interaktionisten vor, indem sie versuchen varianzanalytisch nachzuweisen, wie etwa BOWERS (1973, S. 320f.) anhand einer Literaturübersicht, daß die durch die Haupteffekte (Vpn und Situationen) aufgeklärten Varianzanteile kleiner sind, als die, die durch eine Interaktion beider Faktoren erklärt werden können. Allerdings: "interactionism proved to be no more successful in breaking the presumed .30 barrier than previous approaches" (EPSTEIN & O'BRIEN, 1985, S. 516).

Der Grund hierfür liegt laut EPSTEIN (1977, 1979, 1980) darin, daß Untersuchungen im Rahmen aller drei genannten Ansätze zumeist von einzelnen Items ausgegangen sind und versucht haben, Verhalten in Einzelsituationen vorherzusagen. Das bedeutet, daß in diesen Studien eine relativ hohe Fehlerkomponente in Kauf genommen worden ist, weil Einzelitems fehlerbehafteter sind als aggregierte Werte, die aus mehreren, das gleiche Konstrukt messenden Items bestehen. Solche aggregierten Werte sind reliabler, da der Meßfehler herausgemittelt worden ist. So steigt der o.a. mittlere Koeffizient der Iteminterkorrelationen bei HARTSHORNE & MAY (1928, 1929, zit. bei EPSTEIN & O'BRIEN, 1985, S. 522f.) unter Berücksichtigung der (z.T. bekannten) Reliabilitäten der Items von r=.23 auf r=.42 und ein Testkompositum aus den der Mittelung zugrundeliegenden neun Einzelitems hätte bei einer mittleren Iteminterkorrelation von r=.23 nach der Beziehung von SPEARMAN-BROWN (LIENERT, 1989, S. 119) schon eine Reliabilität von R=.73 (EPSTEIN & O'BRIEN, 1985, S. 522f.). Die Befunde von HARTSHORNE & MAY (zit. ebd.) lassen sich also auch im Sinne einer Stützung der Eigenschaftstheorien auslegen, denn auch bei Intelligenztests liegen die mittleren Iteminterkorrelationen in dem dort genannten Bereich (EPSTEIN & O'BRIEN, 1985, S. 523). Und EYSENCK (1953, zit. bei SCHWENKMEZGER, 1984, S. 266) merkt dazu an, daß Koeffizienten auf Itemebene im Bereich von .3 die Hypothese der Situationsspezifität des Verhaltens geradezu widerlegen.

Um den Fehleranteil an den Testwerten zu reduzieren und auf diesem Wege Korrelationen von über r=.3 zu erreichen, befürwortet EPSTEIN (1977, 1979, 1980, EPSTEIN & O'BRIEN, 1985) mit Testwerten zu arbeiten, die über Vpn, über Stimuli bzw. Situationen (bzw. Fragebogenitems), über die Zeit (Versuchsdurchgänge) oder über verschiedene Meßmethoden aggregiert werden, je

nachdem, in welchen Bereichen eine Reduktion der unsystematischen Fehlervarianz bzw. der spezifischen Varianz erwünscht ist. Traits sollen weiterhin nicht nur relativ stabile und konsistente, sondern auch relativ breit angelegte Konstrukte sein. Eine Aggregation über verschiedene Stimuli bzw. Situationen hinweg ist deshalb wichtig, weil dies nicht einfach nur zu einer Erhöhung der Reliabilität der Testwerte beiträgt, sondern zugleich auch den Generalisationsbereich des Konstrukts erweitert, denn es werden verschiedene Aspekte desselben Konstrukts in verschiedenen Situationen bzw. durch verschiedene Items gemessen. Gleiches gilt für eine Aggregation über Zeitpunkte, die zwar der Erhöhung der Zeitstabilität dient, zugleich aber auch eine Verbreiterung des Generalisationsbereiches mit sich bringt; denn trotz des Versuchs, die situativen Bedingungen hierbei konstant zu halten, werden verschiedene Zeitpunkte für dieselben Vpn auch immer geringfügig verschiedene Situationen bedeuten. Zur Messung einer Eigenschaft sind demnach wiederholte Messungen desselben Verhaltens bzw. Erlebens in verschiedenen Situationen zu verschiedenen Zeitpunkten erforderlich - wie in ESM-Studien verwirklicht. In ihnen sind Situationen und Zeitpunkte konfundiert (EPSTEIN & O'BRIEN, 1985, S. 517). Bei der Ermittlung der Vorhersagevalidität von aggregierten Testwerten ist darauf zu achten, daß die Aggregation möglichst in gleichem Umfang sowohl auf Prädiktor- wie auf Kriterienebene vorgenommen wird, denn nur so ist "eine faire Beurteilung der Validität möglich" (SCHWENKMEZGER, 1984, S. 266).

Üblicherweise sind Vpn bei der Administration von herkömmlichen Traitmeßinstrumenten (z.B. Persönlichkeitsfragebögen) dazu aufgefordert, Auskunft darüber zu geben, wie sie sich "im allgemeinen fühlen" (Traitfragebogen des STAI, LAUX et al., 1981) oder wie ihr habituelles Verhalten oder Erleben sonst beschaffen ist. Was dabei von den Vpn verlangt wird, ist nichts anderes als ein intuitives Mitteln über viele mental repräsentierte Situationen oder Zeitpunkte. Allerdings ist dieses summarische Urteil nicht unbedingt eine zuverlässige Schätzung, denn HEDGES, JANDORF & STONE (1985) konnten zeigen, daß es schon bei einem kurzen Bezugszeitraum von nur einem Tag erhebliche Verzerrungen dieses Urteils geben kann. Sie ließen ihre Vpn viermal am Tag deren momentane Stimmung schätzen und am Abend die Stimmung für den gesamten Tag. Dieses Gesamturteil ähnelte dem Tagesspitzenwert der Stimmung zu einem der vier Meßzeitpunkte mehr, als dem errechneten Durchschnitt aus allen vier Messungen. UNDERWOOD, FROMING & MOORE (1980) fanden andererseits heraus, daß Vpn durchaus in der Lage sind, ihre typische Trait-Befindlichkeit unter Absehung von ihrer momentanen State-Befindlichkeit zum Zeitpunkt der Administration des Traitinstruments anzugeben. In der vorliegenden Arbeit wird die durchschnittliche intraindividuelle Ausprägung einer Stimmung während des Beobachtungszeitraums als Traitwert betrachtet (vgl. dazu CATTELL, 1967, S. 170), ein Verfahren, das auch HEDGES, JANDORF & STONE (1985) empfehlen: "If mood is sampled randomly within days for a large number of days and averaged, the resulting aggregate variable could be treated at a level of conceptual generality comparable to that of a trait." (ebd., S. 433)

2.3.2. State - Trait Unterscheidung

Der Unterscheidung auf der diagnostischen Ebene zwischen Prozeßdiagnostik auf der einen und Status- bzw. Eigenschaftsdiagnostik auf der anderen Seite steht auf der Ebene psychologischer Konstrukte die Unterscheidung zwischen States und Traits entgegen. Statusdiagnostik ist, wie bereits erwähnt (vgl. Kap. 2.1.2.), auf die Erhebung von zeit- und bedingungskonstanten Informationen im Sinne eines überdauernden Ist-Zustands ausgerichtet, Prozeßdiagnostik hat dagegen das Ziel, Veränderungen in psychologischen Merkmalen festzustellen, mithin Informationen zu liefern über die zeit- und bedingungsvariablen Komponenten dieser Merkmale (JÄGER, 1992, JÄGER & SCHEURER, 1992, PAWLIK, 1982). Eigenschaften, also Traits, ermöglichen als statische Konzepte kein Verständnis von dynamischen Prozessen (RUDINGER, 1985), was an den Schwierigkeiten der Veränderungsmessung im Rahmen der KT deutlich wird (vgl. Kap. 2.1.2.). In den so konstruierten Meßinstrumenten werden die Vpn oft, zumindest im Merkmalsbereich ohne Leistungscharakter, gebeten, ihr typisches Verhalten anzugeben (vgl. Kap. 2.3.1.). In ESM-Datenerhebungen kann das, was empirisch über den Beobachtungszeitraum in den Merkmalen konstant bleibt, als Trait aufgefaßt werden (vgl. Kap. 2.1.3. und Kap. 2.3.1.). In der vorliegenden Arbeit werden zusätzlich diejenigen Merkmalsanteile, die über die Zeit und über die Situationen variieren, jedoch über die beiden Messungen innerhalb eines Meßzeitpunkts konstant bleiben, als State aufgefaßt (vgl. Kap. 2.1.3.).

Die States als psychologische Konstrukte "entsprechen in grober Annäherung dem umgangssprachlichen Stimmungsbegriff; es handelt sich hierbei um temporäre Zustände von Aktivation, Entspannung, guter Stimmung und dergl. ... Zumindest in groben Zügen sollte eine gewisse Entsprechung der State- zu den Trait-Faktoren bestehen, soweit diese vergleichbare Konstrukte betreffen." (AMELANG & BARTUSSEK, 1981, S. 63) Traitfaktoren werden üblicherweise ermittelt, indem mehrere Merkmale zu einem Zeitpunkt bzw. in nur einer Situation an vielen Vpn erhoben werden und indem diese Merkmale faktorisiert werden (R-Technik) (vgl. AMELANG & BARTUSSEK, 1981). In solche Faktorenanalysen fließen allerdings auch Statekomponenten der Merkmale mit ein, denn zu einem Meßzeitpunkt unterscheiden sich die Vpn nicht nur in ihren Eigenschaften, sondern auch durch den jeweiligen Zustand, in dem sie sich zu dem Untersuchungszeitpunkt gerade befinden (vgl. Kap. 2.1.3. und BUSE & PAWLIK, 1991). In der hier durchgeführten Untersuchung werden Traitwerte faktorisiert, die als intraindividuelle Itemmittelwerte vorliegen und dadurch relativ statefrei sind (vgl. (46)). Die Frau von CATTELL (1967) war die erste Vp, die Mitte der vierziger Jahre Daten lieferte, mit denen eine Statefaktorenanalyse möglich war. Nötig sind dafür Angaben von mindestens einer Vp in mehreren Merkmalen und über viele Zeitpunkte bzw. Situationen. Diese Merkmale werden dann über die Zeitpunkte faktorisiert (P-Technik) (vgl. AMELANG & BARTUSSEK, 1981) und es ergeben sich Faktoren, die traitfrei sind (da ja nur eine Vp untersucht wurde). Der Vergleich solcher Statefaktorenlösungen auf der Basis derselben Merkmale aber erhoben an verschiedenen Vpn, zeigt eine hohe interindividuelle Übereinstimmung (CATTELL, 1973). Zu ähnlichen Ladungsmustern wie die P-Technik kommt auch die "incremental R-technique" (CATTELL, 1967) oder "differentielle R-Technik" (AMELANG & BARTUSSEK, 1981, S. 338), bei der die an vielen Vpn in mehreren Merkmalen erhobenen Differenzwerte zwischen zwei verschiedenen Meßzeitpunkten faktorenanalysiert werden, ungeachtet der damit verbundenen Schwierigkeiten, wie sie in Kap. 2.1.2.

geschildert wurden (CATTELL, 1967, 1973). Eine Faktorenanalyse der Statewerte aus der hier durchgeführten ESM-Studie erfolgt als Ketten-P-Technik: Zunächst werden pro Vp für jeden Zeitpunkt und für jedes Item die Abweichungswerte vom jeweiligen intraindividuellen Mittelwert bestimmt, wodurch die Trait-Komponente eliminiert wird. Die so gewonnenen Matrizen mit den intraindividuellen Statewerten aller Items zu allen wahrgenommenen Zeitpunkten werden versuchspersonenweise hintereinander gehängt, so daß unter den Statewerten der Vp 1 im Item x die Statewerte der Vp 2 im selben Item x stehen usw. Es wird erwartet, daß die Ladungsmuster aus den Trait- und den Stateanalysen gleiche Strukturen aufweisen.

Dem Konstrukt der Angst ist auf der State- und auf der Traitebene viel Forschungsarbeit gewidmet worden (vgl. dazu AMELANG & BARTUSSEK, 1981, KROHNE & KOHLMANN, 1990, SCHWENKMEZGER, 1985). Es kann als gesichert gelten, daß sich auf beiden Ebenen Ängstlichkeits- bzw. Angstfaktoren ermitteln lassen (AMELANG & BARTUSSEK, 1981). Nach CATTELL (1979, zit. bei SMITH, 1988, S. 690) ist ein Trait "a relatively stable, enduring personality disposition, described as 'that which defines what a person will do when faced with a defined situation'", wohingegen ein State sein soll "a relatively momentary or temporary behavioral influence which results from a situation and is formally defined as the 'result of the effect of a situation upon a state liability'" (ebd.). Von den Forschungen von CATTELL ausgehend entwickelten SPIELBERGER und Kollegen (SPIELBERGER, 1966, 1972, 1977, SPIELBERGER, LUSHENE & McADOO, 1977) eine State-Trait-Theorie der Ängstlichkeit und darauf aufbauend das STAI (LAUX et al., 1981), das mit zwei verschiedenen Teilinventaren von jeweils zwanzig Items sowohl State- als auch Traitangst erfassen kann. Diese Teilinventare unterscheiden sich nicht nur in den sie konstituierenden Items, sondern auch in den Anleitungen zu ihrer Bearbeitung, die sich im Stateteil auf den Augenblick des Ankreuzens bezieht, im Traitteil dagegen auf die allgemeine Befindlichkeit. Der Trait-State-Theorie der Ängstlichkeit zufolge kann Stateangst verstanden werden als ein "transitory emotional state or condition of the human organism that varies in intensity and fluctuates over time" (SPIELBERGER, LUSHENE & McADOO, 1977, S. 249), während sich Traitangst bezieht auf "relatively stable individual differences in anxiety proneness - that is, to differences in the disposition to perceive a wide range of stimulus situations as dangerous or threatening, and in the tendency to respond to such threats with A[nxiety]-state reactions" (ebd.). Ob sich ein Trait in einer konkreten Situation als beobachtbares Verhalten oder Erleben ausdrückt oder nicht, hängt von der individuellen Stärke des Traits und von der Anwesenheit entsprechender Auslösereize in der Situation ab (SPIELBERGER, 1972). In Abhängigkeit von der Situation erleben Ängstliche somit öfter, intensiver und länger Angst als Nichtängstliche (AMELANG & BARTUSSEK, 1981).

In das ESM-Instrument der vorliegenden Untersuchung ist eine auf achtzehn Items reduzierte Version des STAI-Stateteils integriert, so daß von jeder Vp zu jedem Zeitpunkt ein STAI-Statewert vorliegt. Solche Verkürzungen der Itemzahl führen nicht notwendigerweise zu einer schlechteren psychometrischen Skalenqualität (PLUTCHIK & CONTE, 1989). SPIELBERGER (1972, SPIELBERGER, LUSHENE & McADOO, 1977) zufolge erbringen sogar auf vier oder fünf Items reduzierte Versionen des STAI-Stateteils valide Ergebnisse (vgl. auch HECHELTJEN & MERTESDORF, 1973).

In vorsichtiger Annäherung an SPIELBERGERs theoretisches Konzept soll in dieser Arbeit der State als etwas aufgefaßt werden, das sowohl von der Disposition einer Vp (deren Ausprägung allerdings erst retrospektiv im summarischen Überblick über die einzelnen Meßzeitpunkte erschlossen wird) als auch von den situativen Gegebenheiten in Zeit und Raum im Sinne eines interaktiven Wechselspiels abhängt. Dabei soll das Statekonzept allerdings nicht nur, wie z.B. bei AMELANG & BARTUSSEK (1981) angedeutet (s.o.), auf Befindlichkeitsdimensionen eingeschränkt werden. Die Performance in Fähigkeits- oder Leistungstests unterliegt ebenso zeit- und situationskorrelierten Fluktuationen wie die im Stimmungsbereich. PAWLIK & BUSE (1992) haben in ihren Untersuchungen daher auch Pionierarbeit im Bereich der Feldpsychometrie (Darbietung von Leistungstests unter Feldbedingungen) geleistet und Anwendungen des unter Kap. 2.1.3. beschriebenen State-Trait-Modells auf Meßvariablen der kognitiven Leistung sowie der psychophysiologischen Aktivierung aufgezeigt (BUSE & PAWLIK, 1991).

Andere Autoren haben versucht, State- und Traitwerte durch Administration derselben Items, jedoch mit unterschiedlicher Anleitung der Vpn zu gewinnen. Zum Zwecke der Statediagnostik mußte der Zustand im Moment des Testens angegeben werden, für die Traitdiagnostik die Beschreibung dessen, wie sich die Vpn im allgemeinen fühlten (PATRICK, ZUCKERMAN & MASTERSON, 1974, ZUCKERMAN, 1977). Im Bereich der State-Trait-Angstforschung hat STEYER (1987, MAJCEN, STEYER & SCHWENKMEZGER, 1988) in Erweiterung der KT ein Modell der Varianzzerlegung entwickelt, mit dessen Hilfe auf der Basis von mindestens zwei Messungen zu zwei Meßzeitpunkten mit parallelen Instrumenten (oder mit demselben Instrument, z.B. mit dem STAI) die Testwertevarianz in einen personspezifischen und in einen situationsspezifischen (bzw. in einen durch die Interaktion zwischen beiden bedingten) Anteil zerlegt werden kann. Das Modell ermöglicht die Errechnung entsprechender Konsistenz- und Spezifitätskoeffizienten. Eine Übersicht über weitere Inventare zur Erfassung von State- bzw. Traitangst findet sich bei FRÖHLICH (1983).

2.3.3. Stimmungen

Hinsichtlich einer Definition des Stimmungsbegriffs zeichnet sich in der Literatur ein eher diffuses Bild ab, was dem Wesen der Stimmungen zu entsprechen scheint: "Stimmungen sind Gefühlserlebnisse von diffusem Charakter, in denen sich die Gesamtbefindlichkeit eines Menschen ausdrückt" (EWERT, 1983, S. 399). Sie sind Zustandserlebnisse, die nicht intentional auf Personen, Dinge oder Ereignisse gerichtet sind; Erlebnisse eines Zumuteseins, das nicht über sich hinausweist (ebd., S. 400). Sie "sind durch äußeren Anlaß bedingte oder spontan aufsteigende Gefühlserlebnisse" (HECHELTJEN & MERTESDORF, 1973, S. 111), die mehrere Komponenten enthalten und über einen gewissen Zeitraum fortdauern. BOHNER, HORMUTH & SCHWARZ (1991) betonen das Vorübergehende der Stimmungen und damit ihre Veränderbarkeit über die Zeit. HAMPEL (1977, S. 45) sieht in ihnen "allgemeine, temporäre Verhaltensdispositionen ..., die als Informationsquelle über den augenblicklichen Zustand des Organismus fungieren". "Stimmung dient", laut BOTTENBERG (1970, S. 18), "experimenteller Forschung als operative Formulierung einer ... Größe, die es gestattet, ... die Wirkung ... zu antizipieren, welche diverse Reizvariablen wie etwa emotionelle Belastungen ..., Schlafentzug und -qualität ..., pharmakologische Manipulationen ... und psychotherapeutische Behandlung ... auf physiologische-, Erlebnis- und Verhaltens-Reaktionen ausüben ... . Ein

theoretisches Konzept von Stimmung bleibt", so BOTTENBERG (ebd.) weiter, "auf vorläufige, wenig organisierte Merkmale beschränkt, die zum Teil aus offensichtlicher Konsistenz von Erfahrungen, aus experimentellen Befunden ... hergeleitet werden, zum Teil auf Gedanken zur Thematik von Stimmung zurückgehen ... ." EWERT (1983) überträgt die Figur-Grundunterscheidung auf Gefühlserlebnisse, wonach Stimmungen den Grund bilden, von dem sich "klarer umschriebene Erlebnisse als 'Figur' abheben" (ebd., S. 399), "und zwar immer schon unter dem Aspekt und der Färbung einer jeweils vorherrschenden Stimmung" (ebd., S. 400). So liefern Stimmungen für alle anderen Erlebnisse einen Bezugsrahmen, von dem her diese interpretiert werden. Solche Erlebnisse können Gefühle im engeren Sinne, wie z.B. Zorn, Liebe, Trauer usw., sein, die sich dann als Figuren gegen den Hintergrund der Stimmungen abheben (EWERT, 1983).

Bezüglich der Meßbarkeit von Stimmungen bemängelt HAMPEL (1977, S. 43): "Im deutschen Sprachbereich gibt es bislang kaum Instrumente, mit denen man emotionale Erscheinungen wie Stimmungen erfassen kann; es gibt auch weder eine formulierte Theorie der Stimmungen, die für eine experimentelle Stimmungsforschung oder Testkonstruktion richtungsweisend sein könnte, noch ist in einem systematischen Rahmen versucht worden, den Begriff und das Wesen der Stimmung zu klären." Er unterscheidet drei empirische Zugänge zum Stimmungsgeschehen: die Messung physiologischer Korrelate der Stimmungen, die Fremdbeobachtung von Ausdruckserscheinungen der Stimmungen sowie die Selbstbeobachtung des Stimmungserlebens (ebd.). Diese Aufgliederung deckt sich im wesentlichen mit der von PLUTCHIK & CONTE (1989). Zwar wird die subjektive Stimmungsbeschreibung als Ausdruck der Selbstbeobachtung wegen ihres introspektiven Charakters und ihrer damit angeblich verringerten Wissenschaftlichkeit beargwöhnt, doch stellt sie andererseits die wichtigste und differenzierteste Informationsquelle über das Stimmungsgeschehen dar (HAMPEL, 1977). Stimmungsbeschreibung als Ausdruck der Selbstbeobachtung erfolgt zumeist verbal, ist also an das Medium der (Alltags-) Sprache gebunden. "Die Sprache kann per se als ausgezeichneter Indikator für emotionale Zustände fungieren", befindet HAMPEL (1977, S. 44), da Forschungen zum semantischen Differential gezeigt haben, daß der semantische Raum im Grunde ein emotionaler Raum ist (ebd.). Eine ähnliche Auffassung vertritt auch, wie bereits erwähnt, BOTTENBERG (1970, S. 20) und andere dort zitierte Autoren. Um die Nachteile der freien Beschreibung des Stimmungserlebens durch die Vpn zu vermeiden (z.B. mangelnde interindividuelle Vergleichbarkeit), ist es in der Stimmungsforschung üblich geworden, "dem Berichterstatter ein sprachlich nicht zu anspruchsvolles, für den Beurteilungsgegenstand annähernd repräsentatives und einfach handhabbares Beschreibungs- und Beurteilungssystem in Form einer standardisierten Adjektivliste anzubieten" (HAMPEL, 1977, S. 44). Es sind, besonders im englischen Sprachraum, eine Vielzahl von solchen Stimmungslisten oder Mood Adjective Check Lists (MACL) entwickelt worden. (Übersicht über einige von ihnen z.B. bei PLUTCHIK & CONTE, 1989, S. 240f.) "The MACL is based on one limited but easily accessible type of behavior, namely, the tendency of persons in our culture to apply to mood certain adjectives which complete the sentence 'I feel ----'" (NOWLIS, 1965, S. 353). HAMPEL (1977, S. 44) listet die Vorteile der MACLs auf: "Solche Adjektivlisten werden erstens dem mehrdimensionalen Charakter der Stimmung gerecht; sie bieten zweitens Material an, welches zum festen Erfahrungsschatz eines jeden Individuums gehört, und sie gestatten drittens eine Skalierung, somit intra- wie interindividuelle Vergleichbarkeit". Bei näherem Hinsehen kann jedoch bestenfalls der zweite Punkt kommentarlos hingenommen werden. Bezüglich der anderen beiden

Aspekte muß festgestellt werden: Stimmung ist zwar ein mehrdimensionales Konstrukt und eine Skalierung der Adjektive ermöglicht eine Vergleichbarkeit der Vpn - aber nur im Rahmen eines einzigen Instruments. Denn inwiefern die Auswahl bestimmter Items (Adjektive) für eine konkrete Stimmungsliste der Struktur von Stimmung wirklich gerecht wird, ob und in welchem Maße eine bestimmte Art der Skalierung der Items die erhobenen Daten und damit auch die Struktur des auf ihrer Grundlage errechneten Faktorraums beeinflußt oder ob eine Vergleichbarkeit der Positionen der Vpn zueinander auf dem Merkmalskontinuum auch über eine einzelne Stimmungsliste hinaus gegeben ist, bleibt unklar. In der konkreten Ausgestaltung nämlich unterscheiden sich die einzelnen Stimmungslisten z.T. ganz erheblich voneinander, was zu gravierenden Unterschieden in den Reaktionen der Vpn auf die Items bzw. zu sehr unterschiedlichen Faktorlösungen führen kann. So sind in einigen Stimmungslisten die Adjektive, wie bei NOWLIS (1965, S. 353, s.o.) angedeutet, in die Form einfacher Hauptsätze gekleidet (z.B. LAUX et al., 1981 oder auch im Instrument der vorliegenden Untersuchung), in anderen Stimmungslisten liegen die Adjektive ohne Satzeinbettung vor (z.B. BÜNNING, 1984, JANKE & DEBUS, 1978). Zum Teil ist den Adjektiven ein dichotomes ("trifft zu" - "trifft nicht zu") Antwortformat zugeordnet (z.B. JANKE & DEBUS, 1978), z.T. eine mehrstufige Skala (z.B. LAUX et al., 1981 oder auch im Instrument der vorliegenden Untersuchung), z.T. eine stufenlose graphische Skala (z.B. BÜNNING, 1984); manche Skalen sind deutlich symmetrisch (z.B. LORR & SHEA, 1979, MEDDIS, 1972), manche anderen z.T. erheblich unsymmetrisch (z.B. HECHELTJEN & MERTESDORF, 1973). Manchmal liegt pro Skala nur ein Adjektiv bzw. Satz vor, dem mehr oder weniger zugestimmt werden soll (z.B. BÜNNING, 1984, LARSEN & KASIMATIS, 1991, LAUX et al., 1981 oder auch im Instrument der vorliegenden Untersuchung), manchmal ist eine Skala bipolar verankert durch zwei Adjektive, von denen der Konstrukteur des Instruments annimmt, daß sie antonyme Begriffe seien (z.B. ULLRICH DE MUYNCK & ULLRICH, 1981); dann wieder gibt es Autoren, die davon ausgehen, daß die Antworten auf ein semantisch negatives Wort (z.B. "traurig") oder einen grammatisch negativen Satz (z.B. "ich bin nicht glücklich") den Komplementärantworten auf die entsprechenden positiven Itemformulierungen entsprechen, was durchaus nicht zutreffend sein muß (vgl. PLUTCHIK & CONTE, 1989). "It should be recognized that semantic opposites need not be psychological opposites" (LORR & SHEA, 1979, S. 472). Von den allgemeinen Schwierigkeiten der Formulierung, Polung und Selektion von Testitems abgesehen, scheint die Struktur des Antwortformats der Stimmungsitems ein zentrales Problem zu sein. Sehr häufig werden hierbei Ratingskalen eingesetzt, um der Vp die Möglichkeit zur Abgabe eines differenzierten Urteils zu geben. Bei der Konstruktion und der Anwendung solcher Ratingskalen sind mindestens die folgenden Punkte als potentielle Fehlerquellen zu beachten: das Skalenniveau der durch sie generierten Daten, die Anzahl und die Gestaltung der Antwortstufen.

Ratingskalen "zählen zu den in den Sozialwissenschaften am häufigsten verwendeten, aber auch umstrittensten Erhebungsinstrumenten" (BORTZ, 1984, S. 118); McCLELLAND (1959, zit. bei AMELANG & BARTUSSEK, 1981, S. 150) hält sie gar für das "größte Unglück der Persönlichkeitsforschung". Mit ihrer Hilfe werden "den Beobachtungsdaten Zahlen zugeordnet, ohne zu berücksichtigen, ob die Relationen im Zahlenbereich den Relationen im empirischen Relativ entsprechen" (HEIDENREICH, 1987, S. 420). Schwierig ist daher die Bestimmung des Skalenniveaus, wie schon in Kap. 2.1.1.3. angedeutet. Zwar lehrt die Erfahrung, "daß das empirische

Relativ, auf welches man sich bei der Messung von Stimmungen stützen kann, nur eine Ordinal-Relation enthält" (HAMPEL, 1977, S. 53), doch die Pragmatiker unter den empirisch arbeitenden Psychologen vertreten den Standpunkt, daß die Verletzungen der Intervallskaleneigenschaften der mit Ratingskalen gewonnenen Daten nicht so gravierend seien, als daß man bei ihrer Analyse auf den Einsatz parametrischer Prüfverfahren und anderer Rechenoperationen, die nur für Daten mit mindestens Intervallskalenniveau geeignet sind, verzichten müßte (BORTZ, 1984). Meine eigene Erfahrung im Bereich der Datenanalysen hat gezeigt, daß die vergleichende Anwendung eines parametrischen Prüfverfahrens und eines entsprechenden non-parametrischen Verfahrens auf dasselbe mit Hilfe von Ratingskalen gewonnene Datenmaterial (z.B. t-Test versus U-Test) sehr oft zu ähnlichen Ergebnissen führt (bezogen auf die Auftretenswahrscheinlichkeit der jeweils errechneten Prüfgröße unter den Stichprobenbedingungen bei Gültigkeit der Nullhypothese). Im Rahmen der vorliegenden Arbeit sind daher sowohl solche Rechenverfahren angewandt worden, die intervallskalierte und gegebenenfalls auch normalverteilte Daten voraussetzen, als auch solche Prozeduren, die lediglich auf die ordinalen Informationen der Daten zurückgreifen.

Unabhängig von diesen eher grundsätzlichen Fragen nach der Qualität der Daten sind die Anzahl der Antwortkategorien und gegebenenfalls ihre inhaltlich-verbalen Definitionen von Belang, da durch deren Variation die Faktorenstruktur der Untersuchungsergebnisse verändert werden kann. Das vergleichsweise kleinere Problem scheint dabei die Anzahl der Skalenstufen zu sein, deren Variation hinsichtlich der Reliabilität und der Validität einer Ratingskala eher unerheblich ist (BORTZ, 1984, S. 123). Um die Vpn, insbesondere solche aus nicht-akademischen Stichproben, hinsichtlich ihres Differenzierungsvermögens nicht zu überfordern, werden fünfstufige Antwortskalen empfohlen (HEIDENREICH, 1987); die in den Protokollbögen der vorliegenden Arbeit verwendete Ratingskala hat (aus Gründen, die später (vgl. Kap. 3.2.2.2.) erläutert werden) vier Antwortstufen. Nicht immer muß eine Bereitstellung von mehreren Antwortkategorien im Sinne abgestufter Ratingurteile eine Hilfe zur besseren Urteilsdifferenzierung der Vpn sein: AMELANG & BARTUSSEK (1981) sehen darin vor allem einen "Kunstgriff des Untersuchungsleiters" (ebd., S. 148), der eine hinreichend große Varianz in den Variablen sichern soll, die jedoch wesentlich nicht nur Ausdruck einer wahren Merkmalsvarianz ist, sondern auch durch interindividuell unterschiedliche Bevorzugung extremer Urteilskategorien zustande kommen kann.

BÜNNING (1984) konnte zeigen, daß sich die Faktorenstruktur des Stimmungsraumes entscheidend verändert, wenn die Skalierung geändert wird. In einer ähnlichen Studie wie der vorliegenden ließ er die Stimmung seiner Vpn für die Dauer von zwei Wochen täglich zweimal an Hand einer 33 Items umfassenden Adjektivliste einschätzen. Er aggregierte die aus den stufenlosen graphischen Antwortskalen ermittelten Einzelmessungen jedes Items bei allen Vpn zu intraindividuellen Mittelwertprofilen und verglich die aus ihnen errechnete Trait-Faktorenlösung (R-Technik) mit dem Ergebnis einer Faktorenanalyse, bei der zwar dieselben Ursprungsdaten verwendet wurden, allerdings in künstlich dichotomisierter Form. Wurden die aggregierten, nicht-dichotomisierten Rohdaten faktorisiert, so ergaben sich drei bipolare und zwei unipolare Faktoren. Wurden die Daten dagegen vor dem Aggregieren dichotomisiert, so ergab die Trait-Faktorenanalyse acht unipolare Faktoren. An diesem Vorgehen ist zwar zu kritisieren, daß die Dichotomisierung ex post facto erfolgte und daß somit keineswegs sicher ist, ob dieselben Vpn bei der tatsächlichen Vorgabe eines dichotomen

Antwortformats nicht vielleicht ein ganz anderes Antwortverhalten gezeigt hätten. Es liegen aber Hinweise darauf vor, daß ein vergleichbar strukturierter Faktorraum (mit ausschließlich unipolaren Faktoren) aufgespannt wird, wenn mit echt-dichotomen Antworten z.B. aus Datenrecordererhebungen gerechnet wird (BÜNNING, 1984).

Abgesehen von diesem Spezialfall eines zweistufigen Antwortformats stellt sich bei mehrstufigen, verbal verankerten Ratingskalen die Frage nach der Formulierung dieser Antwortkategorien, denn die scheint erheblich mehr Einfluß auf die Zahl und Polarität der ermittelten Faktoren zu haben als die Anzahl der Stufen allein. So benutzten etwa NOWLIS (1965) und auch THAYER (1967, zit. bei MEDDIS, 1972, S. 180) eine vierstufige Antwortskala mit den Bezeichnungen "definitely not", "cannot decide", "feel slightly" und "definitely feel", denen die Skalenwerte eins bis vier zugeordnet waren. NOWLIS (1965) erhielt auf diesem Wege anstelle der von ihm hypothetisierten vier bipolaren Stimmungsfaktoren zwölf unipolare Faktoren. Diese Unipolarität von Stimmungsfaktoren "suggests that aspects of mood commonly believed to be interdependently opposed to each other may be functionally independent" (NOWLIS, 1965, S. 360), ermöglicht also die Vorstellung, daß eine Vp z.B. müde und schläfrig und zugleich wach und voller Elan sein kann und nicht entweder im einen Zustand oder im anderen ist. Diese der Alltagserfahrung und eben auch NOWLIS' (1965) eigenen Hypothesen widersprechende Befundlage ist von vielen anderen Autoren bestätigt worden (vgl. dazu MEDDIS, 1972, RUSSELL, 1979). Die beiden letztgenannten Autoren weisen jedoch darauf hin, daß das Auffinden von fast ausschließlich unipolaren Stimmungsfaktoren durch ein methodisches Artefakt zustande gekommen sein kann. Um bipolare Faktoren zu erhalten, müssen nämlich in der Iteminterkorrelationsmatrix neben positiven Korrelationen auch negative Korrelationen in nennenswerter Anzahl und Höhe zu finden sein. MEDDIS (1972) kritisiert an der von NOWLIS (1965) verwendeten Ratingskala, daß sie zum einen nicht einmal richtige Ordinaldaten liefert und zum anderen unsymmetrisch ist. Zweifel an der Datenqualität nährt die Kategorie "cannot decide", der der Wert zwei zugeordnet ist, obwohl unklar ist, ob diese Kategorie wirklich "midway between yes and no" (MEDDIS, 1972, S. 180) ist, denn eine solche Antwort kann neben einer mittleren Einschätzung auch heißen, daß die Vp das betreffende Stimmungsitem als irrelevant ansieht, seine Bedeutung nicht versteht o.ä. Deshalb sollte diese Kategorie eher als fehlender Wert denn als eine (gültige) mittlere Einschätzung aufgefaßt werden. Die Skala von NOWLIS (1965) ist weiterhin unsymmetrisch, indem sie zwei abgestufte Antwortkategorien für die Zustimmung zum jeweiligen Item, aber nur eine für seine Ablehnung anbietet. Dadurch können negative Iteminterkorrelationen abgeschwächt oder unterdrückt werden, denn "an increase in the scale for 'happy' from 'slightly' to 'definitely' cannot be matched by a decrease in the score for 'sad' from 'no' to 'definitely no', because there is no such second category" (MEDDIS, 1972, S. 180). Und in der Tat konnte MEDDIS (1972) zeigen, daß bei Verwendung einer symmetrischen Kontrollskala (bei der den Werten von eins bis vier die Kategorien "definitely do not feel", "do not feel", "feel slightly" und "definitely feel" zugeordnet waren) signifikant mehr substantiell negative Korrelationen, aber auch mehr substantiell positive Korrelationen zwischen den Items auftraten als bei der Verwendung von NOWLIS' (1965) Skala. Zudem reduzierte sich die Anzahl der Korrelationen, die um null herum lagen. Als Konsequenz zeigte sich eine Zunahme der negativen Itemladungen in den extrahierten Faktoren, also eine verstärkte Bipolarität. Zu einem ähnlichen Ergebnis kommt auch RUSSELL (1979), der die Skalen von MEDDIS (1972) und von NOWLIS

(1965) mit noch zwei weiteren Antwortformaten im Vergleich testete. Auch hier traten bei Verwendung von MEDDIS' (1972) Skala die meisten negativen Korrelationen auf, während bei den anderen Skalen die entsprechenden Korrelationskoeffizienten "shifted in the positive direction - away from evidence of bipolarity" (RUSSELL, 1979, S. 349).

RUSSELL (1979) führt noch weitere mögliche Fehlerquellen auf, die zu einer Verringerung oder Verflachung der negativen Korrelationen auch in solchen Studien führen können, die nicht das von NOWLIS (1965) verwendete Antwortformat benutzen. Dazu gehören u.a. der "proximity error" (RUSSELL, 1979, S. 347), bei dem Items, die in raum-zeitlicher Nähe zueinander vorgegeben werden, unter Absehung ihrer Inhalte ähnlich beantwortet werden, oder eine Itemauswahl seitens des Untersuchers, bei der das eine Ende eines potentiell bipolaren Kontinuums zahlenmäßig unterrepräsentiert ist, oder Response-Sets wie z.B. Akquieszenz. Auch LORR & SHEA (1979) konnten einen Einfluß von Antwortstilen der Vpn einerseits und von der Struktur der verwendeten Ratingskala andererseits auf die Polarität der ermittelten Faktoren aufzeigen: Eliminierung des Response-Sets oder Verwendung einer vierstufigen, streng symmetrischen Antwortskala ("definitely do not feel", "probably do not feel", "probably feel", "definitely feel") anstelle einer fünfstufigen, in eine Richtung ansteigenden Skala ("not at all", "a little", "moderately", "decidedly", "extremely") resultierten in einer Zunahme der Bipolarität der Faktoren.

BUSE & PAWLIK (1991) beobachteten bei ihren Befindlichkeitsitems, die auf einer sechsstufigen Antwortskala eingeschätzt werden mußten, daß zwei Items, die inhaltlich beide entweder positive oder negative Befindlichkeiten indizierten, höher positiv miteinander korrelierten als zwei Items miteinander negativ korrelierten, die hypothetisch antonyme Begriffe bildeten. Sie führten dies auf ein unterschiedliches Antwortverhalten der Vpn zurück; demnach neigten einige Vpn dazu, unter Absehung der Iteminhalte generell etwas höhere, andere dagegen eher etwas niedrigere Werte anzugeben, als es ihrem (im testtheoretischen Sinne) wahren Befinden entsprach. Hierdurch wurde eine zusätzliche Quelle der Kovariation eingeführt, die zu einer generellen Erhöhung der Korrelationskoeffizienten führte: positive Koeffizienten verschoben sich in Richtung eins, negative in Richtung null - mit dem gleichen Effekt einer erschwerten Identifizierung möglicher bipolarer Stimmungsfaktoren.

Selbstverständlich bewirkt eine Beseitigung einiger methodischer Mängel in den Datenerhebungsverfahren nicht, daß plötzlich alle aufgefundenen Stimmungsdimensionen deutlich bipolarer Art wären, doch von der Tendenz her scheint es so zu sein, daß (nahezu) rein unipolare Faktorenlösungen (z.B. bei NOWLIS, 1965) eher artefaktbedingt sind und daß bipolare Dimensionen die Wirklichkeit des alltäglichen Stimmungsgeschehens auch besser wiedergeben können. Später wird auf dieses Problem noch zurückzukommen sein (vgl. Kap. 3.5.1.2.).

3. Empirischer Teil

In diesem Abschnitt wird die der Arbeit zugrundeliegende Untersuchung beschrieben. Sämtliche für die Themenstellung bedeutsamen Ergebnisse und aus ihnen abgeleitete Maßnahmen werden in den Kap. 3.4. bis 3.6. dargestellt. Darüber hinaus sind im Anhang alle im Rahmen dieser Arbeit ad hoc konstruierten Meßinstrumente sowie die relevanten Auswertungsergebnisse für die einzelnen Items bzw. Merkmalsskalen aller überhaupt eingesetzten Meßinstrumente und außerdem noch weitere Untersuchungsparameter detailliert dokumentiert.

Der Anhang gliedert sich in drei Teile: Anhang A beinhaltet die Materialien und Ergebnisse aus der Testeinweisung und Vorbefragung, die der eigentlichen Felduntersuchung vorausging. Anhang B umfaßt Materialien und Ergebnisse aus der eigentlichen Felduntersuchung (Beobachtungszeitraum), Anhang C Materialien und Ergebnisse aus der Nachbefragung, die im Anschluß an die Felduntersuchung durchgeführt wurde.

3.1. Übersicht

Unter den Bedingungen einer ESM-Studie sollten Befindlichkeitsmessungen im Feld vorgenommen werden. Außerdem war anschließend eine Itemanalyse zur Konstruktion eines Instruments zur Messung der Befindlichkeit durchzuführen. Die dafür entwickelten und eingesetzten Protokollbögen enthielten einen Pool von zehn verschiedenen, für diese Untersuchung relevanten Items, die innerhalb jedes Protokollbogens doppelt vorgegeben wurden. Damit sollte am Beispiel dieser zwei mal zehn Items eine Dokumentation empirisch gewonnener Parameter für das zugrundegelegte testtheoretische Modell (vgl. Kap. 2.1.3.) ermöglicht werden. Aufbauend auf diesen Parametern und unter Zuhilfenahme weiterer aus den Ursprungsdaten extrahierter Informationen sollten aus den zwei mal zehn Items diejenigen ausgewählt werden, die sich am besten zur Konstruktion einer oder mehrerer änderungssensitiver State-Trait-Kurzskalen zur Messung der Befindlichkeit eignen. Auch für diese Kurzskalen sollten die testtheoretischen Modellparameter bestimmt werden sowie zusätzlich die Testgütekriterien überprüft werden. Weiterhin sollten die Qualität des ESM-Charakters der Untersuchung und die Möglichkeiten des Einsatzes eines so konstruierten Instruments erkundet werden (vgl. Kap. 1.2.).

Wegen des explorativen und damit hypothesenerkundenden Charakters der Untersuchung lassen sich keine übergeordneten, zu prüfenden Hypothesen formulieren, anhand derer über ein Gelingen der Untersuchung abschließend befunden werden könnte. Dennoch wurden im Zuge der Datenanalyse viele hundert statistische Signifikanztests durchgeführt, von denen jeder einzelne selbstverständlich eine konkrete zugrundeliegende Hypothese überprüft. Diese Hypothesen werden im jeweiligen Textzusammenhang expliziert oder auch einfach implizit bei der Darstellung der Signifikanzprüfung vorausgesetzt.

3.1.1. Untersuchungsablauf

Der Beobachtungszeitraum umfaßte sieben Tage. Täglich waren sechs Protokolltermine vorgesehen (insgesamt also 42), die sich über einen Zeitraum von zwölf Stunden je Tag so verteilten, daß in jedem Zweistundenintervall (innerhalb dieser zwölf Stunden) genau ein Protokolltermin lag. Dabei wurde ein Termin innerhalb eines Zweistundenintervalls grundsätzlich nach Zufall plaziert. Nur wenn der Abstand zwischen zwei Protokollterminen durch die Randomisierung zu kurz gewesen wäre, wurde in die Terminplanung insoweit eingegriffen, als daß ca. 45 Minuten Minimalabstand zwischen zwei Terminen gewährleistet wurden. Es gab also keine festen, sondern variable Beobachtungsintervalle. Der Zwölfstundenzeitraum wurde so gewählt, daß er sich mit der Wachphase der Vpn deckte. (Eingesetzt wurden Signalgeber, die täglich von 9.00h bis 21.00h, von 10.00h bis 22.00h oder von 12.00h bis 24.00h weckten.) Einen Tag vor Beginn des Beobachtungszeitraums fand eine Testeinweisung und Vorbefragung statt. In ihr wurde der gesamte Untersuchungsablauf sowie die Bedienung des Signalgebers erläutert und das Ausfüllen der Protokollbögen geübt. Einer von diesen Protokollbögen, die zum Zwecke des Vpn-Trainings während der Testeinweisung bearbeitet wurden, ging als erster gültiger Protokolltermin in die Auswertung mit ein, so daß am siebten Tag des Beobachtungszeitraums nur noch fünf Protokolltermine vorgegeben werden mußten, um insgesamt auf 42 Termine zu kommen.

Als Signalgeber wurde eine handelsübliche Armbanduhr mit Digitalanzeige (CASIO Multi Planner, Module No. 931) verwendet, die u.a. die Möglichkeit bietet, maximal sechzig voneinander unabhängige Wecktermine einzuspeichern. Zu jedem Wecktermin war die Vp aufgefordert, möglichst sofort, spätestens jedoch nach Ablauf von dreißig Minuten, einen Protokollbogen zu bearbeiten, der Items enthielt, die sich auf die Tätigkeit unmittelbar vor der Protokollierung sowie auf das Setting und die Stimmungslage während des Ausfüllens bezogen. (War ein Ausfüllen innerhalb der dreißig Minuten nicht möglich, sollte der Protokolltermin übersprungen werden.) Auf jedem Protokollbogen waren zehn Stimmungsitems, die im Rahmen dieser Untersuchung besonders relevant sind, doppelt vorgegeben. Das Antwortformat für alle Stimmungsitems war eine vierstufige Ratingskala. Jeder Vp stand ein Ringbuch im DIN-A5-Format mit insgesamt 45 Protokollbögen zur Verfügung. Jeder einzelne Protokollbogen bestand aus drei DIN-A5-Seiten. Alle Protokollbögen waren einander grundsätzlich gleich, lediglich die zwei mal zehn relevanten Stimmungsitems waren auf jedem Protokollbogen in einer zufällig anders rotierten Reihenfolge vorgegeben. Der erste dieser Protokollbögen diente im Rahmen der Testeinweisung lediglich zu Übungszwecken, 42 weitere waren für die Protokolltermine und zwei als Ersatz (für evt. unbrauchbar gewordene Protokollbögen) vorgesehen. Die Vpn mußten also ständig während des gesamten Beobachtungszeitraums (genauer: täglich während des in Frage kommenden Zwölfstundenzeitraums) Signalgeber und Ringbuch mit sich herumtragen; sie sollten sich ansonsten genauso verhalten, wie sie es ohne Teilnahme an der Untersuchung getan hätten. Die bereits ausgefüllten Protokollbögen mußten während des Beobachtungszeitraums von einem Teil der Vpn zweimal an den Untersuchungsleiter (in allen Fällen war ich dies selbst) zurückgegeben werden.

Die Testeinweisung und Vorbefragung, die dem Beobachtungszeitraum unmittelbar vorausging, umfaßte die Instruktion der Vpn hinsichtlich Untersuchungsablauf, Bedienung des Signalgebers und

Handhabung der Protokollbögen. Außerdem fand eine Erhebung soziodemographischer und persönlichkeitspsychologischer Basisdaten statt.

Möglichst bald im Anschluß an den Beobachtungszeitraum erfolgte eine Nachbefragung, in der die Vpn Gelegenheit hatten, ausführlich zu der Felduntersuchung Stellung zu nehmen. Eingesetzt wurde zu diesem Zweck ein selbst konstruierter Fragebogen, den die Vpn bearbeiteten. Er enthielt sowohl offene als auch geschlossene Fragen. Vorher wurden alle noch nicht zurückgegebenen Materialien (Signalgeber, Ringbuch, die restlichen ausgefüllten sowie alle leeren Protokollbögen) eingesammelt. Abschließend wurden die Vpn über das Untersuchungsziel aufgeklärt und erhielten entweder ein kleines Präsent als Anerkennung für ihre Mühe oder es wurden ihnen statt dessen Vp-Stunden bescheinigt.

3.1.2. Abweichungen von der Planung

Die Realisierung der in der Planungsphase entwickelten Untersuchungskonzeption erwies sich während der Durchführungsphase als weitaus unkomplizierter als ursprünglich angenommen.

Insbesondere ließen sich die Vpn ziemlich leicht rekrutieren, so daß insgesamt 74 Vpn zur Teilnahme bewogen werden konnten (anstatt der als Minimum angestrebten Zahl von 40 Vpn). Außerdem war zwar eine Felduntersuchung mit zufällig ausgewählten Protokollterminen vorgesehen, jedoch war zunächst beabsichtigt, diese Termine den Vpn bei Untersuchungsbeginn lediglich mitzuteilen; für deren Einhaltung wären die Vpn selber verantwortlich gewesen. Tatsächlich aber standen unerwartet elektronische Signalgeber in hinreichender Stückzahl für den gesamten Erhebungszeitraum zur Verfügung, so daß die Felduntersuchung als echte ESM-Studie ablaufen konnte.

In allen übrigen Punkten gab es keine Abweichungen vom Planungskonzept.

3.2. Methode

3.2.1. Stichprobe

An dieser Untersuchung konnte im Grunde jeder teilnehmen, der über minimale kognitive und emotionale Qualitäten verfügte und darüber hinaus eine gewisse Grundzuverlässigkeit besaß. Konkret mußte jede Vp in der Lage sein, die Aufgabenstellung zu verstehen und ihre momentane Befindlichkeit mit Hilfe der Protokollbögen zu dokumentieren. Das heißt auch, daß sie in der Lage sein mußte, überhaupt in Kontakt mit ihrer Emotionalität treten zu können. Sie sollte auch bereit sein, unverfälscht über sich Auskunft zu geben und die Testinstruktionen zu befolgen.

Der Beobachtungszeitraum sollte für die Vpn repräsentativ sein, d.h. er sollte einen für sie typischen Alltagszeitraum umfassen, in dem keine außergewöhnlichen Situationen (wie z.B. Krankheit, Urlaub, Weihnachtsfest usw.) auftraten (vgl. Kap. 2.2.3.).

Zur Rekrutierung der Vpn wurde auf Studierende der Psychologie an der Universität Hamburg zurückgegriffen. Psychologiestudenten gehören jedoch zu den am häufigsten untersuchten Teilpopulationen überhaupt und es wurden daher auch Nicht-Psychologiestudenten mit einbezogen, um die Untersuchungsmethode auch außerhalb eines psycho-universitären Umfeldes testen zu können und um eine vielseitigere Datenbasis zu erhalten. Es sollten in etwa gleichem Umfang sowohl Männer als auch Frauen untersucht werden. Ausgeschlossen von der Teilnahme waren Psychologiestudenten mit abgelegtem Vordiplom sowie Diplom-Psychologen.

Die Psychologiestudenten wurden per Aushang in den Fluren des Fachbereichs angeworben, die Nicht-Psychologiestudenten rekrutierte ich durch persönliche Ansprache in meinem Freundes- und Bekanntenkreis. Dabei ließen sich ohne große Mühe sogar solche Menschen als Vpn gewinnen, die ich nur oberflächlich kannte oder durch das Vortragen meines Anliegens (z.B. auf einer Party) erst kennenlernte. Die Psychologiestudenten erhielten als Gegenleistung für ihre Teilnahme alle für die Anmeldung zum Vordiplom erforderlichen 20 Vp-Stunden bescheinigt, die Nicht-Psychologiestudenten erhielten als kleine Anerkennung eine Flasche Sekt.

Erwünscht war eine weitestgehende Homogenität der Stichprobe, insbesondere eine Parallelität der Gruppen der Psychologie- und der Nicht-Psychologiestudenten hinsichtlich aller erhobener Ausgangsparameter, z.B. Geschlechterverteilung, Alter, Meßwerte aus den eingesetzten Instrumenten der Vorbefragung usw., mit dem Ziel, die ESM-Daten aller Vpn aus den Beobachtungszeiträumen zum Zwecke der Konstruktion eines Meßinstruments gemeinsam verarbeiten zu können. Abgesehen vom Geschlechterverhältnis wurde diese angestrebte Parallelität der beiden Gruppen während der Untersuchung jedoch nicht kontrolliert. Indizien für eine relative Repräsentativität der Gesamtstichprobe für die Bevölkerung der Bundesrepublik waren willkommen. Hinsichtlich einer Abschätzung der Einsatzmöglichkeiten und vor allem der Akzeptanz eines solchen Instruments sollten aber auch Unterschiede in der Beurteilung der Methode durch Psychologiestudenten einerseits und Nicht-Psychologiestudenten andererseits ermittelt werden können.

3.2.2. Eingesetzte Instrumente

3.2.2.1. Instrumente der Vorbefragung

Während der Testeinweisung wurde eine Vorbefragung durchgeführt, bei der vier Erhebungsinstrumente eingesetzt wurden:

1. ein ad hoc konstruierter Fragebogen (vgl. Anhang A),

2. die Deutsche Personality Research Form (PRF), Form KA (STUMPF, ANGLEITNER, WIECK, JACKSON & BELOCH-TILL, 1985)

3. der Traitfragebogen des STAI (STAI-G Form X2) (LAUX et al., 1981) sowie

4. der Streßverarbeitungsfragebogen (SVF) (JANKE, ERDMANN & BOUCSEIN, 1985).

Die in der Vorbefragung gewonnenen (Trait-) Daten sollen auf zweifache Weise Verwendung finden: zum einen sollen sie helfen, die mit Hilfe der Felduntersuchung ermittelten Traitwerte (dargestellt als intraindividuelle mittlere Itemwerte) in einen inhaltlichen Kontext zu anderen, klassisch gewonnenen Traitdaten zu setzen ("Validierung"), zum anderen dienen sie der Deskription der Stichprobe, d.h. der Aufdeckung möglicher bedeutsamer Unterschiede zwischen den Untergruppen der Vpn (Psychologiestudenten vs. Nicht-Psychologiestudenten, Männer vs. Frauen) sowie einer Abschätzung der Repräsentativität der Gesamtstichprobe für die bundesrepublikanische Bevölkerung.

Mit dem selbstkonstruierten Fragebogen wurden soziodemographische Merkmale erhoben, insbesondere Alter, Geschlecht, Schulabschluß, abgeschlossene Berufsausbildung, Art der Finanzierung des Lebensunterhalts und Fachsemester (nur bei den Psychologiestudenten).

Die PRF ist ein multivariater Persönlichkeitsfragebogen, der aus 234 Items zur Selbstbeschreibung besteht (z.B. Item 34: "Allein die Aussicht, stundenlang arbeiten zu müssen, macht mich müde."). Stimmt die Vp einem Item zu oder meint sie, daß es auf sie zutrifft, so soll sie die Antwortalternative "richtig" ankreuzen, im gegenteiligen Fall soll "falsch" markiert werden. Das Antwortformat ist also dichotom. Die Auswertung erfolgt mit Hilfe einer Schablone, auf der die symptomatischen Itemantworten gekennzeichnet sind. Bei Vorliegen einer symptomatischen Antwort zählt das Item einen Rohwertpunkt auf der zugehörigen Merkmalsdimension; im Falle einer nicht-symptomatischen Antwort ergibt das Item keinen Punkt. Insgesamt gibt es 14 inhaltliche Merkmalsskalen, denen jeweils 16 Items zugeordnet sind (also min. 0, max. 16 Rohwertpunkte je Skala möglich). Eine weitere Skala ("Infrequenz"- oder "Validitäts"-Skala) dient zur Aufdeckung fehlerhaften Antwortverhaltens, z.B. infolge nachlässiger, unkorrekter oder unkooperativer Fragebogenbearbeitung. Sie umfaßt zehn Items, die so formuliert sind, daß ihnen praktisch jeder zustimmen bzw. nicht zustimmen kann (z.B. Item 176: "Ich denke nie über meine Zukunft nach."). Die 16 Inhaltsskalen sind folgenden Merkmalsbereichen zugeordnet: Leistungsstreben, Geselligkeit, Aggressivität, Dominanzstreben, Ausdauer, Bedürfnis nach Beachtung, Risikomeidung, Impulsivität, Hilfsbereitschaft, Ordnungsstreben, Spielerische Grundhaltung, Soziales Anerkennungsbedürfnis, Anlehnungsbedürfnis und Allgemeine Interessiertheit. Aufgrund bestehender Normierung lassen sich die ermittelten Rohwerte geschlechts- und altersgruppenabhängig in Staninewerte umrechnen (STUMPF et al., 1985).

Der Traitteil des STAI umfaßt 20 Items, die der Feststellung interindividueller Unterschiede im Ausprägungsgrad der Ängstlichkeit dienen, also der "Tendenz, Situationen als bedrohlich zu bewerten und hierauf mit einem Anstieg der Zustandsangst zu reagieren" (LAUX et al., 1981, S. 50) (z.B. Item 35: "Ich fühle mich niedergeschlagen"). Die Vpn sind dabei angewiesen, die Items entsprechend ihrer habituellen Befindlichkeit zu beantworten, also anzugeben, wie sie sich im allgemeinen fühlen. Antwortformat ist eine vierstufige Ratingskala mit Skalenwerten von eins bis vier ("fast nie", "manchmal", "oft" und "fast immer"). Die je Item angekreuzten Skalenwerte werden (bei positiv formulierten Items nach erfolgter Umpolung, also: Skalenwert = 5 - angekreuztem Wert) aufsummiert. Die Auswertung erfolgt hier mit Hilfe einer selbstgefertigten Schablone, auf der die

umzupolenden Items besonders gekennzeichnet sind. Die Rohsummen (min. 20, max. 80 Punkte) können ebenfalls anhand von Normtabellen in geschlechts- und altersgruppenspezifische Staninewerte umgerechnet werden (LAUX et al., 1981).

Der SVF erfaßt mit seinen 114 Items die Tendenz, in Belastungssituationen mit bestimmten Strategien zur Streßbewältigung zu reagieren. Jeder einzelnen der insgesamt 19 Streßverarbeitungsmaßnahmen, die der SVF erfassen soll, sind sechs Items zugeordnet. Bei diesen Merkmalen handelt es sich im einzelnen um: Bagatellisierung, Herunterspielen durch Vergleich mit anderen, Schuldabwehr, Ablenkung von Situationen, Ersatzbefriedigung, Suche nach Selbstbestätigung, Situationskontrollversuche, Reaktionskontrollversuche, Positive Selbstinstruktion, Bedürfnis nach sozialer Unterstützung, Vermeidungstendenz, Fluchttendenz, Soziale Abkapselung, Gedankliche Weiterbeschäftigung, Resignation, Selbstbemitleidung, Selbstbeschuldigung, Aggression und Pharmakaeinnahme. Alle Items sind als Vervollständigung eines einzigen, für alle Items geltenden Satzanfangs konzipiert, der oben auf jeder Fragebogenseite wiederholt wird. (Satzanfang: "Wenn ich durch irgendetwas oder irgendjemanden beeinträchtigt, innerlich erregt oder aus dem Gleichgewicht gebracht worden bin..."; Beispielitem 22: "... denke ich hinterher immer wieder darüber nach") Die Vpn geben auf einer fünfstufigen Ratingskala an, wie wahrscheinlich die im jeweiligen Item beschriebene Reaktion in einer so beschriebenen Belastungssituation für sie ist. Die Skala hat Werte von null bis vier (entsprechend den Antwortkategorien "gar nicht", "kaum", "möglicherweise", "wahrscheinlich" und "sehr wahrscheinlich"); es können somit je Merkmalsskala Rohwertsummen zwischen min. 0 und max. 24 Punkten erreicht werden. Da der SVF noch nicht geeicht ist, liegen für ihn keine Normtabellen vor; für Berechnungen stehen daher nur die Rohwerte zur Verfügung (JANKE, ERDMANN & BOUCSEIN, 1985).

Die PRF wurde eingesetzt mit dem Ziel, ein möglichst breites Spektrum von Persönlichkeitsmerkmalen mit Hilfe eines solide konstruierten Meßinstruments zu erfassen, das aber zugleich den Psychologiestudenten unter den Vpn aus den Lehrveranstaltungen möglichst unbekannt sein sollte. Deshalb fand insbesondere das für eine Untersuchung im Befindlichkeitsbereich besser geeignete Freiburger Persönlichkeitsinventar (FPI) (FAHRENBERG, HAMPEL & SELG, 1984) keine Verwendung. Die im Zentrum der Untersuchung stehenden zehn Items der Protokollbögen sollten die hypothetischen Dimensionen euphorische Stimmung und dysphorische Stimmung (Beispielitem: "Ich bin ängstlich") erfassen (vgl. Kap. 3.2.2.2.). Aufgrund der relativen Nähe dieser zu messenden Dimensionen zu den Streß- und Ängstlichkeitskonzepten (vgl. dazu LAUX, 1983, SPIELBERGER, 1972) bot es sich an, in der Vorbefragung auch Traitdaten aus den Bereichen Ängstlichkeit und Streßverarbeitung zu erheben.

3.2.2.2. Protokollbögen

Mit Hilfe der Protokollbögen wurden die Kerndaten der vorliegenden Untersuchung gewonnen. Zentraler Gesichtspunkt bei der Gestaltung der Protokollbögen war der Aspekt der Ökonomie: Es sollte mit einem Minimum an Aufwand seitens der Vpn, d.h. mit einem nur minimalen Eingriff in den Verhaltensstrom der Vpn, ein Maximum an verwertbaren Informationen gewonnen werden.

Jede Vp erhielt ein Ringbuch im DIN-A5-Format, in dem sich 45 durchnumerierte Protokollbögen befanden. Zu jedem der 42 Protokolltermine sollte ein Protokollbogen bearbeitet werden. Jeder Protokollbogen besteht aus drei DIN-A5-Seiten; oben auf jeder Seite befindet sich die Protokollbogennummer, die Seitennummer innerhalb des Protokollbogens und zusätzlich ein Feld, in das die Vp-Nummer vor Beginn des Beobachtungszeitraums vom Untersuchungsleiter handschriftlich eingetragen wurde. Es läßt sich also jedes einzelne Blatt aller verwendeten Protokollbögen eindeutig einem bestimmten Protokollbogen und einer bestimmten Vp zuordnen. (Anhang B zeigt beispielhaft die Protokollbögen mit den Nummern 6, 19 und 34.) Alle Protokollbögen sind einander grundsätzlich gleich, lediglich die zehn relevanten Stimmungsitems, um die es hier geht, wurden doppelt und in stets anders rotierter, randomisierter Reihenfolge vorgegeben. Für die 45 Protokollbögen ergeben sich somit 90 Zufallsreihenfolgen der zehn relevanten Items. Die Randomisierung erfolgte archaisch, aber wirksam durch Blindziehung von numerierten Glasmarmeln aus einer Urne. Durch sie ist die Äquivalenz aller Protokollbögen sichergestellt, so daß das Beachten einer bestimmten Reihenfolge bei ihrer Verwendung unnötig ist. Der Grund für die doppelte Vorgabe liegt in der beabsichtigten Anwendung des State-Trait-Modells von BUSE & PAWLIK (1991), bei der die doppelte Messung derselben Items zu einem Meßzeitpunkt zur Quantifizierung des Meßfehlers dient (vgl. Kap. 2.1.3.) - ein Ansatz, den schon NOWLIS & GREEN (1964, zit. bei NOWLIS, 1965, S. 368) zur Bestimmung der Retestreliabilitäten einzelner Items innerhalb einer Testvorgabe oder auch YAGI & BERKUN (1961, zit. bei NOWLIS, 1965, S. 368) zur Bestimmung der Glaubwürdigkeit ihrer Vpn verfolgten, als sie in eine lange Stimmungsliste ein paar Items doppelt aufnahmen.

Auf der ersten Seite eines jeden Protokollbogens waren zunächst Angaben zu Datum und Uhrzeit zu machen. (Diese Angaben bezogen sich auf den Moment des Bearbeitens des Protokollbogens, nicht auf den Moment, zu dem das Wecksignal ertönte, denn die Bearbeitung durfte noch bis zu dreißig Minuten nach dem Signal erfolgen; vgl. Kap. 3.2.3.) Als nächstes folgt ein Item, das in grober Rasterung die Tätigkeit der Vp unmittelbar vor Ausfüllen des Protokollbogens erfragt. Drei weitere Items beziehen sich auf das Setting im Moment des Protokollierens ("zu Hause" / "nicht zu Hause"), auf das Erleben des Settings ("gewohnt" / "ungewohnt") und auf den sozialen Charakter dieses Settings ("allein" / "nicht allein"). Es folgt ein weiteres Item (optional für die Bedingung "nicht allein"), das nach dem Vorhandensein bzw. der Art des Kontakts zu den anderen anwesenden Personen fragt. Diese Items dienen zum einen der Abschätzung der ökopsychologischen Gütekriterien (vgl. Kap. 2.2.3.), zum anderen sollten sie helfen, den Untersuchungszweck zu verschleiern.

Im Anschluß an die Erhebung dieser situations- bzw. verhaltensbezogenen Merkmale erfolgt die Vorgabe von 38 Items zur Messung der momentanen Befindlichkeit. Die Items von Nr. 1 bis Nr. 10 sind dieselben wie die von Nr. 29 bis Nr. 38; sie werden auf jedem einzelnen Protokollbogen in einer je Itemblock immer wieder neu variierten Reihenfolge dargeboten. Bei diesen Items handelt es sich um fünf Aussagen, die einer hypothetischen gehobenen Stimmungslage zuzuordnen sind:

1. Ich fühle mich ungezwungen

2. Ich fühle mich gewachsen

3. Ich bin vergnügt

4. Ich bin froh

5. Ich bin erfreut

sowie um fünf weitere Aussagen, die einer hypothetischen gedrückten Stimmungslage entsprechen:

6. Ich bin deprimiert

7. Ich bin sorgenvoll

8. Ich bin ängstlich

9. Ich bin betrübt

10. Ich fühle mich überfordert.

Der erste Zehnerblock stellt die erste, der zweite Zehnerblock die zweite Messung im Sinne des State-Trait-Modells von BUSE & PAWLIK (1991) dar.

Da diese beiden Messungen möglichst unabhängig voneinander sein sollten, war es wünschenswert, daß die Vpn innerhalb eines Protokolltermins dieselben Items zweimal beantworten, ohne dies jedoch zu bemerken oder ohne sich dadurch beeinflussen zu lassen in dem Falle, daß sie es doch bemerken sollten. Die Vpn waren deshalb zwar generell so instruiert, daß sie sich im Falle des Auftretens von sehr ähnlichen oder sich wiederholenden Items innerhalb eines Protokollbogens immer wieder erneut fragen sollten, wie es ihnen gerade jetzt, in diesem Moment des Ankreuzens ergeht, unabhängig davon, was sie vorher angegeben hatten (vgl. Kap. 3.2.3.), doch wäre es eine Überforderung auch der gutwilligsten Vpn gewesen, wenn die zwei mal zehn Items einfach so aufeinander gefolgt wären, womöglich noch alle auf derselben Seite. Auf diese Weise wären die doppelten Items aufgefallen, und ein Abschreiben oder zumindest ein Abgleich mit den zuerst bearbeiteten identischen Items wäre beim Lösen der zweiten Items unvermeidbar gewesen und hätte zu unbrauchbaren Daten geführt. Es war daher dringend geboten, Vorkehrungen zu treffen, die ein Abschreiben, ein Nachschlagen oder sogar ein Erinnern an die bereits dargebotenen Items verhindern oder zumindest unwahrscheinlicher machen sollten. Neben der erneuten Durchmischung der Items auf jedem Protokollbogen, die ein Wiedererkennen der Items von Zehnerblock zu Zehnerblock desselben Bogens, aber auch von Protokollbogen zu Protokollbogen durch Vermeidung von Positionseffekten erschweren sollte, und neben einer entsprechenden Instruktion der Vpn erschien es sinnvoll, einen Itempuffer zwischen die beiden Blöcke mit den relevanten Items zu schieben, so daß die Items aus den beiden Zehnerblöcken

nicht auf derselben Protokollbogenseite zu finden wären. Zusätzlich würden diese Pufferitems die Vpn ablenken und mögliche Lerneffekte innerhalb eines Protokollbogens im Sinne einer retroaktiven Hemmung abschwächen. Dazu war es wünschenswert, daß die Items dieses Pufferblocks sich inhaltlich, von ihrem Erscheinungsbild her, von der Formulierung und vom Antwortformat her nicht von den kritischen Zehnerblöcken abhoben. Es lag nahe, hierfür ebenfalls Befindlichkeitsitems zu verwenden, damit die relevanten Items zusammen mit den Pufferitems eine Itemliste bilden konnten, die wie aus einem Guß erschien. Die Items von Nr. 11 bis Nr. 28 wurden zu diesem Zweck in die Protokollbögen aufgenommen. Es sind die ersten 18 (von 20) Items des STAI-Stateteils; sie werden stets an derselben Stelle in den Protokollbögen dargeboten und in ihrer Reihenfolge nicht rotiert. Dieser itemreduzierte STAI-Statepuffer wird im folgenden als STAI-S18 bezeichnet. (Der Grund für die Reduzierung von 20 auf 18 Items ist simpel: Die letzten beiden Items des originalen STAI-Stateteils lauten "Ich bin froh" und "Ich bin vergnügt"; beide sind im Pool der relevanten Items bereits vertreten.) Konkret bewirkt die Einschaltung der STAI-S18 Items, daß auf der ersten Seite eines jeden Protokollbogens die Items des ersten Zehnerblocks von Nr. 1 bis Nr. 7 stehen, auf der zweiten Seite die von Nr. 8 bis Nr. 10 und direkt daran anschließend die STAI-S18 Items von Nr. 11 bis Nr. 28. Auf der dritten Seite befinden sich dann die Items des zweiten Zehnerblocks von Nr. 29 bis Nr. 38. Die Items des ersten Zehnerblocks und die des zweiten Zehnerblocks befinden sich also unter keinen Umständen gemeinsam auf derselben Seite eines Protokollbogens. Wenn die Protokollbögen so verwendet werden wie geplant, d.h. wenn sie während des Beobachtungszeitraums beim Protokollieren nicht aus dem Ringbuch herausgenommen werden und insbesondere nicht nebeneinander gelegt werden, kann die Vp die Struktur der Itemliste nicht durch unmittelbare Anschauung erkennen und auch keinen direkten Abgleich eines im zweiten Block dargebotenen Items mit demselben im ersten Block dargebotenen Item vornehmen.

Zur Erlangung psychometrisch einwandfreier Daten aus zwei unabhängigen Messungen innerhalb eines Meßzeitpunkts wurden demnach vier Maßnahmen im Sinne einer Verringerung der Durchschaubarkeit des Instruments ergriffen:

1. die Vpn wurden entsprechend instruiert,

2. zwischen den beiden Zehnerblöcken mit den relevanten Items wurden achtzehn Pufferitems eingefügt zur Schaffung einer raum-zeitlichen Distanz bei gleichzeitiger Kompensierung des Lerneffekts aus der ersten Vorgabe durch retroaktive Hemmung, wodurch

3. die Items der beiden Zehnerblöcke immer auf verschiedenen Seiten des Protokollbogens plaziert werden konnten und

4. zur Erschwerung eines unerwünschten Lerneffekts über den gesamten Beobachtungszeitraum hinweg wurden die Items der Zehnerblöcke stets in neu randomisierten Reihenfolgen angeboten.

Die Administration der STAI-S18 Items ermöglicht zusätzlich die Gewinnung eines STAI-Statewertes für jede Vp zu jedem Meßzeitpunkt (zur psychometrischen Qualität itemreduzierter STAI-Statewerte vgl. Kap. 2.3.2.). Sie eröffnet damit die Möglichkeit eines Abgleichs der mit ihrer

Hilfe gewonnenen Meßwerte mit den Meßwerten der übrigen verwendeten Befindlichkeitsitems ("Übereinstimmungsvalidierung" der Statewerte).

Das Antwortformat für alle 38 Befindlichkeitsitems ist eine vierstufige Ratingskala mit den Kategorien "überhaupt nicht", "ein wenig", "ziemlich" und "sehr", denen die Werte eins bis vier zugeordnet sind. Als Konsequenz aus der oben geführten Diskussion (vgl. Kap. 2.3.3.) wäre die Verwendung eines symmetrischen Antwortformats (z.B. das von LORR & SHEA, 1979) gerade im Hinblick auf die faktorenanalytische Konstruktion eines Instruments zur Stimmungsmessung sicherlich geschickter gewesen. Zudem legen auch die vergleichenden Untersuchungen von RUSSELL (1979) zu den Häufigkeitsverteilungen der Antworten der Vpn bei unterschiedlichen Antwortformaten die Verwendung symmetrischer Antwortskalen nahe. RUSSELL (1979) beobachtete nämlich bei der Verwendung der symmetrischen Ratingskala von MEDDIS (1972) eine unimodale, annähernd symmetrische Verteilung der Antworten auf die vier Antwortkategorien, während die Verwendung der asymmetrischen, in eine Richtung ansteigenden Skala von McNAIR & LORR (1974, zit. bei RUSSELL, 1979, S. 348), die der hier verwendeten ähnlich ist, zu einer ebenfalls unimodalen, aber stark linkssteilen Verteilung führte. Symmetrische Skalen rechtfertigen daher eher noch als asymmetrische die Anwendung parametrischer Prüfverfahren. Die Wahl fiel dennoch auf die hier vorliegende Skala, weil es die Originalskala des STAI-Stateteils ist (LAUX et al., 1981). Es wurden dabei nicht nur die Antwortkategorien übernommen, sondern die gesamte Stimmungsliste wurde in ihrem Layout dem Stateteil des STAI nachempfunden, im Vertrauen auf das Renommee eines angesehenen Meßinstruments und im Hinblick auf die Verwertbarkeit der STAI-S18 Meßwerte. Auch die Instruktionen zum Bearbeiten der Protokollbögen wurden wörtlich vom STAI-Stateteil übernommen (vgl. Kap. 3.2.3.).

Von den zehn relevanten Stimmungsitems, die hier zur Konstruktion einer änderungssensitiven State-Trait-Stimmungsliste als Itempool herangezogen werden, sollten je fünf die Merkmale "allgemeine gehobene Stimmung" und "allgemeine gedrückte Stimmung" messen, zwei a priori Konstrukte, die sich so oder so ähnlich schon in einer Reihe anderer Untersuchungen gezeigt haben (vgl. Übersicht bei PLUTCHIK & CONTE, 1989; außerdem BOTTENBERG, 1970, BÜNNING, 1984, HAMPEL, 1977, HECHELTJEN & MERTESDORF, 1973, LORR & SHEA, 1979). Ausgehend von den Untersuchungen von BÜNNING (1984) fiel die Wahl auf die o.a. zehn Adjektive, die für die vorliegende Untersuchung (in Annäherung an den STAI-Stateteil) in die Form einfacher Aussagesätze gebracht wurden. BÜNNING (1984) ließ seine 48 Vpn für die Dauer von zwei Wochen täglich zweimal eine 33 Items umfassende Stimmungsliste unter Vorgabe einer graphischen Ratingskala bearbeiten. Dabei waren die Vpn instruiert, ihre momentane Stimmungslage zu protokollieren. Er führte mit den zu Mittelwertprofilen aggregierten Einzelmessungen eine Trait-Faktorenanalyse (R-Technik) durch und mit den um ihre intraindividuellen Mittelwerte verminderten Einzelmeßwerten aller Vpn zu allen Meßzeitpunkten eine State-Faktorenanalyse (Ketten-P-Technik) (vgl. Kap. 2.3.2.).

Für das vorliegende Erhebungsinstrument sollten je hypothetischem Merkmal (euphorische und dysphorische Stimmung) aus den 33 Items von BÜNNING (1984) diejenigen fünf Items ausgewählt werden, die in beiden Faktorenanalysen jeweils gemeinsam auf einem sinnverwandten Faktor luden

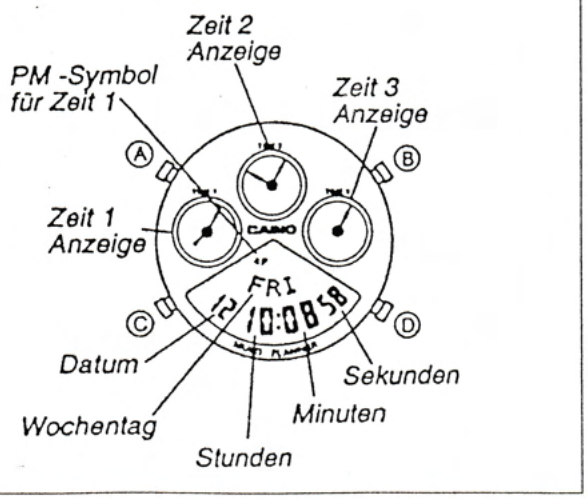
und dabei möglichst hohe Ladungen von mindestens .40 aufweisen konnten. Gleichzeitig sollten sie möglichst auf keinem anderen von BÜNNINGs (1984) Faktoren mit nennenswerten Ladungen vertreten sein. Dies ließ sich bei den letztlich in den Pool der zehn relevanten Items aufgenommenen Adjektive im Rahmen der Stateanalyse verwirklichen, bei der Traitanalyse gab es jedoch z.T. Überschneidungen der Items der gedrückten Stimmungslage, die allesamt auf dem Faktor "Depressivität" hoch luden, mit den Faktoren "Aggressivität" und "Erschöpfung", während ein Item der gehobenen Stimmungslage eine substantiell negative Ladung auf dem Faktor "Depressivität" besaß. Als zusätzliches Auswahlkriterium fungierte die Verteilung der Skalenwerte je Item über alle Meßzeitpunkte und alle Vpn: Die Skalenwerte sollten im Sinne einer mittleren Itemschwierigkeit möglichst breit streuen, denn es erschien unsinnig, ein Item auswählen zu wollen, dem kaum eine von BÜNNINGs (1984) Vpn je hat zustimmen können. Dies ließ sich am ehesten für die Items der gehobenen Stimmungslage realisieren. Die Items der dysphorischen Stimmung erhielten dagegen weitaus häufiger Skalenwerte, die sich am linken Rand der graphischen Skala ("überhaupt nicht zutreffend") befanden; es waren somit Items, die allesamt einen niedrigen Schwierigkeitsindex hatten, also relativ selten in substantiell symptomatischer Richtung angekreuzt worden waren. Eine Augenscheinoptimierung hinsichtlich aller Kriterien führte letztlich zur Auswahl der genannten zehn für die vorliegende Untersuchung relevanten Items.

Zehn Items scheinen als Itempool für eine Testkonstruktion eine geringe Anzahl zu sein, doch war Vorsicht geboten, wo die Gefahr bestand, die Vpn durch ein zu umfangreiches Erhebungsinstrument von der Teilnahme an der Untersuchung abzuschrecken oder sie im Alltag übermäßig zu belasten, was zu einer unkontrollierten Störung des Verhaltensstroms oder zu einer Beeinträchtigung der Compliance hätte führen können mit der Konsequenz einer reduzierten Datenbasis bzw. einer verschlechterten Datenqualität.

3.2.2.3. Signalgeber und Weckterminpläne

Jede Vp erhielt als Signalgeber eine Armbanduhr Marke CASIO (Multi Planner, Module No. 931) (Abb. 1), die die Möglichkeit zum Einspeichern von max. 60 Weckterminen bot. Diese Uhren waren Eigentum der Universität Hamburg und standen in hinreichender Stückzahl (8 Exemplare) während der gesamten Untersuchung zur Verfügung.

Abb. 1 Signalgeber



Jeder Wecktermin mußte mit Monats-, Tages-, Stunden- und Minutenangabe vor Untersuchungsbeginn vom Untersuchungsleiter mit Hilfe der vier Funktionsknöpfe eingegeben werden. Die Alarmfunktion mußte zusätzlich bei Übergabe des Signalgebers aktiviert werden. Darauf erzeugte die Uhr zu jedem eingespeicherten Termin für die Dauer von max. 20 Sekunden einen Piepton. Die Vpn konnten diesen Signalton durch Betätigen irgendeines Funktionsknopfes abstellen. Ansonsten waren zum Handling der Uhr unter Alltagsbedingungen keinerlei Aktivitäten der Vpn erforderlich. Nach Verstreichen eines Wecktermins war die Weckfunktion für den nächsten Termin automatisch aktiviert. Es gab für die Vpn keine Möglichkeit zum Stummschalten der Alarmfunktion; jeder Protokolltermin wurde angezeigt. Insgesamt waren in jeder Uhr 41 Wecktermine eingespeichert (vgl. Weckterminpläne in Anhang B), jeweils sechs für die ersten sechs Untersuchungstage und fünf für den letzten Tag. (Der für die Gesamtzahl von 42 Protokollterminen fehlende Termin lag während der Testeinweisung. Dort wurden zwei Protokolle übungshalber bearbeitet; das zweite davon ging als erstes gültiges Protokoll in die Auswertung mit ein.)

Während der Untersuchung zeigte die Uhr die mitteleuropäische Zeit in der unteren großen Digitalanzeige korrekt an, die drei Analogdisplays darüber zeigten verschiedene Ortszeiten (vgl. Abb. 1). Datum und Wochentag waren falsch eingestellt. Dies hatte seine Ursache darin, daß in die einzelnen Uhren verschiedene Weckterminpläne eingespeichert waren. Jedem dieser Weckterminpläne war ein anderer Monat zugeordnet; die Unterscheidung der einzelnen Uhren wurde so für den Untersuchungsleiter erleichtert. Der Tag der Übergabe der Uhr, der gleichzeitig Tag der Testeinweisung war, erhielt das Datum 1. des jeweiligen Monats, die Untersuchungstage mit den Protokollterminen hatten demzufolge die Daten 2. bis 8. des Monats. Daraus ergaben sich die falschen Datum- und Wochentagangaben auf dem Uhrendisplay. Wurde die Uhr an die nächste Vp weitergegeben, wurde das Datum auf den 1. zurückgestellt und der Weckterminplan erneut durchlaufen.

Eine potentielle Schwierigkeit lag in den Funktionsknöpfen der Uhr, die aus dem Gehäuse herausragten und durch einfaches Drücken betätigt werden konnten. Sie stellten eine Störquelle für den Fall ihrer unbeabsichtigten Betätigung im Feld dar (z.B. durch Anstoßen mit der Uhr an feste Gegenstände wie Tischplatte o.ä.). Insbesondere konnten durch einfaches Betätigen des Knopfes D die nächsten drei Wecktermine in den Analogdisplays sichtbar gemacht werden, was den ESM-Charakter der Untersuchung hätte gefährden können. Eine entsprechende Anleitung zur Behebung einfacher Störungen (bei unbeabsichtigtem Betätigen von nur einem der Knöpfe) wurde den Vpn ausgehändigt (vgl. Anhang A und B).

Da es unsinnig ist, Vpn während ihrer Schlafphase aufzuwecken und ihre Stimmung einschätzen zu lassen, sollten alle Protokollermine grundsätzlich innerhalb der Wachphasen der Vpn plaziert werden. Diese sind jedoch inter- wie intraindividuell variabel und deshalb wurde pro Tag des Beobachtungszeitraums ein Zwölfstundenzeitraum zur Plazierung der Protokollermine so gewählt, daß er sich mit hoher Wahrscheinlichkeit mit den Wachphasen der Vpn während der sieben Untersuchungstage zu decken versprach. Konkret gab es sechs verschiedene Weckterminpläne (vgl. Anhang B), einen, bei dem dieser Zwölfstundenzeitraum von 12.00h mittags bis Mitternacht reichte, drei, bei denen die Vpn zwischen 9.00h morgens und 21.00h abends Signale erhielten, sowie zwei

weitere, bei denen dies zwischen 10.00h morgens und 22.00h abends erfolgte. Da bis zu acht Uhren im Einsatz waren, wurden zwei Weckterminpläne doppelt implementiert. Dabei wurde darauf geachtet, daß zwei Vpn, die z.B. zusammen wohnten, nicht zwei Uhren mit identischen Terminplänen erhielten. Jeder dieser sechs verschiedenen Weckterminpläne kam zwischen 10 und 15 mal zum Einsatz (vgl. Anhang B).

Die täglichen Zwölfstundenintervalle wurden in sechs Intervalle je zwei Stunden gegliedert. Pro Zweistundenintervall sollte ein Wecktermin liegen, der nach Zufall zu plazieren war. Dies erfolgte mittels zwölf numerierter Glasmarmeln (1 bis 12), die blind aus einer Urne gezogen wurden. Die erste gezogene Marmel gab einen Zehnminutenzeitraum (von zwölf) innerhalb der zwei Stunden an; sie wurde zurückgelegt und anschließend wurde die Minute innerhalb dieses Zehnminutenzeitraums gezogen (dabei waren die Marmeln 11 und 12 ungültig). Beispiel: Es galt, für eine Uhr, die zwischen 9.01h und 21.00h wecken sollte, den zweiten Wecktermin zu bestimmen. Der sollte zwischen 11.01h und 13.00h liegen. Die erste gezogene Marmel war die Zehn. Das bedeutete, der Wecktermin sollte im 10. Zehnminutenzeitraum nach 11.00h liegen, also zwischen 12.31h und 12.40h. Die Marmel wurde zurückgelegt. Die zweite gezogenen Marmel war die Elf. Sie wurde, da in diesem Durchgang ungültig, zurückgelegt. Erneutes Ziehen brachte eine Sechs. Der endgültige Wecktermin lautete demnach 12.36h. Lagen zwei auf diese Weise ermittelte Protokolltermine zu dicht beieinander, so wurde auf einen minimalen Abstand zwischen beiden Terminen von ca. 45 Minuten geachtet. Am siebten Tag entfiel der letzte Wecktermin, was den Vorteil hatte, daß zu dem Zeitpunkt bereits die Nachbefragung durchgeführt werden konnte, ohne daß dabei ein Protokolltermin verlorengegangen wäre.

3.2.2.4. Instrument der Nachbefragung

Bevor die Vpn anläßlich der Nachbefragung in den Untersuchungszweck eingeweiht wurden und bevor ein mündlicher Austausch über den Beobachtungszeitraum stattfand, wurden sie gebeten, einen ad hoc konstruierten Nachbefragungsbogen in Gegenwart des Untersuchungsleiters selbständig zu bearbeiten (vgl. Anhang C). Daten aus solcher Art der schriftlichen Befragung, die für alle Vpn unter annähernd gleichen Bedingungen stattfindet (wie etwa Dauer des Bearbeitens des Fragebogens, kein Kontakt der Vpn zu anderen Personen usw.), bei der der Untersuchungsleiter zur Beseitigung von Unklarheiten anwesend ist und bei der er die äußeren Befragungsbedingungen auch hinreichend kontrollieren kann und an der praktisch alle Vpn auch wirklich teilnehmen und sich äußern, Daten solcher Art können eine annehmbare Zuverlässigkeit besitzen, besonders wenn es um Faktenfragen geht (ATTESLANDER & KOPP, 1987). Der Nachbefragungsbogen dient der Informationsgewinnung über das Verhalten der Vpn in der Felduntersuchung und darüber, wie sie die Untersuchung erlebt haben. Insbesondere stehen dabei die Aspekte der Sicherung der psychometrischen Qualität der erhobenen Daten aus den Protokollbögen, der Abschätzung der ökopsychologischen Gütekriterien (vgl. Kap. 2.2.3.), der Durchschaubarkeit der Untersuchung und der Methodenakzeptanz im Vordergrund.

Der Nachbefragungsbogen besteht aus drei Teilen: Im ersten und im letzten Teil werden Fragen gestellt, die mit "ja" oder "nein" beantwortet werden sollen. Für jeweils eine dieser beiden

Antwortmöglichkeiten ist eine nähere Erläuterung in offener Antwortform vorgesehen (Ausnahme: Frage Nr. 14). Der erste Teil mit den Fragen von Nr. 1 bis Nr. 5 ist eher allgemein gehalten (mit Fragen nach der Repräsentativität des Beobachtungszeitraums, nach angenehm und unangenehm erlebten Aspekten der Untersuchung u.ä.), während die Fragen des letzten Teils von Nr. 7 bis Nr. 15 mehr nach speziellen Details der Untersuchung fragen (z.B. nach falsch protokollierten Angaben, nach einem Protokollieren unter Drogeneinfluß, nach dem vermuteten Untersuchungszweck u.ä.). Der mittlere Teil, die Frage Nr. 6, beinhaltet 21 Items, in denen vorformulierte Aussagesätze über die Untersuchung dargeboten werden, denen die Vp entsprechend ihrem Erleben und Verhalten mehr oder weniger zustimmen kann (Beispielitem: "Das Tragen der Uhr war unangenehm."). Die Zustimmung erfolgt auf einer fünfstufigen Ratingskala mit den Antwortkategorien "stimme ganz und gar zu", "stimme ein bißchen zu", "bin nicht sicher", "stimme weniger zu" und "stimme gar nicht zu", denen die Werte eins bis fünf zugeordnet sind.

Mit den Fragen des ersten Teils sollten die Spontanäußerungen der unbefangenen Vpn abgeschöpft werden, ohne ihre Aufmerksamkeit schon auf bestimmte Detailfragen zu lenken. Informationen zu bestimmten untersuchungsrelevanten Themen sollten dann in quantifizierbarer Form durch die Itembatterie in Frage Nr. 6 gewonnen werden. Bei den offenen Fragen des letzten Abschnitts ging es um bestimmte Einzelaspekte, deren Klärung für die Qualität der Protokolldaten z.T. erhebliche Relevanz hatte und deren Beantwortung von den Vpn ein hohes Maß an Aufrichtigkeit verlangte.

3.2.3. Durchführung

Die Datenerhebung fand in der Zeit von Juni 1992 bis Januar 1993 statt. (Zwischen dem 24.12.1992 und dem 3.1.1993 wurden keinerlei Daten erhoben.) Die Vpn wurden einzeln oder in kleinen Gruppen in die Untersuchung eingewiesen; da max. acht Signalgeber zur Verfügung standen, konnten auch immer nur acht Vpn gleichzeitig protokollieren. Die Testeinweisung und Vorbefragung fand ebenso wie die Nachbefragung z.T. in einem Untersuchungsraum im Institut I statt, z.T. aber auch in den Wohnungen der Vpn, in der Wohnung des Untersuchungsleiters oder auch in der Firma des Untersuchungsleiters. Die Nachbefragung erfolgte ebenfalls z.T. in Gruppen, z.T. als Einzelbefragung.

Einen Tag vor Beginn des siebentägigen Beobachtungszeitraums erfolgte eine Testeinweisung und Vorbefragung. Sie fand in jedem Fall in der ersten Wochenhälfte, d.h. an einem Montag, Dienstag oder Mittwoch statt; Beginn des Protokollierens war dann der nächstfolgende Tag, also Dienstag, Mittwoch oder Donnerstag. Nur in zwei Ausnahmefällen begann der Beobachtungszeitraum direkt am Tag der Vorbefragung. In der Testeinweisung und Vorbefragung wurden die Vpn über den Untersuchungsablauf informiert und über ihr erwünschtes Verhalten während des Beobachtungszeitraums instruiert; sie bearbeiteten eine Mappe mit unterschiedlichen Fragebögen (vgl. Kap. 3.2.2.1.), sie hatten Gelegenheit, das Ausfüllen der Protokollbögen zu üben und sich mit der Bedienung des Signalgebers vertraut zu machen. Außerdem konnten Fragen gestellt und Unklarheiten beseitigt werden.

Das Material, das den Vpn ausgehändigt wurde, umfaßte ein DIN-A5-Ringbuch mit 45 dreiseitigen Protokollbögen, den Signalgeber, mehrere mit der Anschrift des Untersuchungsleiters versehene Briefumschläge (Versandtaschen DIN-A5) sowie eine zweiseitige schriftliche Testinstruktion (vgl. Anhang A). Am Ringbuch war ein Kugelschreiber mit einem Bindfaden befestigt und in der Innentasche des Ringbuchs steckte eine zweiseitige Kurzfassung der Instruktionen, die um die wörtlich aus dem STAI-Stateteil übernommene Anleitung zum Ausfüllen der Protokollbögen ergänzt war (vgl. Anhang B).

Die Testeinweisung erfolgte nahezu standardisiert: Der Untersuchungsleiter verlas die Testinstruktion, die zugleich jeder Vp zum Mitlesen vorlag (Anhang A). Die Vpn erhielten darüber hinaus keine weiteren Informationen über Sinn und Fragestellung der gesamten Untersuchung. In dieser Testinstruktion wurden die Vpn dahingehend "gebrieft", daß sie nach Ertönen des Signaltons möglichst sofort, bei Verhinderung zum nächstmöglichen Zeitpunkt innerhalb einer Latenzzeit von max. 30 Minuten danach, den im Ringbuch zuoberst liegenden Protokollbogen bearbeiten sollten. In jedem Falle sollten alle Angaben für den Moment des Protokollierens gemacht werden, nicht für den Moment des Ertönens des Signaltons. Dies bedeutet bei hoher Latenzzeit prinzipiell eine Beeinträchtigung des ESM-Charakters der Untersuchung (vgl. Kap. 2.2.3.), doch wären andernfalls unkontrollierbare Gedächtniseffekte ins Spiel gekommen. Beim Überhören eines Protokolltermins oder beim Überspringen eines oder mehrerer Termine, z.B. wenn mehr als 30 Minuten nach dem Signalton vergangen sein sollten, sollte zum nächsten wahrgenommenen Meßzeitpunkt der jeweils nächste unverbrauchte Protokollbogen im Ringbuch bearbeitet werden. Die Verwendung der Protokollbögen in der vorgegebenen Reihenfolge unter Absehung von evt. ausgelassenen Protokollterminen oder gar ihre Verwendung in anderer als der numerierten Reihenfolge (anläßlich des Herausnehmens der Protokollbögen aus dem Ringbuch z.B. zum bequemeren Transport) ist unproblematisch, da alle Protokollbögen einander äquivalent und dadurch beliebig austauschbar sind (vgl. Kap. 3.2.2.2.). Die Vpn wurden weiterhin darauf hingewiesen, daß sie möglichst alle Protokolltermine wahrnehmen sollten. Von eminenter Wichtigkeit ist der vierte Absatz der schriftlichen Testinstruktion. In ihm werden die Vpn dazu angehalten, sich beim Auftreten von sich ähnelnden oder sich wiederholenden Items innerhalb eines Protokolltermins bei jedem einzelnen Item immer wieder erneut zu fragen, wie ihre Befindlichkeit gerade jetzt, in diesem Moment des Ankreuzens ist, und sich nicht damit zu beschäftigen, was sie eben vorher geantwortet haben. Ergänzt wurde dies noch durch den mündlichen Hinweis darauf, daß es beim Protokollieren für die Vpn nicht darum zu gehen habe, durch das Abgeben von inhaltlich konsistenten Antworten vermeintlich gute Vpn sein zu wollen.

Um eine inhaltliche Bezugnahme auf bereits ausgefüllte Protokollbögen während des Beobachtungszeitraums zu erschweren, aber auch um eine Möglichkeit des geordneten und sicheren Aufbewahrens der Bögen zu schaffen, wurden den Vpn mehrere DIN-A5-Versandtaschen ausgehändigt, in die sie die ausgefüllten Protokollbögen so oft wie möglich hineinstecken sollten. Die Psychologiestudenten unter den Vpn mußten diese gesammelten Protokollbögen als zusätzliche Maßnahme zur Vermeidung einer unerwünschten Orientierung am bereits erhobenen Datenmaterial mindestens zweimal während der sieben Feldtage im Institut I abgeben oder per Post an den Untersuchungsleiter schicken. Diese Maßnahme entfiel für die Nicht-Psychologiestudenten. Ihnen

wurde nämlich unterstellt, hauptsächlich deshalb an der Untersuchung teilzunehmen, weil sie dem Untersuchungsleiter einen freundschaftlichen Gefallen tun wollten und weil sie vielleicht auch ein bißchen neugierig und an der Methode interessiert waren. Ansonsten hatten sie jedoch keinerlei Vorteile durch ihre Teilnahme zu erwarten, die ein unkorrektes Antwortverhalten im Feld sinnvoll gemacht hätten. Im Gegensatz dazu wurde bei den psychologiestudentischen Vpn vermutet, daß sie hauptsächlich wegen der in Aussicht gestellten Vp-Stunden zur Teilnahme bereit waren. Dies machte eine solche vorgezogene Rückgabepflicht lediglich für die Psychologiestudenten erforderlich. Allen Vpn wurde im letzten Absatz der Testinstruktion eindringlich klar gemacht, daß falsche Protokollangaben nicht nur wertlos, sondern für die Untersuchung in hohem Maße destruktiv sein würden. Sie wurden explizit aufgefordert, im Falle eines Absinkens der Compliance ihre Mitarbeit sofort abzubrechen. Den Psychologiestudenten wurde zudem mündlich erläutert, daß sie im Falle nur weniger wahrgenommener Protokolltermine oder im Falle eines Abbruchs ihrer Mitarbeit trotzdem mit der Bescheinigung aller oder zumindest sehr vieler Vp-Stunden rechnen könnten. Dies erschien notwendig, um auch von dieser Seite her die Wahrscheinlichkeit des Datenfälschens zu minimieren. Generell wurden die Vpn aufgefordert, sich während des Beobachtungszeitraums möglichst wenig mit der Untersuchung zu beschäftigen, d.h. sich möglichst wenig mit anderen darüber zu unterhalten, möglichst wenig über sie nachzudenken usw., damit ihr Alltagsleben so wenig wie möglich von der Datenerhebung gestört werden würde.

Nachdem die Vpn über den Ablauf und ihr erwünschtes Verhalten informiert worden waren, wurde ihnen der zweite Teil der Testinstruktion, die Hinweise zur Bedienung der Uhr (Anhang A), verlesen (vgl. Kap. 3.2.2.3.). Sie enthielt eine Übersicht über mögliche einfache Störungen beim Betätigen der Funktionsknöpfe, die z.T. die Alarmfunktion außer Kraft setzen und damit den Weckterminplan deaktivieren konnten. Die Vpn wurden zusätzlich aufgefordert, ein unbeabsichtigtes Betätigen der Funktionsknöpfe zu vermeiden und das Display der Uhr immer wieder auf mögliche aufgetretene Störungen zu überprüfen. Auf Wunsch der Vpn konnten sie durch Drücken der Funktionsknöpfe Störungen simulieren und das Wiederherstellen der Alarmfunktion und des ordnungsgemäßen Displays üben. Regelhaft wurde dies jedoch nicht durchgeführt, um die Vpn nicht durch zuviel Vertrautheit im Umgang mit den Funktionsknöpfen zum spielerischen Manipulieren der Uhr zu animieren.

Zum Schluß wurde den Vpn die private und z.T. auch die geschäftliche Telephonnummer des Untersuchungsleiters mitgeteilt, damit sie bei auftauchenden Unklarheiten oder Schwierigkeiten Betreuung und Hilfestellung erhalten konnten. Durch einen angeschlossenen fernabfragbaren Anrufbeantworter wurde den Vpn auf diesem Wege praktisch eine 24-Stunden-Service-Hotline zur Verfügung gestellt. Die Vpn wurden gebeten, bei Schwierigkeiten sofort von ihr Gebrauch zu machen.

Nach dieser Einführung wurde in der Testeinweisung und Vorbefragung als nächstes das Bearbeiten der Protokollbögen geübt. Dazu öffneten die Vpn das Ringbuch und es wurde ihnen als erstes die in der inneren Einstecktasche befindliche Anleitung vorgelesen (Anhang B). Diese bestand im wesentlichen aus der wörtlich übernommenen Originalanweisung zur Bearbeitung der STAI-Stateitems, ergänzt um eine Kurzfassung der vorher besprochenen allgemeinen Testeinweisung und

der erneuten Angabe der Telephonnummer des Untersuchungsleiters. Hinter diesem Instruktionsblatt steckte ein zweites Blatt, das noch einmal die Hinweise zur Behebung von Störungen des Signalgebers beinhaltete (Anhang B). (Durch diese Informationsredundanz war sichergestellt, daß die Vpn alle wichtigen Informationen auch vor Ort im Feld ständig zur Verfügung hatten.) Die Vpn bearbeiteten darauf den ersten Protokollbogen, der allerdings nicht in die Auswertung einging, sondern später vernichtet wurde. Es wurde auf richtiges und vollständiges Ausfüllen (insbesondere Datum und Uhrzeit) geachtet. Die einzelnen Items wurden inhaltlich nicht weiter besprochen, es sei denn, die Vpn verlangten dies ausdrücklich zur Beseitigung von Verständnisschwierigkeiten. Lediglich bei dem Item "Ich fühle mich gewachsen" wurde regelmäßig darauf hingewiesen, daß es sich dabei um ein "einer Situation gegenüber gewachsen sein" handele und nicht um ein "körperlich größer geworden sein".

Als die Vpn hiermit fertig waren, öffneten sie eine Mappe, in der sich alle Instrumente der Vorbefragung befanden (vgl. Kap. 3.2.2.1.), und zwar in der Reihenfolge: selbstkonstruierter Fragebogen, PRF, STAI-Traitteil und SVF. Jede Vp bearbeitete jedes Instrument selbständig für sich. Wenn sie mit einem Instrument fertig war, wartete sie bei Gruppeneinweisungen bis alle anderen Vpn auch fertig waren, damit der Untersuchungsleiter zu Beginn des jeweils folgenden Instruments (PRF, STAI-Traitteil, SVF) die dazugehörige Anweisung zu seiner Bearbeitung verlesen konnte. Auch bei Vorbefragungen von nur einer einzigen Vp wurde die jeweilige Anweisung vom Untersuchungsleiter verlesen.

Auf dem selbstkonstruierten Fragebogen zur Erhebung soziodemographischer Daten mußten die Psychologiestudenten, die dem Untersuchungsleiter alle vorher nicht bekannt waren, ihren vollen Namen angeben. Dies hatte seinen Sinn zum einen darin, eine Referenzangabe bei der Bescheinigung der Vp-Stunden zu haben, zum anderen geschah es aber auch als eine Art Sicherheitsmaßnahme, denn die ausgehändigten Signalgeber hatten einen Wert von ca. DM 100,- und waren nicht versichert. Auf das Unterschreiben einer formellen Empfangsbescheinigung durch die Vpn wurde jedoch in allen Fällen verzichtet. Außer dieser Namensangabe lag dem Untersuchungsleiter während der Datenerhebung, aber auch noch während der Aufbereitung und Erfassung der Daten, eine einzige Liste vor, die die Codenummern der Vpn zusammen mit ihren entsprechenden vollen Namen enthielt. Ansonsten waren und sind auf allen Frage- und Protokollbögen und sonstigen Materialien zur eindeutigen Zuordnung lediglich die Codenummern notiert. (Auch auf den Titelseiten der Erhebungsinstrumente PRF, STAI-Traitteil und SVF sind statt der Namen nur die Codenummern eingetragen.) Nach Ausstellen der Vp-Scheine und vor Beginn der Datenanalyse wurden die Namen aus den soziodemographischen Fragebögen ausgeschnitten und zusammen mit der Liste der Namen aller Vpn vernichtet. Es gibt also keine Zuordnung von Namen zu Codenummern mehr und von den Psychologiestudenten auch kein namentliches Gesamtverzeichnis aller Vpn mehr. Es sind in keinem Falle Namensangaben im Rechner erfaßt worden. Das gesamte Untersuchungsmaterial liegt also nur noch in namentlich völlig anonymisierter Form vor.

Nach dem Bearbeiten der Fragebögen wurde abschließend das zweite Stimmungsprotokoll erstellt, indem die Vpn den zweiten Protokollbogen im Ringbuch ausfüllten. Dieser ging als erster gültiger Protokolltermin in die Auswertung ein. Dies ist zwar einerseits ungünstig, denn es handelt sich bei

der Situation der Testeinweisung und Vorbefragung nicht um eine Situation aus dem Alltag der Vpn und außerdem ist sie eine künstlich weitgehend gleichgehaltene Situation für alle Vpn, womit das erhobene Stimmungsprotokoll für eine ESM-Studie eher untauglich ist. Da im voraus aber weder die Resonanz bei der Rekrutierung der Vpn klar war noch die Compliance im Feld, insbesondere die Zahl wahrgenommener Protokolltermine, absehbar war, war andererseits geplant, dieses Protokoll aus der Testeinweisung doch zu verwenden, um dadurch die gesicherte Datenbasis zu verbreitern. Die 45 im Ringbuch vorhandenen Protokollbögen sollten somit folgendermaßen verwendet werden: einer als reiner Übungsbogen, einer als zu Übungszwecken eingesetzter, aber auszuwertender Protokollbogen, 41 für die im Signalgeber eingespeicherten Termine sowie zwei weitere als Ersatz für eventuell verdorbene Protokollbögen. Der Zeitaufwand für die Testeinweisung und Vorbefragung betrug etwa eine bis eineinhalb Stunden.

Während der folgenden sieben Tage des Beobachtungszeitraums nahm der Untersuchungsleiter ein- bis zweimal mit den Vpn telephonisch Kontakt auf, um sich vom ordnungsgemäßen Ablauf zu überzeugen bzw. um ein allgemeines Feedback zu erhalten. Ansonsten lief die Protokollierung im Regelfall ohne weitere Supervision oder gar Eingriffe ab. Zwischendurch sammelte der Untersuchungsleiter die im Institut I abgegebenen, schon bearbeiteten Protokollbögen ein.

Frühestens am Abend des siebten Tages des Beobachtungszeitraums fand die Nachbefragung statt. Dieses Debriefing erfolgte in den gleichen Räumlichkeiten wie die Testeinweisung und Vorbefragung. Die Vpn brachten zu diesem Termin das gesamte sich noch in ihrem Besitz befindliche Material mit. Die Signalgeber wurden dabei auf ordnungsgemäßen Zustand überprüft und die Protokollbögen gezählt, sortiert und auf Vollständigkeit geprüft. Noch vor jeder mündlichen Diskussion der Erlebnisse der Vpn während des Beobachtungszeitraums wurde ihnen der Nachbefragungsbogen (vgl. Kap. 3.2.2.4.) mit der Aufforderung ausgehändigt, die Anleitung zum Ausfüllen durchzulesen und ihn dann selbständig zu bearbeiten. Im Anschluß daran wurden die Vp-Scheine eingesammelt und den Nicht-Psychologiestudenten eine Flasche Sekt als Anerkennung übergeben. Allen Vpn wurde darauf mündlich die im Nachbefragungsbogen nicht enthaltene Frage gestellt, ob ihnen während der Untersuchung an der Struktur der Protokollbögen etwas aufgefallen sei, und speziell, ob sich vielleicht in den Protokollbögen Items wiederholt hätten, und wenn ja, welche das gewesen seien. Im weiteren Verlauf der Diskussion dieser Frage wurden die Vpn über die tatsächlichen Strukturen der Protokollbögen und über die Fragestellung der Untersuchung informiert. Der Zeitaufwand für die Nachbefragung betrug zwischen einer halben und einer Stunde. Legt man außerdem einen zeitlichen Aufwand zum Bearbeiten eines einzigen Protokollbogens von zwei bis vier Minuten zugrunde und berücksichtigt man für die Psychologiestudenten den zweimaligen Weg zum Institut I zur Zwischenabgabe der Protokollbögen, so läßt sich ein Gesamtnettozeitbedarf vom Beginn der Testeinweisung und Vorbefragung an bis zum Ende der Nachbefragung von etwa vier bis sechs Stunden je Vp schätzen. Die Psychologiestudenten erhielten dennoch alle 20 Vp-Stunden bescheinigt, da ein Ausgleich dafür erfolgen mußte, daß sie neben dieser Zeit, in der sie tatsächlich mit der Untersuchung beschäftigt waren, noch an sieben Tagen jeweils zwölf Stunden in Bereitschaft sein mußten. Schließlich wußten sie nicht, wann ein Protokolltermin eintreten würde und mußten deshalb den Signalgeber und das Ringbuch ständig mit sich führen.

3.3. Aufbereitung und Erfassung der Daten

Die Aufbereitung der Daten und die Eingabe in den Rechner wurden z.T. vom Untersuchungsleiter selbst, z.T. aber auch von studentischen Hilfskräften durchgeführt. Die Datenaufbereitung umfaßte alle vorbereitenden Maßnahmen, die nötig waren, damit die Angaben aus den Protokoll- und Fragebögen schnell und präzise in den Rechner eingegeben und ökonomisch verrechnet werden konnten.

Das Wichtigste war dabei die Kontrolle der bearbeiteten Protokollbögen hinsichtlich der Einhaltung der dreißigminütigen Latenzzeit sowie der Richtigkeit ihrer chronologischen Reihenfolge, die in Einzelfällen tatsächlich von der numerischen abwich. Beides mußte für jeden einzelnen Protokollbogen jeder einzelnen Vp anhand des Weckterminplans und anhand des Datums der Testeinweisung überprüft werden. Protokolle, die nach Überschreiten der maximalen Latenzzeit ausgefüllt worden waren oder bei denen aus anderen Gründen eine Zuordnung zu Datum und Uhrzeit eines vorgesehenen Protokolltermins nicht erfolgen konnte, wurden aussortiert und vernichtet. Die Protokolle wurden für jede Vp chronologisch geordnet, die Latenzzeit wurde in Minuten als Differenz zwischen der tatsächlichen Zeit des Protokollierens und der vorgesehenen Weckzeit auf dem Protokollbogen notiert. Über die ausgelassenen Protokolltermine wurde eine Liste angelegt.

Die Kontrolle der Instrumente aus Vor- und Nachbefragung beinhaltete eine Überprüfung der formalen Richtigkeit der Angaben, wo z.B. nur eine Nennung vorgesehen war, wurden Mehrfachnennungen als fehlende Angabe markiert, oder wenn bei einer Frage des Nachbefragungsbogens die Antwortalternative "ja" näher begründet werden sollte, so sollten die Eintragungen in freier Form auch tatsächlich eine Begründung des "ja" darstellen und nicht etwa Argumente für "nein" liefern oder sich auf ganz andere Sachverhalte beziehen. Gegebenenfalls wurden dann einzelne Angaben für die logisch korrekte Erfassung verändert oder ganz gestrichen.

Im Rahmen der Datenaufbereitung wurden die verbalen Antworten auf die offenen Fragen aus Vor- und Nachbefragung in Zahlen umgewandelt und die Angaben aus PRF, STAI-Traitteil und SVF zu Summenwerten aggregiert. Die Codierung der offenen Fragen erfolgte mittels eines nach inhaltsanalytischen Gesichtspunkten angelegten Codeplans, der empirisch aufgrund der schriftlichen Äußerungen der Vpn erzeugt wurde. Dabei wurden je Frage über die bearbeiteten Fragebögen aller Vpn hinweg den semantisch gleichen oder sehr ähnlichen Kommunikationseinheiten (z.B. einzelnen Worten, Teilsätzen, ganzen Sätzen oder auch mehreren ganzen Sätzen auf einmal) Codenummern zugeordnet, die auf den Fragebögen neben den dazugehörigen Fragen notiert und später erfaßt wurden. Bei der PRF wurden die symptomatischen Antworten mit Hilfe einer Schablone ermittelt, in ein Schema übertragen und dort je Merkmalsskala zu Rohwerten addiert. Beim STAI-Traitteil wurden die Werte der Ratingskalen aller Items, gegebenenfalls nach Umpolung, zu einem STAI-Trait-Rohwert addiert. Anschließend wurden die Rohwerte beider Instrumente anhand der entsprechenden Normtabellen in Staninewerte umgerechnet. Bei dem SVF wurden durch Übertragen der Ratingskalenwerte jedes Items in ein Additionsschema die Merkmalsskalenwerte ermittelt. Hier konnten keine Normwerte errechnet werden (vgl. Kap. 3.2.2.1.).

Die Erfassung der Daten erfolgte getrennt für die Vorbefragung, die Nachbefragung und die Protokollbögen. Grundsätzlich wurden dabei fehlende Angaben, unzulässige doppelte Angaben oder anderweitig unklare Angaben als fehlende Werte erfaßt. Die Protokollbögen wurden im ASCII-Format so erfaßt, daß je Vp und je Protokolltermin eine Zeile vorgesehen war. Ausgelassene oder ungültige Protokolltermine (s.o.) wurden als Leerzeile erfaßt, so daß jeder Protokolltermin an der Stelle seines vorgesehenen chronologischen Auftretens positioniert wurde. Die so entstandenen 42-zeiligen Matrizen je Vp wurden nacheinander erfaßt, d.h., daß nach den 42 Protokollterminen der ersten Vp die 42 Protokolltermine der zweiten Vp kommen usw. Die Variablen stehen dabei stets spaltenweise untereinander.

3.4. Datenanalyse und Ergebnisse

Von wenigen Ausnahmen abgesehen, bei denen ein Taschenrechner verwendet wurde, erfolgte die gesamte Datenanalyse rechnergestützt auf einem IBM-kompatiblen PC mit dem Programmpaket SPSS/PC+ (BROSIUS, 1988, 1989, SCHUBÖ, UEHLINGER, PERLETH, SCHRÖGER & SIERWALD, 1991, UEHLINGER, 1988), Version 4.0.1 .

Bei der statistischen Überprüfung von Unterschieds- oder Zusammenhangshypothesen wurde generell ein Signifikanzniveau von 5% zugrundegelegt. Betrug die Wahrscheinlichkeit für das Auftreten der aus den empirischen Stichprobenwerten errechneten Prüfgröße bei Gültigkeit der Nullhypothese $P \leq .05$ (Irrtumswahrscheinlichkeit), wurde der Unterschied bzw. der Zusammenhang als signifikant eingestuft und die Nullhypothese zugunsten der (ungerichteten) Alternativhypothese abgelehnt. Die Irrtumswahrscheinlichkeit beschreibt dabei die maximale Wahrscheinlichkeit, mit der eine gültige Nullhypothese bei einem signifikanten Ergebnis fälschlicherweise verworfen werden kann (α-Fehler). Eine Irrtumswahrscheinlichkeit von $P \leq .10$ (aber $P > .05$) wurde als Tendenz interpretiert und beschreibt damit einen Zusammenhang bzw. Unterschied, der weniger aussagekräftig ist, da er noch mit maximal zehnprozentiger Wahrscheinlichkeit hätte zufällig zustande gekommen sein können. Demgegenüber wurde ein Ergebnis mit einer Irrtumswahrscheinlichkeit von $P \leq .01$ als hochsignifikant eingeordnet (vgl. hierzu, zur deskriptiven Statistik und zu den gängigen inferenzstatistischen Verfahren nebst ihren Anwendungsvoraussetzungen BORTZ, 1984, 1989, BORTZ, LIENERT & BOEHNKE, 1990, CLAUSS & EBNER, 1985, SACHS, 1974).

3.4.1. Homogenität und Repräsentativität der Stichprobe

Das hier zu konstruierende Instrument zur Messung der Befindlichkeit ist zunächst nur für eine Anwendung bei Angehörigen der Normalpopulation unter Alltagsbedingungen vorgesehen und nicht für Untersuchungen von präselegierten Populationen wie etwa im klinischen Bereich. Ziel der Analyse der Daten aus der Testeinweisung und Vorbefragung war daher die Überprüfung der Homogenität der Gesamtstichprobe und eine bescheidene Abschätzung ihrer Repräsentativität für die Gesamtpopulation der Bundesrepublik (Ergebnisübersichten vgl. Anhang A). Die 74 Vpn setzten sich aus 41 (55%) Frauen und 33 (45%) Männern zusammen; 43 (58%) von ihnen waren Studierende der Psychologie, 31 (42%) gehörten zur Gruppe der Nicht-Psychologiestudenten.

Die Homogenität der Stichprobe wurde durch Analyse der Meßwerte der Nicht-Psychologiestudenten versus Psychologiestudenten und Frauen versus Männer überprüft. Dabei zeigt sich, wenn man zunächst das Alter der Teilstichproben untersucht, daß es zwischen diesen Untergruppen keine bedeutsamen Unterschiede im Mittelwert gibt (t-Test für unabhängige Stichproben). Es liegt in allen Untergruppen in der Nähe des Gesamtmittelwerts von 29.8 Jahren (Gesamtverteilung: Range = 28 Jahre, kleinster Wert 21, größter Wert 49 Jahre). Prüft man die Altersverteilung aber mit Hilfe einer zweifaktoriellen Varianzanalyse (BACKHAUS, ERICHSON, PLINKE & WEIBER, 1990, BORTZ, 1989, BROSIUS, 1988) mit dem Alter als der abhängigen und den Geschlechtern und den (Nicht-) Psychologiestudenten als den unabhängigen Variablen, so zeigt sich neben der Insignifikanz für die Haupteffekte eine hochsignifikante Wechselwirkung zwischen den unabhängigen Variablen ($P<.01$). Sie äußert sich darin, daß die männlichen Nicht-Psychologiestudenten mit 32.3 Jahren und die weiblichen Psychologiestudentinnen mit 31.3 Jahren im Mittel deutlich älter sind als die weiblichen Nicht-Psychologiestudentinnen mit 27.1 und die männlichen Psychologiestudenten mit 28.5 Jahren. Darüber hinaus besteht jedoch kein Zusammenhang zwischen dem Geschlecht und der Zugehörigkeit zu den Gruppen der (Nicht-) Psychologiestudenten (Vierfelder Chi-Quadrattest). Die Gruppe der Nicht-Psychologiestudenten war prinzipiell auch für Vpn mit einem anderen Schulabschluß als dem Abitur offen, doch zeigt es sich, daß nur 6 Vpn eine solche andere Schulbildung angaben, davon immerhin noch 4 Vpn mit Fachhochschulreife. Es läßt sich somit per Augenschein sagen, daß es zwischen beiden Gruppen keine relevanten Unterschiede in der Schulbildung gibt; dies gilt in verstärktem Maße auch für die beiden Geschlechtergruppen.

Etwas anders sieht es bei der beruflichen Vorbildung aus: 39% aller Vpn, und zwar genau derselbe Prozentsatz für Frauen und Männer, geben an, keine abgeschlossene Berufsausbildung zu haben. Von den Nicht-Psychologiestudenten geben aber mehr als die Hälfte (52%) an, keine Berufsausbildung zu haben, während es bei den Psychologiestudenten nur weniger als ein Drittel (30%) sind. Betrachtet man diejenigen, die eine abgeschlossene Berufsausbildung vorweisen können, so zeigt sich (in Relation zum Gesamt aller Vpn der jeweiligen Untergruppe) im Bereich der nicht-akademischen Ausbildungen (Lehre, Berufsfachschule etc.) eine Dominanz der sozialen, pädagogischen und Heilberufe (Krankenpfleger, Erzieher, Krankengymnast, Heilpraktiker, Schule für Atmung) bei den Psychologiestudenten (30%) gegenüber den Nicht-Psychologiestudenten (3%). Auch bei den kaufmännischen Berufen (Kaufmann, Buchhändler, Wirtschaft, Hotelfach) zeigt sich ein Übergewicht auf der Seite der Psychologiestudenten (22%) verglichen mit den Nicht-Psychologiestudenten (9%). Demgegenüber haben 19% aller Nicht-Psychologiestudenten eine handwerklich-technische Ausbildung (Handwerker, Laborant, Techniker), bei den Psychologiestudenten sind dies nur 2%. Im Bereich der akademischen Ausbildungen liegen die Nicht-Psychologiestudenten mit 24% klar vor den Psychologiestudenten mit nur 6%. Zwischen Männern und Frauen gibt es keine so deutlichen Unterschiede in der beruflichen Vorbildung. Wegen des hohen Prozentsatzes ungültiger oder verweigerter Antworten (16% aller Vpn) auf die Frage nach dem Lebensunterhalt läßt sich zu diesem Punkt nur wenig sagen. Immerhin leben 63% aller Männer von der eigenen Berufs- / Erwerbstätigkeit, während es bei den Frauen nur 44% sind; etwa gleich viele Frauen (41%) leben von der Unterstützung anderer, während dies nur 18% der Männer tun. Gut die Hälfte aller Psychologiestudenten (51%) befand sich zur Zeit der Untersuchung im ersten Fachsemester, weitere 37% waren im zweiten Fachsemester.

Die zu Staninewerten umgerechneten Meßwerte der einzelnen Merkmalsskalen der PRF wurden zwischen den beiden Paaren von Teilstichproben unter Verwendung ihrer ordinalen Informationen auf Unterschiede hinsichtlich ihrer zentralen Tendenz untersucht (U-Test nach MANN & WHITNEY). Vorher erfolgte jedoch die Auswertung der Validitäts- oder Infrequenzskala. Sie ergab für 99% aller Vpn unauffällige Werte zwischen 0 und 2 Rohpunkten; lediglich eine Psychologiestudentin erreichte den für diese Skala kritischen Grenzwert von 3 Punkten, bei dem die gesamte PRF nur unter Vorbehalt zu interpretieren ist (bei mehr als 3 Punkten sollte auf die Auswertung der PRF gänzlich verzichtet werden) (STUMPF et al., 1985). Eine Inspektion der übrigen Daten dieser Vp zeigte aber keinerlei Auffälligkeiten, so daß auch ihr PRF-Fragebogen mit in die Gesamtdatenanalyse einging. Für den Geschlechtersplit zeigt sich, daß es signifikante Unterschiede in den Merkmalen Aggressivität, Dominanzstreben und Anlehnungsbedürfnis gibt sowie von der Tendenz her einen Unterschied in der Skala Geselligkeit. In allen Fällen sind diese Unterschiede so beschaffen, daß den Frauen die höheren Werte und damit die stärkeren Merkmalsausprägungen zukommen. Die Nicht-Psychologiestudenten unterscheiden sich dagegen in nur zwei Merkmalen von den Psychologiestudenten, indem sie im Bereich Spielerische Grundhaltung signifikant höhere und im Merkmal Hilfsbereitschaft von der Tendenz her niedrigere Werte haben.

In der Traitangst, gemessen mit dem STAI-Traitteil, unterscheiden sich die Frauen durch signifikant höhere Werte von den Männern. Auch hier wurden die aus den Rohwerten gewonnenen Staninewerte verwendet (U-Test). Zwischen Psychologie- und Nicht-Psychologiestudenten gibt es keine Unterschiede in diesem Merkmal.

Bei der Analyse der SVF-Daten mußte in Ermangelung von Normtabellen auf die Rohwerte zurückgegriffen werden; bei Vorliegen von normalverteilten Daten sollte der t-Test für unabhängige Stichproben verwendet werden. Zur Überprüfung der angenommenen Normalverteilung wurde je Merkmalsskala ein KOLMOGOROV-SMIRNOV-Anpassungstest mit den empirisch aus der Stichprobe errechneten Parametern (Mittelwert und Standardabweichung) durchgeführt, ohne daß das Programm SPSS/PC+ allerdings die Schrankenwerte von LILLIEFORS (1967, zit. bei BORTZ, LIENERT & BOEHNKE, 1990, S. 322) berücksichtigt. Deren Anwendung zur Signifikanzbestimmung wäre hier aber indiziert, da Mittelwert und Standardabweichung beider Verteilungen bei dieser Verfahrensweise identisch sind und der KOLMOGOROV-SMIRNOV-Test nur noch auf andere Verteilungsunterschiede ansprechen kann als die, die durch diese beiden Parameter determiniert sind. Dies führt dazu, daß die Prüfgröße Z, die aus der größten absoluten Ordinatendifferenz zwischen den kumulierten empirischen und theoretischen Verteilungen bestimmt wird, eher zu niedrig und damit ihre Auftretenswahrscheinlichkeit zu hoch ausfällt und somit eher eine Ergebnisinterpretation im Sinne einer Stützung der Nullhypothese nahegelegt wird (BORTZ, LIENERT & BOEHNKE, 1990, SCHUBÖ et al., 1991). Von den für alle 19 SVF-Merkmalsskalen durchgeführten KOLMOGOROV-SMIRNOV-Anpassungstests wurde nur ein einziger signifikant. Deshalb wurden zur Prüfung der Gruppenunterschiede in diesen Skalen trotz der eben beschriebenen methodischen Mängel t-Tests angewandt. Signifikante Mittelwertunterschiede zeigen sich einmal in der Skala Herunterspielen durch Vergleich mit anderen, in der die Männer einen höheren Mittelwert haben als die Frauen. In den Skalen Bedürfnis nach sozialer Unterstützung, Fluchttendenz und Aggression haben die Frauen im Schnitt signifikant und in den Skalen Gedankliche

Weiterbeschäftigung sowie Resignation immerhin noch tendenziell höhere Werte als die Männer. Die Nicht-Psychologiestudenten unterscheiden sich von den Psychologiestudenten nur in einem SVF-Merkmal signifikant, nämlich in der Pharmakaeinnahme, wo sie einen höheren Mittelwert aufweisen. Von der Tendenz her neigen sie auch stärker zur Selbstbeschuldigung, während sie im Merkmal Schuldabwehr tendenziell niedrigere Werte haben als die Psychologiestudenten.

Zusammenfassend läßt sich sagen, daß hinsichtlich der Alterszusammensetzung zwischen den Unterstichproben im Mittel keine Unterschiede bestehen, daß es jedoch eine Wechselwirkung zwischen den beiden Gruppierungsfaktoren gibt. Diese läßt vermuten, daß es ähnliche Verteilungen für solche Merkmale in der Gesamtstichprobe gibt, die hoch mit dem Alter korreliert sind. (Das Alter ist aber lediglich mit zwei der hier erhobenen Persönlichkeitsmerkmale, die als Rohwerte verrechnet wurden, höher als zu r=.30 korreliert, einmal zu r=.46 und einmal zu r=.31.) Ansonsten gibt es zwischen den beiden Geschlechtern mehr Unterschiede hinsichtlich der erhobenen Persönlichkeitsmerkmale als zwischen den Psychologiestudenten und den Nicht-Psychologiestudenten. Letztere unterscheiden sich jedoch erheblich in Umfang und Art der beruflichen Vorbildung. Das Psychologiestudium als Sozialisationsfaktor ist bei den Psychologiestudenten zum Erhebungszeitpunkt offensichtlich noch nicht lange wirksam gewesen: 88% von ihnen befanden sich im 1. oder 2. Semester.

Es sei bereits hier darauf hingewiesen, daß die Interpretation einer größeren Anzahl jeweils unabhängig voneinander durchgeführter Signifikanzprüfungen bei gegebenem Signifikanzniveau vorsichtig zu interpretieren ist: Werden aus derselben Grundgesamtheit (also bei gültiger Nullhypothese) für dasselbe oder verschiedene Merkmale wiederholt Paare von Stichproben gleichen Umfangs gezogen und auf Unterschiede z.B. in ihrer zentralen Tendenz untersucht, so beträgt die Wahrscheinlichkeit, unter diesen Stichprobenpaaren auch solche zu finden, deren Mittelwerte sich zufällig so sehr voneinander unterscheiden, daß die Wahrscheinlichkeit für das Auftreten der unter den Stichprobenbedingungen errechneten Prüfgröße noch unter die Irrtumswahrscheinlichkeit sinkt, bei einer vorgegebenen Irrtumswahrscheinlichkeit von z.B. 5% eben genau 5%. Mit anderen Worten: Gibt es in Wirklichkeit keine Unterschiede z.B. zwischen den Psychologie- und den Nicht-Psychologiestudenten in den erhobenen Merkmalen, so werden trotzdem bei zwanzig durchgeführten t-Tests 5% von ihnen, also einer, zufällig signifikant werden.

Eine Abschätzung der Repräsentativität der Stichprobe erfolgte für zwei Bereiche: Zum einen für das Verhältnis von Frauen zu Männern, zum anderen für die normierten Erhebungsverfahren (PRF und STAI-Traitteil), bei denen Vergleichswerte aus Eichstichproben vorliegen. Das Geschlechterverhältnis in der Gesamtstichprobe beträgt 55.4% Frauen zu 44.6% Männern. Das Verhältnis in der Bevölkerung beträgt dagegen 51.7% Frauen zu 48.3% Männern (Statistisches Bundesamt, 1992). Gefragt ist danach, ob die empirisch beobachteten Häufigkeiten in der Stichprobe nur zufällig von den aufgrund der Populationsnorm zu erwartenden Häufigkeiten abweichen (Nullhypothese). Der durchgeführte Chi-Quadrattest zeigt in der Tat, daß die empirischen Häufigkeiten nur zufällig von den erwarteten abweichen (Chi-Quadrat = .4, df=1, nicht signifikant), wodurch die Annahme der Repräsentativität in diesem Punkt gestützt wird. Als nächstes soll überprüft werden, ob die in der Stichprobe gefundenen empirischen Verteilungen der Staninewerte bei den

Merkmalsskalen der PRF und beim STAI-Traitteil von der theoretisch zu erwartenden abweichen. Unter der Voraussetzung nämlich, daß die hier untersuchte Gesamtstichprobe repräsentativ für die Eichstichproben der beiden Instrumente ist und diese wiederum repräsentativ für die Bevölkerung sind, müßten die empirischen Staninewerte der theoretisch zu erwartenden (Normal-) Verteilung mit einem Mittelwert von $\bar{x}$=5 und einer Standardabweichung von annähernd s=2 (genauer: s=1.96) (LIENERT, 1989, S. 342) folgen. Dies wurde für die 14 PRF-Merkmalsskalen und die STAI-Traitskala mit Hilfe des KOLMOGOROV-SMIRNOV-Anpassungstests überprüft. Der Anpassungstest ergibt dabei für drei der 15 Skalen lediglich tendenzielle Abweichungen von der angegebenen Normalverteilung; bei allen anderen 12 Skalen sind die Abweichungen jedoch (z.T. hoch) signifikant. Darüber hinaus dürfte es, da die Staninenormierung geschlechtsspezifisch und auch altersspezifisch erfolgt (was eine unmittelbare Vergleichbarkeit der Staninewerte auch von solchen Vpn ermöglicht, die sich im Hinblick auf Alter und Geschlecht unterscheiden), keine bedeutsamen Mittelwertunterschiede zwischen den Männern und Frauen geben, von Zufallssignifikanzen (s.o.) abgesehen. Dieser Punkt wird für die PRF und das STAI ebenfalls nicht bestätigt, denn von den 15 durchgeführten U-Tests auf Geschlechtsunterschiede wurden immerhin 4 (27%) signifikant.

Hinsichtlich der Repräsentativität der Stichprobe für die (Normal-) Bevölkerung läßt sich somit feststellen, daß sie für die untersuchten Persönlichkeitsmerkmale offenbar auch nicht annähernd gegeben ist. Lediglich das Verhältnis von Männern zu Frauen entspricht dem in der Grundgesamtheit.

3.4.2. Ergebnisse zum State-Trait-Modell von BUSE & PAWLIK (1991)

Zentrale Voraussetzung für das der Instrumentenentwicklung zugrundeliegende testtheoretische Modell (vgl. Kap. 2.1.3.) ist die Erhebung von jeweils zwei Meßwerten (pro Merkmal und Vp) zu jedem einzelnen von mehreren aufeinanderfolgenden Meßzeitpunkten. Dabei sollen alle einzelnen Meßwerte möglichst unabhängig von dem Vorgang des Messens selbst sein. Insbesondere unterstellt das Modell, daß die Ausprägung des wahren Statewertes je Meßzeitpunkt während der beiden Einzelmessungen konstant bleibt und nicht einer (z.B. meßbedingten) systematischen Veränderung unterworfen wird (vgl. auch Kap. 2.2.3.); denn nur dann macht es Sinn, entsprechend dem Reliabilitätskonzept der KT (vgl. Kap. 2.1.1.2.) aus der gemessenen Fluktuation der Testwerte über beide Meßzeitpunkte mittels Korrelation der ersten mit der zweiten Meßwertreihe auf die unsystematisch wirkende Fehlerkomponente schließen zu wollen. Die ESM-Methode erfordert darüber hinaus, daß sich die Merkmalsausprägung über den gesamten Beobachtungszeitraum hinweg nicht infolge der Datenerhebung selbst systematisch verändert. Die Interpretation der ermittelten Modellparameter und damit die Qualität eines auf ihnen aufbauenden Meßinstruments hängt entscheidend vom Vorhandensein und dem Umfang entsprechender Reaktivitätseffekte in der durchgeführten Feldstudie ab. Sie werden im Anschluß an die Darstellung der Modellparameter untersucht.

3.4.2.1. Modellparameter auf Itemebene

In dem Modell von BUSE & PAWLIK (1991) (vgl. Kap. 2.1.3.) geht es um die Aufspaltung der Testwerte in eine wahre Trait-, eine wahre State- und in eine unsystematische Fehlerkomponente und damit einhergehend um die Aufspaltung der Testwertevarianz in die Varianz der Trait-, der State- und der Fehlerwerte. Die drei letztgenannten Parameter beziehen sich je Item auf die gesamte Vpn x Zeitpunkte-Matrix und gelten nicht etwa für einzelne Vpn oder Zeitpunkte allein. (Für jedes der in Kap. 3.2.2.2. aufgeführten 10 relevanten Befindlichkeitsitems liegen zwei solcher Meßwertmatrizen vor, eine für alle ersten Messungen je Meßzeitpunkt, d.h. je Protokolltermin, und eine für alle zweiten Messungen.) Aus ihnen lassen sich die State- und Trait-Charakteristiken ableiten, die der Bestimmung der Änderungssensitivität eines jeden Items dienen. Außerdem läßt sich mit der so aufgespaltenen Testwertevarianz eine Reliabilitätsbestimmung vornehmen.

Zur Errechnung der drei Varianzanteile je Item werden benötigt:

1. die interindividuelle Varianz der je Vp über alle Meßzeitpunkte gemittelten Testwerte (d.h. die Varianz der intraindividuellen Mittelwerte),

2. die über alle Vpn gemittelte intraindividuelle Varianz der Testwerte zwischen allen Meßzeitpunkten (d.h. die mittlere intraindividuelle Testwertevarianz) sowie

3. die über alle Vpn gemittelte intraindividuelle Korrelation zwischen der Testwertereihe der jeweils ersten Messung je Meßzeitpunkt und der Reihe der jeweils zweiten Messung je Protokolltermin (d.h. die mittlere intraindividuelle Meßwertekorrelation zwischen erster und zweiter Messung).

Außerdem sind Informationen über die Zahl der von jeder einzelnen Vp wahrgenommenen Protokollterminе und damit auch über die Gesamtzahl aller überhaupt wahrgenommenen Meßzeitpunkte erforderlich. Von den 3108 möglichen Protokollterminen (74 Vpn mit jeweils 42 möglichen Protokollterminen) gingen 2669 (85.9%) als gültige Protokolle in die Auswertung ein, was einer mittleren Reaktionsrate von 36.1 wahrgenommenen Protokollterminen je Vp entspricht. Dabei nahmen die Vpn min. 23 und max. alle der 42 angebotenen Protokollterminе wahr.

Die Errechnung der Varianz der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ war simpel: Die Mittelwerte wurden unter Beachtung der unterschiedlichen Anzahl wahrgenommener Protokollterminе je Vp gebildet und aus ihnen interindividuell ein mittlerer Mittelwert mit der dazugehörigen Verteilungsvarianz bestimmt. Zur Errechnung der mittleren intraindividuellen Testwertevarianz $\bar{s}^2_i$ wurde zunächst für jede Vp gesondert deren intraindividuelle Testwertevarianz bestimmt und diese mit der Zahl der (intraindividuellen) Freiheitsgrade (df = Zahl der wahrgenommenen Protokollterminе - 1) gewichtet. Diese 74 intraindividuell gewichteten Varianzen wurden anschließend addiert und durch die Summe der Freiheitsgrade (d.h. 2669 - 74 = 2595) dividiert. Für die Bestimmung der mittleren intraindividuellen Korrelation $\bar{r}_{ii'}$ wurde zunächst die intraindividuelle

Testwertekorrelation für jede Vp einzeln ermittelt. Diese Korrelationskoeffizienten wurden in FISHER's Z-Werte transformiert (mit Z = FISHER's Z):

$$(56) \quad Z = \frac{1}{2} * \ln \left[\frac{1+r}{1-r} \right] \qquad \text{(BORTZ, 1989, S. 262)}$$

Wie aus dieser Beziehung deutlich wird, ist für r=1 kein Z definiert. Dieser Fall trat durchaus gelegentlich auf, nämlich immer dann, wenn eine Vp tatsächlich zu jedem einzelnen Protokolltermin bei der zweiten Messung exakt denselben Wert angegeben hat wie bei der ersten Messung. Außerdem kam es bei einigen Items (z.B. bei "Ich bin ängstlich") vor, daß intraindividuell über alle Protokolltermine einer Vp hinweg überhaupt keine Varianz der Meßwerte auftrat, mithin daß die Vp stets und ohne Ausnahme über alle Protokoltermine denselben Wert angegeben hat. Wenn aber die Varianz null ist, ist auch die Standardabweichung null und damit, wie etwa aus (12) zu ersehen ist, die Korrelation nicht definiert. Bei der Errechnung der mittleren intraindividuellen Korrelation konnte also nicht in allen Fällen auf eine Basis von n=74 gültigen Ausgangskoeffizienten zurückgegriffen werden. In jedem Falle wurden die erfolgreich transformierten Korrelationskoeffizienten addiert und durch ihre jeweilige Basiszahl dividiert, um so mittlere Z-Werte zu erhalten. Anschließend erfolgte die Rücktransformation in Korrelationskoeffizienten (BORTZ, 1989, S. 263).

Im nächsten Schritt wurden aus den so aufbereiteten Basiswerten die Varianzen der Trait-, State- und Fehlerwerte gemäß den in Kap. 2.1.3. angegebenen Rechenvorschriften bestimmt. Die Varianz der Traitwerte s^2_t wurde wie in (46) errechnet; dabei wurde für n (Zahl der Protokolltermine) deren mittlere Anzahl (36.1; s.o.) eingesetzt. Die beiden anderen Varianzen ergaben sich durch Anwendung von (51) für die Varianz der Statewerte s^2_s und von (52) für die der Fehlerwerte s^2_e.

Die State-Charakteristik S ist der hier entscheidende Parameter zur Bestimmung der Änderungssensitivität der Items, zu dessen Errechnung in (53) die Varianz der Statewerte zur Summe aus der Varianz der State- und der Traitwerte ins Verhältnis gesetzt wird. Ein Item kann als änderungssensitiv eingestuft werden, wenn deutlich mehr als 50% der Varianz der wahren Werte auf die Varianz der Statewerte zurückzuführen ist. Die State-Charakteristik addiert sich mit der Trait-Charakteristik aus (54), ihrem Komplementärwert, zu eins. Schließlich kann die Traitvarianz noch zur Summe aus Trait- plus Fehlervarianz entsprechend (55) in Beziehung gesetzt werden, um ein statefreies (Trait-) Reliabilitätsmaß im Sinne von (24) zu erhalten. Tabelle 1 listet die hier besprochenen Kennwerte für alle zwei mal zehn Items auf. Die in der ersten Spalte stehenden mittleren intraindividuellen Mittelwerte entsprechen den Gesamtmittelwerten über alle Vpn und alle Zeitpunkte.

Tab. 1 Parameter auf der Basis des Modells von BUSE & PAWLIK (1991)

		$\bar{\bar{x}}$	$s^2_{\bar{x}}$	$\bar{s}^2_i$	$\bar{r}_{ii'}$	s^2_t	s^2_s	s^2_e	T	S	R
1. Messung	deprimiert	1,24	0,072	0,225	0,708	0,07	0,16	0,07	0,29	0,71	0,50
	sorgenvoll	1,39	0,090	0,299	0,641	0,08	0,19	0,11	0,30	0,70	0,43
	überfordert	1,37	0,085	0,326	0,690	0,08	0,22	0,10	0,25	0,75	0,43
	ängstlich	1,18	0,047	0,150	0,608	0,04	0,09	0,06	0,32	0,68	0,42
	betrübt	1,33	0,089	0,291	0,704	0,08	0,20	0,09	0,28	0,72	0,48
	ungezwungen	2,92	0,130	0,576	0,615	0,11	0,35	0,22	0,24	0,76	0,34
	gewachsen	2,95	0,222	0,420	0,673	0,21	0,28	0,14	0,43	0,57	0,61
	vergnügt	2,20	0,219	0,581	0,807	0,20	0,47	0,11	0,30	0,70	0,64
	froh	2,29	0,212	0,559	0,782	0,20	0,44	0,12	0,31	0,69	0,62
	erfreut	2,21	0,189	0,600	0,789	0,17	0,47	0,13	0,27	0,73	0,58
2. Messung	deprimiert	1,26	0,093	0,242	0,708	0,09	0,17	0,07	0,33	0,67	0,55
	sorgenvoll	1,41	0,117	0,289	0,641	0,11	0,19	0,10	0,37	0,63	0,51
	überfordert	1,39	0,097	0,335	0,690	0,09	0,23	0,10	0,28	0,72	0,46
	ängstlich	1,21	0,074	0,168	0,608	0,07	0,10	0,07	0,40	0,60	0,51
	betrübt	1,37	0,120	0,309	0,704	0,11	0,22	0,09	0,34	0,66	0,55
	ungezwungen	2,82	0,174	0,555	0,615	0,16	0,34	0,21	0,32	0,68	0,43
	gewachsen	2,87	0,227	0,393	0,673	0,22	0,26	0,13	0,45	0,55	0,63
	vergnügt	2,13	0,230	0,629	0,807	0,21	0,51	0,12	0,30	0,70	0,64
	froh	2,29	0,240	0,580	0,782	0,22	0,45	0,13	0,33	0,67	0,64
	erfreut	2,19	0,202	0,584	0,789	0,19	0,46	0,12	0,29	0,71	0,60

$\bar{\bar{x}}$	mittlerer intraindividueller Mittelwert	s^2_t	Varianz der wahren Traitwerte
$s^2_{\bar{x}}$	Varianz der intraindividuellen Mittelwerte	s^2_s	Varianz der wahren Statewerte
$\bar{s}^2_i$	mittlere intraindividuelle Varianz	s^2_e	Varianz der Fehlerwerte
$\bar{r}_{ii'}$	mittlere intraindividuelle Korrelation (zwischen 1. und 2. Messung)	T	Traitcharakteristik
		S	Statecharakteristik (Änderungssensitivität)
		R	Reliabilität der Traitwerte

Betrachtet man zunächst die mittleren intraindividuellen Mittelwerte $\bar{\bar{x}}$, so fällt auf, daß die Vpn sich anläßlich der zweiten Messung im Schnitt als deprimierter, sorgenvoller, überforderter, ängstlicher und betrübter beschreiben. Auf der anderen Seite geben sie bei der zweiten Messung an, weniger ungezwungen, weniger vergnügt, weniger erfreut zu sein und sich weniger gewachsen zu fühlen. Legt man die Gesamtverteilungen mit ihren jeweils 2669 einzelnen Meßwerten je Item zugrunde und überprüft die Mittelwertdifferenzen zwischen erster und zweiter Messung auf Zufälligkeit (t-Test für korrelierte Stichproben), so zeigt sich, daß alle neun aufgeführten Differenzen von der ersten zur zweiten Messung (z.T. hoch) signifikant sind. Das gleiche Bild zeigt sich für diese Items, wenn statt der Einzelwerte die Verteilungen der intraindividuellen Mittelwerte aus der ersten und der zweiten Messung (mit einem Mittelwert je Item, Vp und Messung) einander gegenübergestellt werden (t-Test für korrelierte Stichproben) (Ausnahme: "sorgenvoll" wird hier nicht signifikant). Über die bloße statistische Signifikanz hinaus gewinnt dieser Befund erheblich an Gewicht, wenn man die Iteminhalte betrachtet: Die durchschnittliche Befindlichkeit der Vpn verändert sich demnach im Laufe der Administration eines einzelnen Protokollbogens in Richtung einer Zunahme der negativen Iteminhalte (deprimierter usw.) und in Richtung einer Verringerung der positiven Befindlichkeiten (weniger ungezwungen usw.) (vgl. Kap. 3.4.2.2.). Lediglich das Item "Ich bin froh" macht eine Ausnahme: Es verändert sich auch, allerdings nur marginal, in Richtung einer Verringerung der Merkmalsausprägung (dritte Nachkommastelle; nicht dokumentiert). Bei einer Irrtumswahrschein-

lichkeit von P=.74 kann diese Differenz aber als zufällig gelten.

Die interindividuellen Varianzen der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ nehmen von der Tendenz her mit Anwachsen der mittleren intraindividuellen Mittelwerte $\bar{\bar{x}}$ zu, sowohl bei der ersten als auch bei der zweiten Messung - ein zunächst unplausibel erscheinender Effekt, der jedoch mit den unterschiedlichen Gesamtverteilungen der Items zusammenhängen dürfte (vgl. Anhang B), denn ein Item, dessen schmaler Verteilungsgipfel am Rand der vierstufigen Ratingskala liegt (z.B. "Ich bin deprimiert") wird auch eher intraindividuelle Mittelwerte liefern, die nahe am Rand der Skala angesiedelt sind und die eine relativ geringe interindividuelle Varianz aufweisen. Demgegenüber wird ein Item mit einem breiteren Verteilungsgipfel in der Mitte der Skala (z.B. "Ich bin erfreut") auch eher intraindividuelle Mittelwerte mit einer größeren Varianz liefern. Die Varianzen der intraindividuellen Mittelwerte nehmen außerdem bei allen Items von der ersten zur zweiten Messung zu. Dies deutet an, daß die durchschnittlichen Merkmalsausprägungen, d.h. die wahren Traitwerte, bei der zweiten Messung breiter streuen, daß sich die Vpn bei der zweiten Messung also hinsichtlich ihrer habituellen Komponente stärker voneinander unterscheiden.

Die mittlere intraindividuelle Varianz $\bar{s}^2_i$ ist unterschiedlich hoch, was sowohl auf eine hohe Variabilität der wahren Statewerte hinweisen kann als auch auf eine hohe Fehlerkomponente. Dies läßt sich erst je Item näher bestimmen, wenn dessen mittlere intraindividuelle Korrelation $\bar{r}_{ii'}$ zur Analyse mit herangezogen wird. Je höher diese ist, desto mehr wird die mittlere intraindividuelle Varianz $\bar{s}^2_i$ durch wahre Statevarianz bestimmt. Wäre die mittlere Korrelation gleich eins, bestünden die gesamten intraindividuellen Varianzen aller Vpn ausschließlich aus wahrer Statevarianz. (In einem solchen Falle müßte man jedoch daran zweifeln, ob die Angaben unabhängig voneinander zustande gekommen sind.) Die hier dokumentierten mittleren Korrelationskoeffizienten sind sehr hoch, bedenkt man, daß es sich dabei um Korrelationen auf Itemebene handelt. Zudem sind es Mittelwerte, was bedeutet, daß etwa die Hälfte aller intraindividuellen Koeffizienten noch höher liegen.

Bei Betrachtung der errechneten Varianzen der Traitwerte s^2_t fällt auf, daß sie sich, ähnlich wie die Varianzen der intraindividuellen Mittelwerte, ausnahmslos von der ersten zur zweiten Messung vergrößern, während sich die Varianzen der Statewerte s^2_s und der Fehlerwerte s^2_e hierbei unspezifisch verhalten, z.T. verkleinern sie sich, z.T. werden sie größer oder bleiben auch gleich. Die Vermutung einer erhöhten Variabilität der Traitwerte anläßlich der zweiten Messung bestätigt sich also. Daß die zweite Messung mehr die habituelle Komponente des Erlebens zum Tragen bringt, zeigt sich auch ganz deutlich an den Trait-Charakteristiken T, die von der ersten zur zweiten Messung ebenfalls anwachsen, vom Item "Ich bin vergnügt" abgesehen, bei dem der T-Wert gleich bleibt. Wegen der Komplementarität verringern sich die State-Charakteristiken von einer Messung zur anderen um exakt den Betrag, um den die Trait-Charakteristiken zunehmen. Es zeigt sich bei allen Items, daß die State-Charakteristiken als Maßzahlen der Änderungssensitivität über .50 liegen, daß also bei allen Items mehr als 50% der Varianz der wahren Werte durch die Varianz der Statewerte determiniert ist. Dabei erreichen neun der zehn Items zumindest in der ersten Messung State-Charakteristiken von annähernd .70 oder mehr. Lediglich bei dem Item "Ich fühle mich gewachsen" sind hier nur gut die

Hälfte (57%) der gesamten wahren Varianz Statevarianz. Einer Zunahme der Varianz der wahren Traitwerte entspricht bei annähernd gleichbleibender Fehlervarianz auch eine Zunahme der statefreien (Trait-) Reliabilität R, was durch die Ergebnisse bestätigt wird.

Zusammenfassend bleibt festzuhalten: Die State-Charakteristiken belegen für neun der zehn relevanten Items zumindest bei der ersten Messung eine nicht unbefriedigende Ausprägung der Änderungssensitivität als Itemmerkmal. Diese Änderungssensitivität verringert sich jedoch von der ersten zur zweiten Messung innerhalb eines Meßzeitpunkts; im Gegenzug vergrößert sich die Trait-Charakteristik. Verantwortlich dafür ist eine Zunahme der interindividuellen Varianz der intraindividuellen Mittelwerte und einhergehend damit eine Vergrößerung der Varianz der wahren Traitwerte. Die zweite Messung spricht also mehr auf das habituelle Moment des Stimmungserlebens an, d.h. auf Unterschiede zwischen den Vpn, und nur noch in verringertem Maße auf situativ bedingte Unterschiede innerhalb der Vpn.

Die Ursache für diesen Befund ist einleuchtend und liegt auf der Hand: Jede Messung mißt das im jeweiligen Moment gerade aktuelle Erleben. Die erste Messung wird dabei stärker durch das beeinflußt, was die Vpn unmittelbar vor dem Bearbeiten der Protokollbögen getan und erlebt haben und was ihrem hier interessierenden, unbeeinflußten Alltagsverhalten und -erleben entspricht. Sie unterscheiden sich in dieser Meßsituation, die dem Alltagserleben alleine schon zeitlich näher liegt als die der zweiten Messung, in geringerem Maße voneinander als in der zweiten Meßsituation. Bei dieser hingegen sind Unterschiede zwischen den Vpn stärker ausgeprägt. Die zweite Messung wird nämlich genauso wie die erste von dem beeinflußt, was die Vpn gerade tun und erleben, doch in der zweiten Meßsituation machen alle das gleiche: sie bearbeiten einen Protokollbogen (und das heißt präzise formuliert: sie haben alle unmittelbar vorher die STAI-S18 Items bearbeitet) und befinden sich daher alle in der gleichen, quasi-standardisierten Situation. Das bedeutet, die erste Messung mißt das Alltagsverhalten und -erleben, die zweite dagegen die Erhebungssituation - ein Effekt, der, obschon schwach, unerwünscht ist, und einer Grundannahme dieses Untersuchungsansatzes widerspricht. Annahme war nämlich (vgl. Kap. 2.1.3. und Kap. 2.2.3.), daß sich die Befindlichkeit während des Messens nicht verändert und zwar am allerwenigsten auf systematische Art und Weise durch den Meßvorgang selbst.

3.4.2.2. Reaktivitätseffekte auf Itemebene

Durch das Vorliegen eines solchen möglichen Reaktivitätseffekts und die mit ihm einhergehenden Zweifel an dieser Grundannahme könnte die Basis zur Quantifizierung der wahren Statevarianz und der Fehlervarianz, nämlich die mittlere intraindividuelle Korrelation zwischen erster und zweiter Messung, in Frage gestellt werden. Von einem solchen Effekt darf nämlich nicht angenommen werden, daß er sich bei allen Vpn im Sinne einer linearen Veränderung aller Meßwerte auf dieselbe Art und Weise auswirkt und sich so hinsichtlich der Korrelationsberechnung wieder neutralisiert. Treten Effekte zwischen beiden Messungen auf, dann ist die angeblich unsystematische Fehlervarianz von systematischer Varianz in den wahren Werten überlagert.

Generell muß überall dort mit Reaktivitätseffekten gerechnet werden, wo Meßdaten von informierten Vpn erhoben werden, die um ihre Teilnahme an einer Datenerhebung wissen (CAMPBELL, 1957, zit. bei GACHOWETZ, 1987, S. 259). Überträgt man die verhaltensorientierte Begriffsbestimmung von STERN (1986, S. 14) auf das Erleben, so kann Reaktivität als die Beeinflussung des zukünftigen Auftretens von Stimmungen und ihren Modalitäten durch mit der Stimmungswahrnehmung und -protokollierung zusammenhängende Stimuli (z.B. Signalton und Items der Protokollbögen) sowie durch Vorgänge im Bereich der Informationsverarbeitung verstanden werden. Bei dem hier zur Diskussion stehenden möglichen Reaktivitätseffekt handelt es sich um eine sehr kurzfristige, im Minutenbereich liegende Beeinflussung von Stimmungen oder Stimmungsausprägungen. Sie findet während oder gleich im Anschluß an die Rezeption der Protokollbogenitems als Stimuli (bis zum letzten STAI-S18 Item) und infolge parallel dazu ablaufender hypothetischer Informationsverarbeitungsprozesse statt. Ihr Effekt wird als Meßwertdifferenz zwischen erster Messung (als Bezugsgröße) und zweiter Messung dargestellt. Es handelt sich hier also nicht um langfristige, über den gesamten Beobachtungszeitraum ablaufende systematische Veränderungen im Stimmungserleben und im Protokollierverhalten (auf die wird weiter unten eingegangen), sondern um einen häufig wiederkehrenden, sich kurzfristig innerhalb von wenigen Minuten ereignenden Effekt.

Zur weiteren Überprüfung eines solchen Effekts wurde neben den oben durchgeführten Überprüfungen der Mittelwertdifferenzen zwischen erster und zweiter Messung (Kap. 3.4.2.1.) noch eine andere, differenziertere Strategie angewandt: Für jede Vp und jedes Item wurde ein Vergleich der intraindividuellen Mittelwerte aus erster und zweiter Messung durchgeführt (t-Test für korrelierte Stichproben). Für diese 740 vorgenommenen Signifikanzprüfungen gilt, wie bereits dargelegt (vgl. Kap. 3.4.1.), daß unter Zugrundelegung eines Signifikanzniveaus von 5% genau dieser Prozentsatz (entsprechend einer Gesamtzahl von 37 einzelnen t-Tests bzw. 3.7 t-Tests je Item) zufällig signifikant werden kann. Wie die Datenanalyse zeigt, werden jedoch fast zweieinhalb mal so viele dieser t-Tests, nämlich 91, signifikant (entsprechend 12.3% von 740). Die Tabelle 2 schlüsselt die signifikant gewordenen t-Tests nach Items auf und dokumentiert außerdem die Richtungen der Mittelwertänderungen.

Tab. 2 Häufigkeit und Richtung signifikanter t-Tests auf Itemebene zwischen 1. und 2. Messung (Basis: 74 t-Tests je Item) (P≤.05)

	Gesamt	2. Mittelwert ist größer	kleiner
betrübt	8	7	1
deprimiert	6	5	1
sorgenvoll	6	4	2
überfordert	6	4	2
ängstlich	4	4	-
vergnügt	17	3	14
gewachsen	15	1	14
ungezwungen	15	-	15
froh	8	4	4
erfreut	6	2	4
SUMME	91		

Dabei zeigen sich wiederum die bereits beobachteten Veränderungen in semantisch bedeutsamer Richtung (vgl. Kap 3.4.2.1.). Bei vielen Vpn steigen die mittleren Ausprägungen der negativen Befindlichkeiten eher an, während sie bei den Items "vergnügt", "gewachsen" und "ungezwungen" eher absinken (im Extremfall des Items "Ich fühle mich ungezwungen" sogar bei insgesamt 15 Vpn, was 20% aller Vpn entspricht). Bei dem Item "Ich bin erfreut" ist diese Veränderung nicht sehr ausgeprägt, wenngleich die Richtung deutlich wird. Nur das Item "Ich bin froh" verhält sich insgesamt neutral, was die Befunde aus Kap. 3.4.2.1. stützt. In insgesamt 30 Fällen verändern sich die negativen Befindlichkeiten, in 61 Fällen jedoch die positiven Befindlichkeiten. In symptomatischer Hinsicht verändern sich die Items der negativen Befindlichkeiten in 24 Fällen (Zunahme des Mittelwerts), die der positiven in 51 Fällen (Verringerung des Mittelwerts).

Es bleibt angesichts dieser Befundlage festzustellen, daß es offensichtlich eine systematische Veränderung von der ersten zur zweiten Messung in den Ausprägungen der mittleren Befindlichkeiten auf Itemebene gibt, die sich inhaltlich in völlig plausibler Form mitteilt. Demnach geben die Vpn bei der zweiten Messung zwar eher an, in gedrückterer Stimmung als bei der ersten Messung zu sein, doch in weit größerem Umfang beschreiben sie sich bei der zweiten Messung als weniger euphorisch. Der Reaktivitätseffekt läßt sich daher nicht so sehr als ein vermehrtes "Genervtsein" der Vpn in Folge des Bearbeitens der Protokollbögen beschreiben, er äußert sich vielmehr in einer Dämpfung der guten Laune.

Ein weiterer Punkt, der bei der Diskussion möglicher Reaktivitätseffekte beachtet werden muß, ist die bereits erwähnte langfristige Veränderung der Meßwerte über den gesamten Beobachtungszeitraum. Dieser Effekt hat zwar keine Auswirkungen auf die empirische Gültigkeit des Modells von BUSE & PAWLIK (1991), sein mögliches Auftreten soll hier dennoch kurz überprüft werden. Konkret geht es dabei um ein "pattern of response decay" (STONE, KESSLER & HAYTHORNTHWAITE, 1991, S. 600) im Zuge einer ständigen Wiederholung desselben Stimulusmaterials, das mit nichts anderem als der Länge der Zeit, die die Vpn an der Untersuchung bereits teilgenommen haben, korreliert zu sein scheint und das sich in einer systematischen Veränderung der Auftretenshäufigkeiten oder Ausprägungen des protokollierten Verhaltens bzw. Erlebens über die Zeit äußert. So ein Antwortverfallmuster deutet auf Lern- oder Ermüdungseffekte hin (CATTELL, 1967, STONE, KESSLER & HAYTHORNTHWAITE, 1991) oder kann auch in Folge verstärkter Selbstbeobachtung und -auseinandersetzung bzw. verstärktem Bewußtwerden von bisher unbeachtet gebliebenen Verhaltensweisen bzw. Erlebnisinhalten auftreten. "But to date, the potential influence of daily experience studies on daily experience has not been adequately addressed and remains a methodological challenge to this area of inquiry." (TENNEN, SULS & AFFLECK, 1991, S. 321) Eine methodisch grundlegend neue und angemessene Herangehensweise an diese Problematik wurde allerdings auch hier nicht entwickelt, statt dessen wurde das Naheliegende unternommen: Die durchschnittlichen Merkmalsausprägungen der ersten Hälfte der Beobachtungszeiträume wurden mit denen aus der zweiten Hälfte verglichen. Dazu wurden für jede Vp und für jedes Item (getrennt für beide Messungen) zwei intraindividuelle Mittelwerte gebildet, einer aus den Protokollterminen 1 bis 21, der andere aus den Protokollterminen 22 bis 42. Je Item und Messung wurden alle 74 intraindividuellen Mittelwerte auf der Basis der ersten 21 Protokolltermine gegen diejenigen aus den letzten 21 Protokollterminen getestet (t-Tests für korrelierte Stichproben). Als Ergebnis zeigen sich

keinerlei signifikante Mittelwertunterschiede; dabei sind die empirischen Mittelwertunterschiede oft so gering, daß sie z.T. mit großer Wahrscheinlichkeit zufällig zustande gekommen sind.

Ein Reaktivitätseffekt über den Beobachtungszeitraum hinweg kann somit auf der Ebene von Mittelwertvergleichen nicht nachgewiesen werden.

3.4.3. Ergebnisse zur Untersuchungsmethode

In diesem Abschnitt werden die Angaben der Vpn aus den Nachbefragungsbögen (vgl. Anhang C) sowie alle Angaben aus den Protokollbögen ausgewertet, die nicht unmittelbar der Befindlichkeitsmessung dienen (vgl. Anhang B). Die Datenanalyse erfolgt dabei unter den Aspekten der Sicherung der (psychometrischen) Qualität der erhobenen Felddaten, der Abschätzung der ökopsychologischen Gütekriterien sowie der Akzeptanz der Methode seitens der Vpn.

Die Nachbefragungen fanden nach Abschluß der Beobachtungen statt, d.h. frühestens am Abend des letzten Tages der Felddatenerhebung. Fast zwei Drittel aller Nachbefragungen (64%) wurden an diesem Tag oder dem darauffolgenden durchgeführt, weitere 18% fanden zwischen dem zweiten und vierten Tag nach Ende der Beobachtungen statt und noch einmal der gleiche Prozentsatz zu einem noch späteren Zeitpunkt (vgl. Anhang B).

3.4.3.1. Datenqualität

Die Frage nach der Datenqualität berührt das Problem der Durchführungsobjektivität, das in Kapitel 3.6.1. abgehandelt wird. Die hier untersuchten Sachverhalte können nur ex post eine Abschätzung der Compliance der Vpn und anderer Faktoren während der Beobachtungszeiträume aufgrund der Selbstauskünfte der Vpn leisten, die nicht mit einer tatsächlichen Kontrolle der Erhebungsbedingungen durch den Untersuchungsleiter (wie z.B. in Laboruntersuchungen) verwechselt werden darf. Als qualitätsmindernd müssen dabei alle potentiellen Störeffekte angesehen werden, d.h. alle die Einflüsse, die möglicherweise während des Beobachtungszeitraums unkontrolliert wirksam geworden sind und dadurch zu solchen Veränderungen in der Befindlichkeit der Vpn oder in ihrem Protokollierverhalten geführt haben, wie sie ohne diese Einflüsse nicht aufgetreten wären. In diesem Zusammenhang muß zunächst auf den eben diskutierten Reaktivitätseffekt hingewiesen werden (vgl. Kap. 3.4.2.2.; aber auch Kap. 3.5.2.3.), der zur Minderung der Datenqualität beiträgt, indem er als Konsequenz aus dem messungsbedingten Eingriff in das Alltagserleben der Vpn zu einer systematischen Veränderung der zu messenden Merkmale in der oben beschriebenen Richtung während der Protokollierung führt. Speziell sollen im folgenden die von den Vpn selbst wahrgenommenen Reaktionen auf die Felduntersuchung, mögliche Verstöße gegen die Instruktionen und die Gewährleistung der Undurchschaubarkeit der Protokollbögen und der gesamten Untersuchung überprüft werden.

Zur Sicherstellung der Undurchschaubarkeit wurden eine Reihe von Maßnahmen getroffen (vgl. Kap. 3.2.2.2.). Zur Feststellung ihrer Wirksamkeit wurden alle Vpn im Anschluß an die Bearbeitung der Nachbefragungsbögen mündlich gefragt, ob ihnen an der Struktur der Protokollbögen etwas aufgefallen sei, ob sie bemerkt hätten, daß sich vielleicht einige Items wiederholt hätten u.ä. Keine einzige Vp gab darauf zur Antwort, daß die ersten zehn Items dieselben waren wie die letzten zehn. Es ist auch in keinem Falle aufgefallen, daß nur die ersten und die letzten zehn Items ständig rotiert wurden und die STAI-S18 Items immer an derselben Stelle standen. Die Positionen einzelner STAI-S18 Items wurden allerdings gelegentlich erinnert, insbesondere "nervös", "zappelig" und "verkrampft", die aufeinander folgen und, so die Vpn, sich eher auf körperliche Zustände beziehen würden. Von entscheidender Bedeutung in diesem Zusammenhang ist auch die Beantwortung der Frage 13 des Nachbefragungsbogens (vgl. Anhang C). In ihr werden die Vpn aufgefordert, Angaben zu dem vermuteten Zweck oder Ziel der Untersuchung zu machen: 82% gaben an, sich Gedanken dazu gemacht zu haben. Ein Blick in die entsprechende Tabelle im Anhang C zeigt, daß in den weitaus meisten Fällen das Thema der Untersuchung im Bereich der Situationsabhängigkeit des Erlebens, in der Erforschung von zyklischen Merkmalsveränderungen, im Bereich der Testentwicklung oder der Reliabilitätsermittlung im weitesten Sinne angesiedelt worden ist, und nur in vier Fällen wurden Antworten in der Nähe des tatsächlichen Themas gegeben ("interessant ist, wie man bei den wiederholten Fragen geantwortet hat; mehrfache Erfassung von Items zur Befindlichkeit; Antwortkonstanz; Einfluß der vorhergehenden Fragen auf die nächsten Antworten"). (Es ist bei allen offenen Fragen wie dieser hier mit Mehrfachnennungen zu rechnen.) Die Undurchschaubarkeit des Instruments und die Verschleierung des Untersuchungsziels sind also offenbar erfolgreich gewährleistet worden, keine Vp hat das Ziel oder Thema annähernd richtig beschrieben und in keinem Fall ist der Aufbau der Protokollbögen erkannt worden. (Dazu kann noch angemerkt werden, daß die Vpn auch nach Offenlegung der Untersuchungsabsicht zumeist weiterhin einen uninformierten Eindruck machten und daß selbst der Versuch, das Thema fortgeschrittenen Kommilitonen oder sogar fertigen Diplom-Psychologen zu erläutern, nur in seltenen Fällen von Erfolg gekrönt war.)

Der vorgezogenen Rückgabepflicht der Protokollbögen während des Beobachtungszeitraums für die Psychologiestudenten wurde in allen Fällen nachgekommen und ein unzulässiges Abschreiben in großem Stil oder Orientieren an bereits ausgefüllten Protokollbögen wurde dadurch sehr erschwert oder unmöglich gemacht. Weitere Hinweise auf die Datenqualität ergeben sich aus den Fragen 9, 10 und 12 des Nachbefragungsbogens. Vor der Bearbeitung der Fragen 9 bis 12 wurde im Fragebogentext an die Vpn appelliert, besonders hier wahrheitsgemäß zu antworten. Frage 9 erkundigt sich danach, ob die Vpn beim Bearbeiten der Protokollbögen öfter versucht haben, sich an im selben oder in früher ausgefüllten Protokollbögen gemachte Angaben zu erinnern. Immerhin knapp ein Viertel der Vpn (24%) bejahte dies. Von denen gab die Hälfte an, sich an vorherige Antworten erinnert zu haben, ein Drittel hat zurückgeblättert und ein Drittel gab an, sich nur innerhalb desselben Protokollbogens orientiert zu haben. Manche haben dies aber erst nach Bearbeiten des Protokollbogens getan, andere haben versucht, sich eben nicht zu erinnern und ganz neu zu antworten. Drei Vpn gaben an, dies nur selten oder ein- bis fünfmal getan zu haben, doch immerhin zwei Vpn sagten, daß sie es etwa 7 bis 15mal getan haben. Frage 10 versucht herauszufinden, ob beim Bearbeiten der Protokollbögen falsche Angaben gemacht wurden oder von

bereits ausgefüllten Bögen abgeschrieben wurde. Nur 4 Vpn (5%) bejahten diese Frage und alle sagten, daß sie falsche Uhrzeiten angegeben haben. In einem Falle wurde die Tätigkeit unmittelbar vor dem Bearbeiten nicht korrekt angegeben. Und nur 2 Vpn (3%) antworteten auf Frage 12, daß sie beim Ausfüllen der Protokollbögen häufiger, und d.h. bei beiden Vpn ca. zwei- bis dreimal, unter deutlicher Einwirkung von Alkohol, Medikamenten oder anderen Drogen gestanden haben. Weitere Hinweise auf mangelnde Compliance oder andere im Feld aufgetretene Störeffekte können auch aus einigen Items der Frage 6 des Nachbefragungsbogens gewonnen werden. Den dort formulierten Aussagesätzen zur Untersuchung sollten die Vpn auf einer fünfstufigen Verbalskala mehr oder weniger zustimmen (1 = stimme ganz und gar zu / 5 = stimme gar nicht zu, Mittelkategorie: 3 = bin nicht sicher). Die meisten dieser Items zeigen dabei deutlich bimodale Verteilungen der Antworten auf die fünf Antwortkategorien, was an schlecht oder mehrdeutig formulierten Items oder an unklaren Antwortkategorien liegen kann oder was auch ein Zeichen für ein heterogenes Meinungsbild innerhalb der Stichprobe sein kann. Jedenfalls ist die Mittelkategorie bei den meisten Items nur sehr schwach besetzt, was genau so gut auch als Zeichen der Entscheidungsstärke der Vpn interpretiert werden kann: Die meisten hatten zu fast allen Items eine so klare Meinung, daß sie die neutrale Kategorie nicht bemühen mußten. Dem Item Nr. 1 ("Das Befolgen der Untersuchungsanweisungen war anstrengend.") stimmten immerhin 42% aller Vpn ein bißchen oder ganz und gar zu, 38% der Vpn taten dies beim Item Nr. 3 und fanden es demzufolge anstrengend, ständig das Ringbuch mit den Protokollbögen herumzutragen, während 47% die Bearbeitung der Protokollbögen als Störung in ihrem Tagesablauf empfunden haben (Item Nr. 9). Dies macht deutlich, daß ein großer Teil der Vpn die Teilnahme an der Untersuchung im täglichen Vollzug vor Ort zumindest partiell als mehr oder minder belästigend angesehen hat, was für die Qualität der Daten nicht förderlich ist. Dies gilt aber offenbar nicht für die Länge eines einzelnen Protokollbogens: Daß die Bearbeitung eines einzelnen Protokollbogens zu lange dauert, fanden nur 11% der Vpn, die anderen 89% konnten dem entsprechenden Item (Nr. 8) nur weniger oder gar nicht zustimmen. Auch während des Debriefings in der Nachbefragung bestätigten viele Vpn mündlich, daß der Zeitaufwand je Protokolltermin durchaus nicht zu groß gewesen sei. Auf Nachfrage erklärten die Psychologiestudenten unter ihnen z.T. sogar explizit, daß die vergebenen 20 Vp-Stunden eine Vergrößerung des Itemumfangs um ca. 20 Items durchaus zulassen würden. Die Eindeutigkeit der Befindlichkeitsitems in den Protokollbögen bemängelten 32% der Vpn (Item Nr. 2), 50% hatten daran eher nichts auszusetzen, doch 16% waren sich hinsichtlich der Eindeutigkeit nicht sicher. Mündlich wurde zusätzlich von einigen Vpn berichtet, daß sich die Bedeutungen mancher Items im Laufe des Beobachtungszeitraums für sie verändert haben oder daß die Antwortkategorien der Ratingskala unklar waren. Daß sich die Befindlichkeitsitems z.T. sehr ähnelten oder sich wiederholten, störte zudem die Hälfte der Vpn (Item Nr. 5).

Die Vpn sollten sich laut Instruktionen (vgl. Anhang A) möglichst wenig mit anderen über die Untersuchung unterhalten, um nicht durch zuviel Aufmerksamkeit und gedankliche Reflexion Prozesse in Gang zu setzen (z.B. verstärkte Selbstbeobachtung im Hinblick auf die untersuchten Merkmale, Nachdenken über den Untersuchungszweck u.ä.), die zu systematischen Veränderungen in den Merkmalen oder im Protokollierverhalten hätten führen können. Ein knappes Viertel der Vpn (24%) stimmte dem Item Nr. 21 ein bißchen oder ganz und gar zu und bestätigte damit, sich während des Beobachtungszeitraums öfter oder intensiv mit anderen über die Untersuchung unterhalten zu

haben. 59% der Vpn haben sich im Hinblick auf die Gefühle, die im Fragebogen angesprochen wurden, im Laufe der Untersuchung immer aufmerksamer beobachtet (Item Nr. 18), was aber, wie bereits berichtet (vgl. Kap. 3.4.2.2.; aber auch Kap. 3.5.2.3.), nicht zu einer Veränderung in den mittleren Merkmalsausprägungen über die Beobachtungszeiträume geführt hat.

Zusammmenfassend kann festgestellt werden, daß auf der Basis der hier ausgewerteten Informationen die Datenqualität als hinreichend gut eingeschätzt werden kann, sieht man einmal von dem bereits analysierten Reaktivitätseffekt ab. Es ist offensichtlich gelungen, die Vpn während der Datenerhebung über die Struktur der Protokollbögen und über den Untersuchungszweck im unklaren zu lassen, wobei letzteres vermutlich nicht ausschließlich eine Folge gezielter Maßnahmen war, sondern vor allem durch den hohen Abstraktionsgrad des Themas sichergestellt wurde. Zwar hat fast ein Viertel aller Vpn angegeben, sich beim Ausfüllen der Protokollbögen an bereits protokollierte Angaben erinnert zu haben, doch hat fast niemand gefälschte Angaben gemacht und auch der Einfluß von Drogen oder Medikamenten auf das Antwortverhalten scheint unbedeutend gewesen zu sein. Hinsichtlich der 24% der Vpn, die sich beim Ausfüllen an frühere Angaben erinnert haben, muß gefragt werden, wie gering der Anteil solcher Vpn gehalten werden kann, wenn man bedenkt, daß die im Alltag zum großen Teil automatisch arbeitenden Gedächtnisfunktionen sich nicht einfach per Anweisung abstellen lassen, zumal ihr Funktionieren ansonsten von existentieller Bedeutung für die Bewältigung des Alltags ist. Zwischen einem guten Drittel und knapp der Hälfte der Vpn fanden Teile der Untersuchung anstrengend oder störend. Dies ist prinzipiell bedenklich, denn es ist zu vermuten, daß Protokollierungen, die in einem unlustvollen, vielleicht von Pflichterfüllung bestimmten Zusammenhang erfolgen, nicht nur zu einer Veränderung der Befindlichkeitsmerkmale, sondern auch zu generell unreliableren Angaben führen können. Ein knappes Drittel der Vpn hatte an der Eindeutigkeit der Items oder auch an der Ratingskala etwas auszusetzen, die Hälfte der Vpn fühlte sich durch die Ähnlichkeit einiger Items oder durch Itemwiederholungen gestört. Auch das ist nicht positiv zu bewerten, doch gilt hierbei ebenso wie bei der Beurteilung des zuvor dargestellten Sachverhalts, daß zum einen nicht klar ist, ob und in welchem Umfang diese Faktoren wirklich einen Einfluß auf das Protokollierverhalten ausgeübt haben und darüber hinaus fehlen Vergleichsangaben aus anderen ESM-Untersuchungen. Ein knappes Viertel der Vpn hat sich während der Felddatenerhebung öfter oder intensiv mit anderen über die Untersuchung ausgetauscht und bei knapp zwei Drittel hat eine verstärkte Selbstbeobachtung hinsichtlich der erhobenen Merkmale stattgefunden. Auch dies kann zu einer Veränderung im Protokollierverhalten geführt haben, doch auch hier gilt, daß beides Verhaltensweisen sind, die bei ESM-Studien nur schwer unterbunden werden können. Trotz der aufgezeigten Indizien für mögliche Störeinflüsse und der Unklarheit über deren tatsächliches Wirksamwerden können die erhobenen Daten als interpretierbar angesehen werden.

3.4.3.2. Ökopsychologische Gütekriterien

Die ökopsychologischen Gütekriterien sind die ökologische Validität und die ökologische Repräsentativität (vgl. Kap. 2.2.3.). Eine empirische Untersuchung ist ökologisch valide, wenn in ihr nur solche Situationen vorkommen, die dem Alltag der untersuchten Vpn entstammen, d.h. wenn alle und ausschließlich solche Stimuli wirksam werden, die auch sonst die verhaltenswirksamen Biotope

der Vpn bilden. Es dürfen in einer ökologisch validen Untersuchung keine derartigen Stimuli grundsätzlich fehlen und auch keine neu hinzukommen, die normalerweise nicht wirksam werden (Krankheit, Weihnachtsfest usw.). Wenn die untersuchten Situationen, z.B. in Feldstudien, diese Stimulusvariablen in für die Alltagsumwelt repräsentativen Kombinationen und Häufigkeiten zur Wirkung bringen, ist die Untersuchung ökologisch repräsentativ. Das Kriterium der ökologischen Validität kann also bereits auf eine einzelne Erhebungssituation angewandt werden aber im Falle von ESM-Untersuchungen auch auf den gesamten Beobachtungszeitraum. Das Kriterium der ökologischen Repräsentativität hingegen bezieht sich stets auf eine Stichprobe von mehreren ökologisch validen Einzelsituationen. Alle im Alltag gezogenen Situationsstichproben, wie in ESM-Studien, sind notwendigerweise ökologisch valide - vorausgesetzt, es handelt sich wirklich um Alltagssituationen, was nicht als selbstverständlich gegeben angesehen werden darf. Eine Vp, die zusagt, Protokollierungen im Alltag vorzunehmen, und tatsächlich in der Mitte des Beobachtungszeitraums von einem kritischen Lebensereignis überrascht wird und die Protokollierung fortsetzt, liefert für die zweite Hälfte keine ökologisch validen Daten mehr. Die ökologische Repräsentativität kann, wie im vorliegenden Falle, durch einen randomisierten Zeitstichprobenplan (d.h. durch zufällig ausgewählte, für die Vpn nicht vorhersehbare Stichproben aus ihren alltäglichen Verhaltensströmen) und durch eine Begrenzung der Latenzzeit, hier z.B. auf dreißig Minuten, gewährleistet werden. Je höher aber die tatsächlichen Latenzzeiten im Feld sind und je weniger Protokolltermine wahrgenommen werden, je mehr die Vpn also in die Realisierung des Zeitstichprobenplans eingreifen, desto geringer wird die ökologische Repräsentativität ausfallen.

Die Überprüfung der ökologischen Validität erfordert also eine Abschätzung der Alltäglichkeit der Untersuchungsbedingungen, d.h. der Beobachtungszeiträume. Der Nachbefragungsbogen enthält dazu die Frage 1, in der die Vpn danach gefragt werden, ob der Untersuchungszeitraum für sie repräsentativ war, d.h. ob sie eine für sie typische Woche verbracht haben. Das Wort "repräsentativ" wird hier im Fragenkontext in der Bedeutung "repräsentativ für alle Wochen, in denen Alltagsleben stattfindet" verwendet und meint damit "ökologisch valide" und nicht "ökologisch repräsentativ"! Gut zwei Drittel der Vpn bejahten diese Frage, immerhin 32% verneinten sie jedoch. Betrachtet man die freien Antworten für die Begründung der Verneinung der Frage 1, so läßt sich jedoch kein Schwerpunkt erkennen. Das Antwortspektrum ist breit gefächert und umfaßt quantitativ oder qualitativ ungewöhnliche kleinere Ereignisse, überdurchschnittlich starke berufliche Beanspruchung, Übergang von Semesterferien zum Veranstaltungsbeginn, häufiges Alleinsein, zuviel Freizeit und fehlenden Streß ebenso wie zuviel Streß oder familiäre Spannungen. Daneben werden auch Trennungssituation vom Partner, neuer Job, Kurzurlaub, Umzug und Krankheit genannt. Die zuletzt aufgeführten Faktoren können durchaus als kritische Lebensereignisse angesprochen werden, die einen hohen Grad an sozialer Wiederanpassung erfordern können (HOLMES & RAHE, 1967, zit. bei BASTINE, 1990, S. 165) und in dem Beobachtungszeitraum einer ESM-Untersuchung nicht vorkommen sollten. Die davor aufgelisteten Abweichungen von typischen Wochenabläufen können dagegen eher als unbedenklich eingestuft werden, wenngleich es in der subjektiven Einschätzung von Ereignissen und ihrer Bedeutung für das Verhalten und Erleben der Vpn selbstverständlich eine erhebliche interindividuelle Variationsbreite gibt. Insgesamt scheint die überwältigende Mehrzahl von Vpn in alltäglichen oder alltagsähnlichen Zeiträumen untersucht worden zu sein; die Untersuchung kann daher als ökologisch valide gelten. Bedenkt man zusätzlich, daß der Alltag mancher Vpn Zyklen

von Verhaltens- oder Erlebensweisen beinhalten kann, deren Ablauf den einer Woche übersteigt (etwa mehrwöchentlich angelegte Wechselschichtdienste oder vierzehntägig stattfindende Sportveranstaltungen, deren Auswirkungen ein ganzes Wochenende beeinflussen können, usw.), so zeigt sich, daß es auch für die Vpn selbst nicht immer einfach sein muß, den Erhebungszeitraum als typisch oder untypisch einzustufen. Weitere Hinweise zur ökologischen Validität können sich aus den Protokollbogenitems ergeben, die vor den eigentlichen Stimmungsitems bearbeitet wurden. So zeigt die Tätigkeit der Vpn unmittelbar vor dem Protokollieren eine breite Verteilung auf die vorgegebenen Kategorien. "Freizeit" wurde am häufigsten (41% der Protokolle), "Schlafen / Ruhen" am seltensten (7% der Protokolle) angegeben, was für eine (größtenteils) studentische Stichprobe nicht unplausibel ist und davon zeugt, daß der größte Teil der Protokolltermine während der Wachzeit der Vpn plaziert wurde. "Anderes" wurde aber auffallend häufig (20%) markiert, was ein Indiz für schlecht gewählte oder formulierte Antwortkategorien sein mag. 84% der Protokolle wurde in gewohnter Umgebung angefertigt, das Verhältnis von zu Hause und nicht zu Hause ausgefüllten Bögen ist ausgewogen (46% zu 53%) und in 65% der Protokolltermine waren die Vpn nicht allein. Unter der Bedingung "nicht allein" wurde die Qualität des Kontakts in nur 5% der Fälle als unangenehm bezeichnet, was etwas zu niedrig erscheint. Dennoch sprechen auch diese Befunde für eine hohe Alltagsähnlichkeit der Beobachtungszeiträume, besonders der Umstand, daß in nur 16% der Termine eine Protokollierung in ungewohnter Umgebung stattgefunden hat.

Die ökologische Repräsentativität wird im wesentlichen durch eine Analyse der Latenzzeiten sowie der Zahl der wahrgenommenen Protokolltermine eingeschätzt. Von den 3108 möglichen Protokollterminen (74 Vpn mit jeweils 42 möglichen Terminen) gingen 2669 (85.9%) als gültige Protokolle in die Auswertung ein, was einer mittleren Reaktionsrate von 36.1 wahrgenommenen Protokollterminen je Vp entspricht. Dabei nahm jede Vp zwischen min. 23 (54.8%) und max. 42 (100%) der angebotenen 42 Protokolltermine wahr. In keinem Falle ist es zu einem vorzeitigen Abbruch der Protokollierung gekommen, lediglich bei der Vp mit den 23 wahrgenommenen Terminen gab es am letzten Untersuchungstag keine Protokollierungen mehr (und zwar als Folge eines zwischenzeitlich eingetretenen kritischen Lebensereignisses). Ein Blick auf die Verteilung der Latenzzeiten zeigt, daß mehr als die Hälfte (51%) aller Protokolle unmittelbar nach dem Ertönen des Signals (vor Ablauf von einer Minute) bearbeitet wurde und weitere 14% bis zum Ablauf von 2 Minuten danach. Insgesamt wurden in den ersten 5 Minuten nach dem Wecksignal 80% aller Protokollbögen ausgefüllt, in den ersten 10 Minuten insgesamt 86%, in den ersten 20 Minuten insgesamt 93%, weitere 4% folgten bis zum Ablauf der Latenzzeit. (Für das von jeder Vp während der Testeinweisung und Vorbefragung bearbeitete Protokoll gibt es keine Zeitabweichung, weil es dafür keinen expliziten Termin gab.) Obwohl sich aus diesen Befunden schon eindeutige Belege für das Vorliegen einer ausreichenden ökologischen Repräsentativität infolge einer hohen Reaktionsrate und geringen Latenzzeiten ergeben (die Vpn nahmen sehr viele Protokolltermine wahr, die meisten davon unmittelbar nach Ertönen des Signaltons, und hatten somit nur geringen Einfluß auf die Auswahl der konkreten Protokolliersituationen), sollen zusätzlich noch ein paar Angaben aus den Nachbefragungsbögen ausgewertet werden.

In Item Nr. 19 (aus Frage 6) stimmten 79% der Vpn der Aussage weniger oder gar nicht zu, bestimmte Situationen - mit oder ohne Absicht - in Erwartung eines Protokolltermins vermieden oder

verzögert aufgesucht zu haben. 92% der Vpn gaben weiter an, daß sie bestimmte Situationen - mit oder ohne Absicht - nicht unter anderem deshalb aufgesucht haben, weil sie einen Protokolltermin für wahrscheinlich hielten (Item Nr. 20), und nur 17% der Vpn sagten in Item Nr. 15, daß sie manchmal die Mitnahme des Signalgebers oder des Ringbuchs mit den Protokollbögen vergessen haben. Nur 9% aller Vpn gaben in Item Nr. 12 an, daß es sie gereizt hat, an den Funktionsknöpfen der Uhr herumzuspielen, während alle anderen 91% dieser Aussage gar nicht zustimmten. Das paßt zu den Angaben in Frage 11, wonach nur 2 Vpn (3%) aussagten, daß es ihnen einmal oder häufiger gelungen ist, die nächsten Protokolltermine durch Manipulieren der Uhr herauszufinden, was sich wiederum mit der informellen Kontrolle der Signalgeber durch den Untersuchungsleiter anläßlich der Nachbefragung deckt: In einem Falle war die Uhrzeit verstellt (was beim Auswerten der Protokollbögen berücksichtigt werden konnte) und in einem anderen Falle war der Fehler (Anzeigen der nächsten drei Protokolltermine in den Analogdisplays) schon während der Datenerhebung nach telefonischer Rücksprache behoben worden. Ansonsten wurden alle Signalgeber in einwandfreiem Modus zurückgegeben. Soweit sind diese Befunde mit den oben gemachten Aussagen zur ökologischen Repräsentativität konform. Andererseits stimmten 87% der Vpn dem Item Nr. 10 aus Frage 6 zu und bekundeten damit, daß es ihnen nicht immer möglich war, den Fragebogen sofort nach Ertönen des Signaltons zu bearbeiten, was sich mit den Angaben aus Frage 7 deckt, in der 78% der Vpn sagten, daß es Situationen gegeben hat, in denen das Bearbeiten eines Protokollbogens besonders schwer oder unmöglich war (häufig genannte Situationen: während der Arbeit, Freizeit, Friseur, Sport, Autofahrt, Therapiesitzung, Schlaf, Kino, Vorlesung, Beerdigung, Kontakt- bzw. Streitsituationen). Das bedeutet, daß die Vpn zwar insgesamt viele Protokolltermine mit kurzen Latenzzeiten wahrnehmen konnten, daß es aber bei fast allen Vpn auch solche Situationen, wie die beschriebenen, gegeben hat, in denen das Protokollieren eben nicht oder nicht so einfach möglich gewesen ist. Eine Beeinträchtigung der ökologischen Repräsentativität scheint ihre Ursachen also weniger in Personparametern als mehr in situativen Gegebenheiten zu haben, die bei einer Protokollierung im Feld nur schwer zu vermeiden sind. Es gibt offensichtlich Alltagssituationen, die mit einer spontanen Protokollierung nicht vereinbar sind. (Allerdings wurden von den Vpn auch Wege gefunden, diese Schwierigkeiten zu umgehen: So ließ eine autofahrende Vp, als ein Protokolltermin bei 180 km/h auf der Autobahn eintrat, ihre mündlich gegebenen Antworten auf die Protokollbogenitems von ihrem Beifahrer notieren.)

3.4.3.3. Methodenakzeptanz

Um die Einsatzmöglichkeiten eines Instruments, wie des hier vorgestellten, abschätzen zu können, müssen Anhaltspunkte für die Akzeptanz der Methode durch die Vpn gewonnen werden, denn eine zuverlässige Datenerhebung mit Papier und Bleistift unter Feldbedingungen bei doppelter Itemvorgabe je Protokolltermin erfordert neben einer positiven a priori Einstellung der Vpn zur Untersuchung und einer hohen Compliance insgesamt eine hohe Akzeptanz der Methode in der konkreten Handhabung im Feld. Dabei sollen auch mögliche Unterschiede zwischen den Splitgruppen berücksichtigt werden.

Eine zentrale Rolle spielt in diesem Zusammenhang die Frage 14 des Nachbefragungsbogens, in der die Vpn gefragt werden, ob sie sich erneut für so eine Untersuchung zur Verfügung stellen würden.

Darauf antworteten 77% mit "ja", doch immerhin 18% mit "nein" und in 5% der Fälle konnte keine Auswertung erfolgen. Es gab bei dieser Frage nicht die Möglichkeit, im Falle einer Nein-Antwort in offener Form eine Begründung anzugeben. Dies geschah deshalb, um mögliche Nein-Sager nicht durch die Notwendigkeit einer ihnen vielleicht unangenehmen Begründung von der ehrlichen Beantwortung der Frage abzubringen. Die 5% der Vpn, die als "keine Angabe" erfaßt wurden, sind daher auch alles solche Teilnehmer, die sich nicht für eine der Antwortalternativen bedingungslos entscheiden wollten. Frage 2 erkundigt sich danach, ob die Vpn mit den Befindlichkeitsitems in den meisten Fällen ihre tatsächlich erlebten Gefühle zum Ausdruck bringen konnten. Hintergrund dieser Frage ist der Gedanke, daß bei der Konstruktion eines solchen Erhebungsinstruments nicht nur das Forschungsinteresse des Untersuchers eine Rolle spielen darf, sondern daß den Vpn auch das Gefühl gegeben werden sollte, in den Protokollbögen nach dem gefragt zu werden, was für sie wirklich wichtig ist. Diese Aspekte könnten dann z.B. bei der Auswahl der Pufferitems berücksichtigt werden oder auch direkt in die Konstruktion einer Stimmungsliste mit einfließen. Knapp zwei Drittel (65%) der Vpn bejahte diese Frage, ein gutes Drittel dagegen verneinte sie. Als Gefühle, die in der Liste gefehlt haben, wurden vor allem Ärger, Frustration, Erschöpfung und Traurigkeit genannt. Daneben bemängeln viele Vpn die generelle Brauchbarkeit der dargebotenen Items zur Beschreibung ihrer Stimmung (Items sind zu undifferenziert oder passen nicht oder sind mißverständlich, Zwischentöne fehlen u.ä.).

Für etwa genausoviele Vpn (66%) gab es an der Untersuchung irgendetwas, das sie als besonders positiv erlebt haben (Frage 3). Im einzelnen wurde dabei vor allem genannt, daß die Protokollbögen als Hilfe zur besseren Reflexion des Befindens gedient haben, aber es wurden auch vielfältige Erfahrungen im Umgang mit und im Erleben von Stimmungen genannt, z.B. daß sich Stimmungen schnell ändern können u.ä. Außerdem stand der Aspekt der Auseinandersetzung mit sich selbst, d.h. des Bewußtseinszuwachses und des Nachdenkens, im Vordergrund. Vereinzelt wurde auch erwähnt, daß die Untersuchung Spaß gemacht hat oder daß es sich um eine nette kleine Selbsterfahrung gehandelt hat. Die sich anschließende Frage 4 erkundigt sich danach, ob den Vpn an dieser Untersuchung irgendetwas eher unangenehm aufgefallen ist. Darauf antworteten wiederum 66% mit "ja". Wichtigste Argumente hierbei sind: Warten auf den Signalton, Angst ihn zu überhören, Piepen stört. Außerdem wurden die Plazierung der Protokolltermine moniert, die den Tagesablauf unterbrachen, sowie vielfältige Beanstandungen im Zusammenhang mit dem Signalgeber vorgebracht (z.B. häßlich, unhandlich, Plastikarmband u.ä.). Gleiche Items in denselben Protokollbögen führten daneben zu zwanghafter Orientierung an den bereits bearbeiteten oder sie gestalteten das Protokollieren langweilig und nervig. Immerhin 61% der Vpn gaben weiterhin in Frage 5 an, daß es irgendetwas an dieser Untersuchung gegeben hat, das für sie eine Bereicherung ihrer Lebenserfahrung gewesen ist (Frauen mit 73% signifikant häufiger als Männer mit 45%; Chi-Quadrattest). Als Präzisierung dieser Angabe wurden vor allem Argumente aus den freien Antworten in Frage 3 wiederholt: Bewußtwerden über eigene Gefühle bzw. die Schwankungsbreite und die Dynamik der Schwankungen von Stimmungen sowie bessere Reflexion des Befindens.

Die Analyse der verbleibenden Items der Frage 6 ergibt weitere Informationen zur Abschätzung der Akzeptanz dieser Untersuchungsmethode. So fühlten sich durch die Untersuchung nur 28% der Vpn kontrolliert (Item Nr. 4), davon allerdings die Psychologiestudenten signifikant stärker als die Nicht-

Psychologiestudenten (U-Test). Der Aussage "Die Untersuchung war interessant" (Item Nr. 7) stimmten 73% der Vpn ein bißchen oder ganz und gar zu, während sich 12% nicht entscheiden konnten und angaben, nicht sicher zu sein. Allerdings war sich nur eine Vp so sicher, die Untersuchung nicht interessant gefunden zu haben, daß sie "stimme gar nicht zu" angab. Dem Item Nr. 6 konnten 59% der Vpn zustimmen und somit aussagen, daß die Menschen in ihrer Umgebung auf ihre Teilnahme an der Untersuchung positiv reagiert haben, und immerhin 30% waren sich hinsichtlich dieses Items nicht sicher. Hierbei stimmten die Frauen und die Psychologiestudenten signifikant stärker zu als die jeweiligen anderen Splitgruppen (U-Test).

Zur Akzeptanz des Signalgebers gab es ebenfalls einige Items in Frage 6. So gaben 57% der Vpn in Item Nr. 11 an, daß das Tragen der Uhr nicht unangenehm war; den Frauen war es dabei tendenziell weniger unangenehm als den Männern (U-Test). Daß die Handhabung der Uhr umständlich war (Item Nr. 13), konnte fast niemand finden: 91% stimmten weniger oder gar nicht zu. Es ist auch bei kaum einer Vp vorgekommen, daß sie die Uhr verlegt hatte und sie nicht auf Anhieb wiederfinden konnte (Item Nr. 14): Diesem Item stimmten 94% weniger oder gar nicht zu. 71% der Vpn war das Piepen des Signalgebers in Gegenwart anderer Leute nicht unangenehm (Item Nr. 16); 60% der Vpn fanden nicht, daß der Weckton der Uhr zu leise war (Item Nr. 17), während doch 35% dieser Aussage zustimmten. Zur Klärung der Akzeptanz des Signalgebers trägt auch Frage 8 bei. Sie erkundigt sich danach, ob die Vpn die Uhr tagsüber immer am Handgelenk getragen haben. 76% antworteten mit "ja", die Psychologiestudenten hochsignifikant häufiger als die Nicht-Psychologiestudenten (Chi-Quadrattest). Wenn nicht am Handgelenk, so befand sich der Signalgeber zumeist neben der Vp (auf Schreibtisch, auf Schrank) oder in einer Tasche. Einige Vpn wählten auch so exotische Orte wie den Jackenkragen oder trugen ihn in der Hand. Es ist verwunderlich, daß es trotzdem weder Beschädigungen der Uhren noch Totalverluste gegeben hat. Als Gründe für ein Tragen der Uhr an anderer Stelle als dem Handgelenk wurde vor allem genannt, aus Überzeugung kein Uhrenträger zu sein. Daneben wurden Schutzmaßnahme während der Arbeit, Nickel- oder Plastikallergie oder spezielle Anlässe wie Sport usw. genannt.

Zu allerletzt wurden die Vpn im Nachbefragungsbogen aufgefordert anzugeben, ob es etwas gäbe, wonach sie noch nicht gefragt worden seien, was sie aber gerne noch äußern möchten (Frage 15). Darauf antworteten 22% der Vpn mit "ja"; als Äußerungen wurden dann eine Reihe von Bemerkungen gemacht, die bereits in anderen Zusammenhängen angebracht gewesen wären. Neu waren allerdings Sorgen um den Verbleib der erhobenen Daten (z.B. hoffentlich nicht für kommerziellen Zweck) und der Wunsch nach Aufklärung über das Untersuchungsziel, von einigen bereits vor Beginn der Untersuchung gewünscht.

Zusammenfassend läßt sich generell sagen, daß eine Datenerhebung im Feld als ESM-Studie unter Einsatz eines elektronischen Signalgebers mit einer Papier- und Bleistiftprotokollierung von Stimmungen bei doppelter Vorgabe einiger Items eine praktikable Methode der Informationsgewinnung mit einer hohen Akzeptanz auf Seiten der Vpn ist. 77% der untersuchten Vpn würden so eine Untersuchung noch einmal mitmachen, knapp zwei Drittel konnten mit den Befindlichkeitsitems ihre Stimmungen hinreichend gut ausdrücken. Ein großer Teil der Wünsche der restlichen Vpn hinsichtlich einer Modifikation der Itemsammlung könnte im Prinzip berücksichtigt

werden. Jeweils zwei Drittel der Vpn gaben an, positive aber auch negative Erlebnisse durch ihre Teilnahme gehabt zu haben; fast ebenso viele empfanden Teile der Untersuchung als eine Bereicherung ihrer Lebenserfahrung. Den geäußerten unangenehmen Aspekten muß große Aufmerksamkeit entgegengebracht werden, doch können sicher nur ein Teil der vorgebrachten Argumente durch entsprechende Modifikationen der Untersuchungskonzeption entkräftet werden. Weiter zeigt sich, daß die Mehrheit der Vpn sich durch die Untersuchung nicht kontrolliert fühlte und sie interessant fand. Außerdem reagierte das Umfeld der meisten Vpn ebenfalls positiv auf die durchgeführte Untersuchung. Zum Signalgeber gab es Beschwerden in den freien Antworten auf die Frage 4, doch es zeigt sich in der Itembatterie der Frage 6, daß die Mehrheit der Vpn weder das Tragen der Uhr unangenehm noch ihre Handhabung umständlich fand. Es hat auch fast niemand die Uhr so verlegt, daß er sie nicht auf Anhieb wiederfinden konnte; die Mehrheit der Vpn fand ihr Piepen in Gegenwart anderer Personen nicht unangenehm und außerdem den Weckton nicht zu leise. Weniger als ein Viertel der Vpn trug die Uhr anders als am Handgelenk. Den Sorgen der Vpn wegen des späteren Verwendungszwecks der Daten muß Rechnung getragen werden.

Splitgruppenunterschiede sind auf der Basis der hier besprochenen Fragen eher inkonsistent, es sollen deshalb alle Fragen des Nachbefragungsbogens noch einmal insgesamt auf signifikante Unterschiede zwischen den Splitgruppen untersucht werden. Hinsichtlich der Geschlechtsunterschiede läßt sich feststellen (signifikante Unterschiede, wenn nicht anders angegeben): Die Männer haben tendenziell eher während einer typischen Woche protokolliert als die Frauen, welche die Untersuchung wiederum eher als Bereicherung ihrer Lebenserfahrung angesehen haben und weniger geneigt waren, an den Funktionsknöpfen der Uhr herumzuspielen. Die Umgebung von Frauen reagierte zudem positiver auf ihre Teilnahme an der Untersuchung. Männer hat es tendenziell mehr gestört, daß sich die Items z.T. wiederholten, und sie fanden das Tragen der Uhr tendenziell unangenehmer, während sie sich andererseits häufiger über den Untersuchungszweck Gedanken gemacht haben. Abgesehen von gängigen Klischees läßt sich auch in diesen Unterschieden keine Systematik erkennen.

Die Psychologiestudenten fanden die Items der Protokollbögen weniger eindeutig als die Nicht-Psychologiestudenten, fühlten sich durch die Untersuchung stärker kontrolliert, störten sich mehr an den sich wiederholenden oder einander ähnelnden Items, hatten andererseits aber ein Umfeld, das positiver auf ihre Teilnahme an der Untersuchung reagierte. Psychologiestudenten dauerte die Bearbeitung eines einzelnen Protokollbogens eher zu lange, sie empfanden die Bearbeitung der Protokollbögen tendenziell eher als Störung in ihren Tagesabläufen, vergaßen aber andererseits nicht so leicht, das Ringbuch oder den Signalgeber mitzunehmen. Sie haben sich während des Beobachtungszeitraums seltener oder nicht so intensiv mit anderen über die Untersuchung unterhalten wie die Nicht-Psychologiestudenten. Schließlich haben sie die Uhr hochsignifikant häufiger am Handgelenk getragen als die Nicht-Psychologiestudenten. Diese Befunde lassen sich sinnvoll in zwei Richtungen interpretieren: Einerseits sieht es so aus, als seien die Psychologiestudenten die kritischeren Vpn mit einer höheren Bereitschaft zur Einhaltung der Instruktionen gewesen, andererseits könnte man argumentieren, sie seien einfach die desinteressierteren Vpn gewesen, die alles und kein bißchen mehr getan haben als das, was sie unbedingt tun mußten, um die Vp-Stunden bescheinigt zu bekommen, und die sich somit durch die Untersuchung auch leichter beeinträchtigt gefühlt haben. Anzunehmen ist, daß in beiden Gruppen von Vpn sehr unterschiedliche

Motivationslagen bestanden. Die Psychologiestudenten wollten ihre Pflichtscheine bekommen, die Nicht-Psychologiestudenten wollten mir bei meiner Diplomarbeit helfen und bekundeten auch darüber hinaus überwiegend ein großes Interesse an der Untersuchung, wobei der Umstand, Vp in einer wissenschaftlichen Feldstudie sein zu können, gewiß ihren Teil zur Motivation beigetragen hat. Möglicherweise waren die Psychologiestudenten sowohl die kritischeren, mit empirischer Forschungsarbeit auch vertrauteren Vpn, als auch die weniger interessierten.

Zur Abschätzung der Einsatzmöglichkeiten eines Erhebungsinstruments wie des hier vorgestellten läßt sich feststellen, daß es außerhalb des psycho-universitären Umfeldes mindestens eine genauso große Akzeptanz aufweist wie unter den Psychologiestudenten. Dabei darf nicht übersehen werden, daß seine Handhabung ein hohes eigenverantworliches Engagement seitens der Vpn ebenso voraussetzt wie ein großes Vertrauen des Untersuchers in Compliance und Aufrichtigkeit der Vpn. Es kann daher generell nur bei solchen Vpn eingesetzt werden, die zu einer überdurchschnittlich intensiven Mitarbeit bereit sind. Übersehen werden darf in diesem Zusammenhang auch nicht, daß aus den genannten Motivationsgründen die hier gezogene Stichprobe der Nicht-Psychologiestudenten nicht für alle Nicht-Psychologiestudenten repräsentativ sein kann (ganz abgesehen von den in Kap. 3.4.1. diskutierten Sachverhalten).

3.5. Konstruktion einer änderungssensitiven Stimmungsliste

3.5.1. Itemanalyse

Die Anforderungen für die Auswahl eines (der zehn relevanten) Protokollbogenitems zur Konstruktion einer änderungssensitiven Stimmungsliste (auf der Basis der in dieser Untersuchung gewonnenen Daten) werden wesentlich durch das zugrundegelegte testtheoretische Modell bestimmt. Das wichtigste Auswahlkriterium stellt die State-Charakteristik S dar, mithin die Sensitivität des Items für intraindividuelle Veränderungen der wahren Statewerte. Voraussetzung dafür ist eine hohe mittlere intraindividuelle Korrelation zwischen den Testwerten der ersten und der zweiten Messung (hohe intraindividuelle State-Retest-Reliabilität $\bar{r}_{ii'}$), denn durch sie wird entsprechend (51) der Anteil der wahren State-Varianz s^2_s an der mittleren intraindividuellen Testwertevarianz $\bar{s}^2_i$ groß, der Anteil der Fehlervarianz s^2_e gemäß (52) jedoch klein gehalten. Da das hier zu konstruierende Instrument nicht nur zur Diagnostik von intraindividuellen Stateveränderungen tauglich sein soll, sondern außerdem auch zur interindividuellen Unterscheidung der Vpn auf der Traitebene dienen soll, wird auch eine entsprechende interindividuelle Varianz der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ in den erhobenen Daten erwartet. In jedem Falle ist eine niedrige Fehlervarianz unabdingbar. Ungeeignet erscheinen daher solche Items, die nahezu keine Traitvarianz aufweisen, in denen also praktisch alle Vpn gleiche intraindividuelle Mittelwerte aufweisen, ungeachtet der Höhe der intraindividuellen Varianz. Ist die letztere dabei hinreichend groß, wären solche Items möglicherweise sogar zur Diagnostik intraindividueller Veränderungen in den untersuchten Merkmalen geeignet, weil sie auf Unterschiede in den zeit- und bedingungsabhängigen Merkmalskomponenten innerhalb der Vpn, den States, ansprechen; sie wären aber zur Diagnostik von

Unterschieden auf der Traitebene völlig unbrauchbar. Andererseits müssen solche Items als ungeeignet ausgesondert werden, die nahezu frei von intraindividueller Varianz sind, auch wenn sie über substantielle interindividuelle Varianz verfügen. Diese Items hätten von allen Vpn zu allen Protokollterminen (fast) immer denselben Wert erhalten, allerdings von jeder Vp einen anderen. Sie würden damit vor allem auf traitbedingte Unterschiede zwischen den Vpn ansprechen und nicht (so sehr) von statebedingten intraindividuellen Unterschieden beeinflußt werden.

Je höher demnach die intraindividuelle Varianz und je niedriger die interindividuelle Varianz ist, umso mehr handelt es sich um ein änderungssensitives Item mit einer hohen State-Charakteristik, dessen Meßwertevarianz fast ausschließlich auf Unterschiede innerhalb der Vpn zurückzuführen ist. Liegen die Verhältnisse genau umgekehrt (niedrige intraindividuelle, hohe interindividuelle Varianzen), so handelt es sich um ein traitabhängiges Item mit entsprechend hoher Trait-Charakteristik, dessen Lösung durch die Vpn fast ausschließlich durch zeit- und bedingungsstabile interindividuelle Unterschiede bedingt ist. Für den hier verfolgten Zweck der Testkonstruktion ist neben einer hohen State-Charakteristik auch eine geringe, von null verschiedene Trait-Charakteristik nötig, damit eine hinreichende interindividuelle Differenzierung erreicht werden kann. Dadurch verringert sich jedoch aus algebraischen Gründen gleichzeitig die State-Charakteristik, so daß diese nicht gleich eins werden kann (vgl. (53) und (54)).

Neben diesen modellbezogenen Anforderungen an die geeigneten Items ist eine nicht allzu extreme Itemschwierigkeit (nicht zu niedrig und auch nicht zu hoch) wünschenswert. Sie sollte im mittleren Bereich liegen und nicht für alle Items gleich hoch sein. Es wurde auf eine detaillierte Berechnung der Schwierigkeitsindizes für die einzelnen Items verzichtet (vgl. dazu MOOSBRUGGER, 1992, S. 312) und statt dessen mit einer Inspektion der Gesamtverteilungen der Meßwerte über alle 2669 gültigen Protokollbögen vorliebgenommen (vgl. Anhang B). Die Meßwerteverteilungen machen demnach bei allen Items einen passablen Eindruck, mit Ausnahme der Items "Ich bin ängstlich" und "Ich bin deprimiert", bei denen zwischen 79% und 85% der Antworten (1. und 2. Messung) auf die Extremkategorie "überhaupt nicht" entfallen.

Zur Bestimmung derjenigen Items, die aufgrund hoher korrelativer Zusammenhänge in den erhobenen Daten als zu einer gemeinsamen Merkmalsdimension gehörig betrachtet und damit auch zu einer gemeinsamen Skala zusammengefaßt werden können, wurden Faktorenanalysen durchgeführt. Items, die so miteinander korrelieren, werden gemeinsam auf demselben Faktor hohe Ladungen aufweisen und können daher als zusammengehörig betrachtet werden. Faktorisiert wurden zum einen die intraindividuellen Itemmittelwerte (Traitanalyse), zum anderen die um ihre intraindividuellen Mittelwerte reduzierten Einzelmeßwerte (Stateanalyse) und zwar getrennt für die erste und die zweite Messung. Erwartet wird, daß sich die Zahl der ermittelten Faktoren sowie die Ladungsmuster der Items auf den Faktoren sowohl in der Trait- wie in der Stateanalyse für jeweils beide Messungen ähneln.

3.5.1.1. Faktorenanalyse der Traitwerte

Als erstes wurden die intraindividuellen Itemmittelwerte als relativ state- und meßfehlerfreie Schätzgrößen der Traitwerte (vgl. Kap. 2.1.3.) faktorisiert. Den Ausgangspunkt bildeten dabei zwei Matrizen mit jeweils 10 intraindividuellen Itemmittelwerten für jede der 74 Vpn, wobei die eine dieser Matrizen die Werte aus der ersten und die andere die aus der zweiten Messung enthielt. Die zehn Items wurden für jede Matrix getrennt über die 74 Vpn interkorreliert und anschließend faktorenanalysiert (R-Technik).

Die Faktoren wurden nach der Methode der Hauptkomponentenanalyse extrahiert, der eine orthogonale Rotation nach der Varimax-Methode folgte. Durch die Varimax-Rotation wird eine Maximierung der Varianz der quadrierten Ladungen pro Faktor bewirkt, wobei die aufgeklärte Gesamtvarianz aller Items jedoch unverändert bleibt. Zugleich soll so eine Einfachstruktur der Faktorladungsmatrix erreicht werden, um die Interpretierbarkeit der gefundenen Faktoren zu verbessern. Die Einfachstruktur zeigt sich darin, daß die einzelnen Items möglichst in nur einem Faktor eine sehr hohe Ladung, in den anderen Faktoren dagegen möglichst niedrige Ladungen aufweisen. Die Anzahl der endgültig ermittelten Faktoren wurde durch das KAISER-Kriterium vorgegeben, wonach nur solche Faktoren extrahiert werden sollten, deren Eigenwert größer als eins war (BACKHAUS et al., 1990, BORTZ, 1989, BROSIUS, 1989). Zur Überprüfung der Qualität der Ausgangsdaten wurden der BARTLETT-Test auf Nicht-Sphärizität durchgeführt sowie das KAISER-MEYER-OLKIN-Maß bestimmt (BROSIUS, 1989, S. 143ff). Der BARTLETT-Test auf Nicht-Sphärizität überprüft die Nullhypothese, nach der die der Faktorenanalyse zugrundeliegenden Iteminterkorrelationen in der Grundgesamtheit gleich null sind. Dieser Test wird für die beiden hier faktorisierten Matrizen hochsignifikant, d.h. die 10 Items weisen auch in der Grundgesamtheit von null verschiedene Korrelationen auf. Mit dem KAISER-MEYER-OLKIN-Maß wird das Ausmaß der Iteminterkorrelationen nach Herauspartialisierung der linearen Einflüsse der anderen Variablen bestimmt. Je mehr die Varianzen zweier Variablen durch einen gemeinsamen Faktor determiniert sind, desto niedriger sind ihre Partialkorrelationskoeffizienten. Je niedriger diese sind, desto besser sind die Variablen für eine Faktorenanalyse geeignet und desto größer wird das KAISER-MEYER-OLKIN-Maß. Es kann höchstens einen Wert von eins annehmen; Werte über .8 zeigen an, daß die ausgewählten Items für die Durchführung einer Faktorenanalyse recht gut geeignet sind. Das ist bei beiden hier faktorisierten Traitmatrizen der Fall.

Die Faktorenanalyse liefert je Matrix zwei Faktoren mit Eigenwerten größer eins. Nach Varimax-Rotation ergeben sich die in den Tabellen 3 und 4 dargestellten Ladungsmatrizen.

Tab. 3 Trait-Faktorenanalyse / 1. Messung
Faktorladungsmatrix (varimax-rotiert)

	Dysph. Stimm.	Euphor. Stimm.
betrübt	0,90684	-0,18691
sorgenvoll	0,86502	-0,12168
ängstlich	0,86365	0,08882
deprimiert	0,85381	-0,20629
überfordert	0,73197	-0,25631
erfreut	-0,08874	0,95871
froh	-0,15388	0,95041
vergnügt	-0,09820	0,94027
ungezwungen	-0,39185	0,46337
gewachsen	-0,43166	0,33587
Eigenwert	3,96	3,20
aufgeklärte s^2 (%)	39,6	32,0
Σ aufgeklärte s^2 (%)	71,6	

Tab. 4 Trait-Faktorenanalyse / 2. Messung
Faktorladungsmatrix (varimax-rotiert)

	Dysph. Stimm.	Euphor. Stimm.
betrübt	0,89661	-0,19727
ängstlich	0,87724	0,04165
sorgenvoll	0,87050	-0,12618
deprimiert	0,86482	-0,22242
überfordert	0,71782	-0,29260
erfreut	-0,09559	0,95631
froh	-0,16682	0,94079
vergnügt	-0,08746	0,93896
ungezwungen	-0,42184	0,49932
gewachsen	-0,40372	0,40014
Eigenwert	3,98	3,28
aufgeklärte s^2 (%)	39,8	32,8
Σ aufgeklärte s^2 (%)	72,6	

Beide Faktoren zusammen klären in beiden Faktorenlösungen etwa gleich viel Traitvarianz auf, nämlich einmal 71.6% (1. Messung) und einmal 72.6% (2. Messung). Wie aus den Eigenwerten der Faktoren abzulesen ist, klären diese von der Größenordnung her ähnlich viel Varianz auf. Die inhaltlichen Bezeichnungen der Faktoren als dysphorische Stimmung bzw. euphorische Stimmung ergeben sich aus den Ladungen der Markieritems und entsprechen im großen und ganzen den aufgrund der Auswahlkriterien für diese Items erzeugten Erwartungen (vgl. Kap. 3.2.2.2.). Für die fünf Items der gedrückten Stimmungslage ergibt sich ein klares, den Erwartungen exakt entsprechendes Bild. Für drei der fünf Items der gehobenen Stimmungslage ("erfreut", "froh", "vergnügt") kann ebenfalls eine eindeutige Zuordnung zu dem Faktor euphorische Stimmung vorgenommen werden. Nur die Items "ungezwungen" und "gewachsen" zeigen ein unklares, nicht erwartungskonformes Muster. Zum einen haben sie in beiden Analysen auf beiden Faktoren substantielle Ladungen, zum anderen zeigt sich bei dem Item "gewachsen", daß die Ladung auf dem Faktor der dysphorischen Stimmung in beiden Fällen höher ist als auf dem Faktor der gehobenen Stimmung, wenngleich immerhin mit negativem Vorzeichen. Letzteres macht inhaltlich Sinn, doch legt dieser Befund zunächst eine bipolare Struktur des Faktors der gedrückten Stimmung und eine eher unipolare Struktur des anderen Faktors nahe. Da sich eine Zweifaktorenlösung ohne Informationsverlust und ohne Inanspruchnahme des räumlichen Vorstellungsvermögens in einer Ebene, d.h. auf dem Papier, abbilden läßt, wurden die hier ermittelten Positionen der Items im Faktorraum zusätzlich in den Abbildungen 2 und 3 veranschaulicht.

Abb. 2 Trait-Faktorenanalyse
(varimax-rotiert)
Faktorraum für 1. Messung

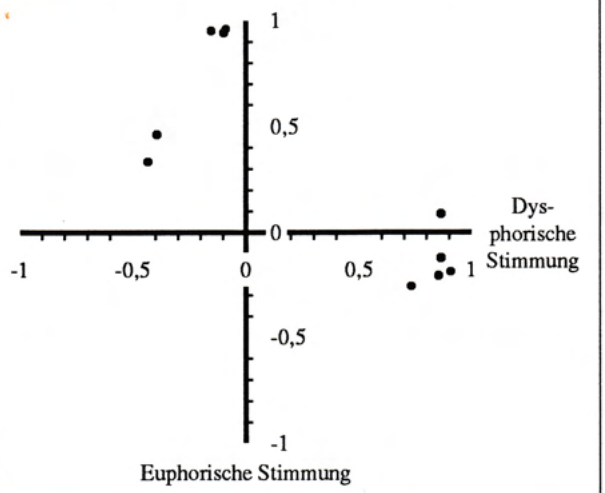
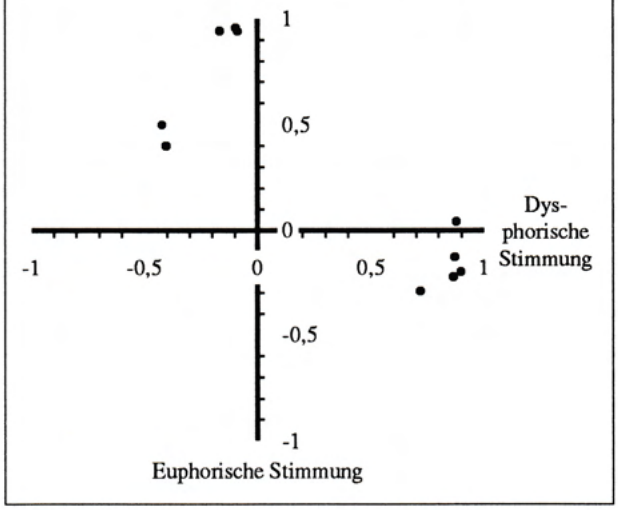


Abb. 3 Trait-Faktorenanalyse
(varimax-rotiert)
Faktorraum für 2. Messung



Es läßt sich die Ähnlichkeit der Faktorenlösungen für beide Messungen leicht per Augenschein erkennen. Das unklare Ladungsmuster der Items "ungezwungen" und "gewachsen" bekommt bei der graphischen Darstellung einen tieferen Sinn: Es zeigt sich, daß die zehn hier untersuchten Items drei Cluster im Faktorraum bilden. Die fünf eindeutig und erwartungsgemäß dysphorischen Items gruppieren sich im oberen positiven Bereich der Abszisse, die drei eindeutig euphorischen Items finden sich im oberen positiven Bereich der Ordinate und die beiden unklaren Items bilden ein Paar nahe dem Zentrum des linken oberen Quadranten, mit gleichermaßen hohen bzw. niedrigen Ladungen auf beiden Faktoren.

Dies legt nahe, versuchsweise noch einen weiteren Faktor zu extrahieren. Tabelle 5 weist das Ergebnis einer Dreifaktorenlösung aus, die sich ergibt, wenn drei zu extrahierende Faktoren als Vorgabe in die SPSS-Prozeduranweisung aufgenommen werden. Alle anderen Parameter sind gegenüber der oben durchgeführten Analyse gleich geblieben.

Tab. 5 Trait-Faktorenanalyse / 1. Messung / Vorgabe: 3 Faktoren
Faktorladungsmatrix (varimax-rotiert)

	Dysph. Stimm.	Freude	Kompetenz
betrübt	0,92034	-0,17999	-0,12210
deprimiert	0,88043	-0,21231	-0,07515
sorgenvoll	0,86553	-0,10445	-0,14594
ängstlich	0,81663	0,14699	-0,26040
überfordert	0,70550	-0,21385	-0,24269
erfreut	-0,09968	0,95085	0,16764
vergnügt	-0,11959	0,94236	0,13088
froh	-0,16046	0,93742	0,18966
gewachsen	-0,22217	0,12576	0,81877
ungezwungen	-0,20312	0,27125	0,77010
Eigenwert	3,68	2,92	1,51
aufgeklärte s^2 (%)	36,8	29,2	15,1
Σ aufgeklärte s^2 (%)	81,1		

Beim Betrachten dieser Ladungsmatrix springt sofort ins Auge, daß sich die beiden unklaren Items von den alten Faktoren abgespalten haben und einen neuen Faktor eröffnet haben. Die beiden so entstandenen neuen Faktoren haben hier die Arbeitstitel "Freude" und "Kompetenz" erhalten. Insgesamt wird mit dieser Lösung gegenüber dem Zweifaktorenmodell 9.5% mehr Itemvarianz ausgeschöpft. Für die zweite Messung zeigt sich ein ähnlich deutliches Ergebnis, das hier nicht dokumentiert ist.

Bis hierher kann zusammenfassend festgehalten werden, daß die zehn Items in beiden Messungen auf der Traitebene zwei bis drei Konstrukte zu messen scheinen. Dabei macht die (hypothetisierte) allgemeine gedrückte Stimmungslage einen recht soliden Eindruck, ebenso wie die mit dem Ausdruck von Freude verbundenen drei Items, die zwar einerseits sehr eng miteinander kovariieren, sich jedoch andererseits wider Erwarten nicht mit den Items "ungezwungen" und "gewachsen" zu einem Faktor der (hypothetisierten) allgemeinen gehobenen Stimmungslage vereinen.

Die Dreifaktorenlösung legt drei voneinander relativ unabhängige Konstrukte nahe, die jedoch durch ihre deutliche Unipolarität den Verdacht erwecken, möglicherweise artefaktbedingt zustande gekommen zu sein. Als Artefaktquellen kommen mindestens alle oben besprochenen Einflußgrößen in Frage (vgl. Kap. 2.3.3.), d.h. die Struktur des Antwortformats könnte dafür ebenso verantwortlich sein wie z.B. ein für die hypothetisierten Konstrukte mangelhaft bestückter, weil viel zu kleiner, Itempool oder auch ein möglicher Response Set.

3.5.1.2. Response Set

Einflüsse der Skalierung bzw. der Formulierung der Antwortkategorien können allerdings an dieser Stelle aus einleuchtenden Gründen genauso wenig ex post untersucht werden, wie etwa der Itempool vergrößert werden kann. Eine Untersuchung der Iteminterkorrelationen auf das Vorliegen eines Response Sets ist aber durchaus möglich. Wie oben dargelegt (vgl. Kap. 2.3.3.), ist das Auftreten von bipolaren, dem Alltagsverständnis genauso wie den a priori Hypothesen vieler Stimmungsforscher (z.B. NOWLIS, 1965) entsprechenden Befindlichkeitsfaktoren an das Vorhandensein von hinreichend vielen und genügend hohen negativen Korrelationen zwischen hypothetisch antonymen Items gebunden. Vorhandene negative Korrelationen können jedoch, etwa durch das Vorliegen eines Response Sets, gegenüber dem wahren Zusammenhang der Variablen vermindert und positive Korrelationen dadurch sogar noch erhöht werden. Wie eine Inspektion von Tabelle 6 zeigt, in der die Interkorrelationen der intraindividuellen Itemmittelwerte für die erste Messung dargestellt sind, ergibt sich für die vorliegenden Daten in der Tat, daß die negativen Korrelationskoeffizienten höchstens bis -.363 ("deprimiert" / "gewachsen") reichen (und damit gerade 13% der gegenseitigen Varianz determinieren), während die positiven Koeffizienten bis .926 ("froh" / "erfreut") reichen (entsprechend knapp 86% Varianzaufklärung).

Tab. 6 Interkorrelationen der intraindividuellen Mittelwerte (1. Messung)

	deprimiert	ungezw.	sorgenvoll	vergnügt	überfordert	froh	erfreut	gewachsen	ängstlich
ungezwungen	-0,266								
sorgenvoll	0,697	-0,359							
vergnügt	-0,282	0,383	-0,247						
überfordert	0,633	-0,358	0,560	-0,317					
froh	-0,348	0,428	-0,261	0,893	-0,333				
erfreut	-0,294	0,366	-0,211	0,903	-0,297	0,926			
gewachsen	-0,363	0,447	-0,287	0,240	-0,356	0,309	0,310		
ängstlich	0,594	-0,322	0,734	-0,039	0,561	-0,065	-0,015	-0,320	
betrübt	0,918	-0,349	0,798	-0,264	0,626	-0,347	-0,279	-0,350	0,683

Diese Struktur ist zwar nicht maximal stark ausgeprägt, es liegen immerhin mehrere Koeffizienten um -.35 herum, während die schwächste positive Beziehung auch nur bei .240 ("vergnügt" / "gewachsen") liegt, doch ist sie deutlich erkennbar. (Die Korrelationsmatrix für die zweite Messung sieht im übrigen ähnlich aus.) Um das Wirksamwerden eines hypothetischen Response Sets nachzuweisen, wonach einige Vpn dazu neigen, auf der Skala stets etwas zu hohe oder stets etwas zu niedrige Werte anzugeben, wird die unrotierte Faktorenlösung betrachtet. Dazu wurde die gleiche Prozedur wie bei der ersten Faktorenanalyse angewandt (vgl. Tabellen 3 und 4 bzw. Abbildungen 2 und 3), jedoch ohne Rotation. Ein auf alle Variablen gleich wirksamer Einfluß des hypothetisierten Response Sets müßte sich in einem Faktor äußern, auf dem alle Items positiv laden. Zugleich sollte sich bei den Befindlichkeitsdimensionen ein deutlich bipolarer Faktor zeigen. Daß es sich dabei nur um einen einzigen Faktor handeln kann, zeigt das Ergebnis der ersten, varimax-rotierten Analyse, in der auch nur zwei Faktoren extrahiert wurden. Da die Varimax-Lösung aus der unrotierten gewonnen wurde, kann diese ebenfalls nur aus zwei Faktoren bestehen; auch die aufgeklärte Gesamtvarianz muß gegenüber der rotierten Lösung unverändert bleiben. Die Tabellen 7 und 8 zeigen die unrotierten Faktorenlösungen für die erste und die zweite Messung auf der Traitebene.

Tab. 7 Trait-Faktorenanalyse / 1. Messung
Faktorladungsmatrix (unrotiert)

	Content Faktor	Set Faktor
betrübt	0,84170	0,38578
deprimiert	0,81041	0,33878
sorgenvoll	0,76940	0,41363
überfordert	0,74177	0,22638
ängstlich	0,64376	0,58254
gewachsen	-0,54672	0,01543
ungezwungen	-0,59005	0,14177
vergnügt	-0,63542	0,69999
erfreut	-0,63871	0,72045
froh	-0,68632	0,67522
Eigenwert	4,85	2,31
aufgeklärte s^2 (%)	48,5	23,1
Σ aufgeklärte s^2 (%)	71,6	

Tab. 8 Trait-Faktorenanalyse / 2. Messung
Faktorladungsmatrix (unrotiert)

	Content Faktor	Set Faktor
betrübt	0,82949	0,39342
deprimiert	0,81976	0,35407
sorgenvoll	0,76529	0,43361
überfordert	0,74659	0,20854
ängstlich	0,66779	0,57038
gewachsen	-0,56423	0,06890
ungezwungen	-0,63931	0,13618
vergnügt	-0,64440	0,68851
erfreut	-0,66146	0,69723
froh	-0,70825	0,64133
Eigenwert	5,03	2,23
aufgeklärte s^2 (%)	50,3	22,3
Σ aufgeklärte s^2 (%)	72,6	

Erwartungsgemäß ergeben sich für beide Messungen jeweils zwei Faktoren, die, wie erwähnt, insgesamt genauso viel Varianz erklären wie die rotierten. Der Vergleich mit den Tabellen 3 und 4 läßt aber erkennen, daß bei den hier gefundenen unrotierten Lösungen die jeweils zweiten Faktoren erheblich niedrigere Eigenwerte als die ersten Faktoren aufweisen und damit deutlich weniger Varianz als diese erklären. Zwar sind auch in der rotierten Lösung die ersten Faktoren varianzstärker (deshalb stehen sie ja an erster Stelle), doch die Differenzen zu den zweiten Faktoren sind bei weitem nicht so groß wie hier. Die Ladungsmuster entsprechen ebenfalls den Erwartungen: In den ersten Faktoren beider Messungen, hier "Content Faktoren" genannt, laden die Items der dysphorischen Stimmung substantiell positiv, genau wie in der rotierten Lösung. Die Items der euphorischen Stimmung laden hier negativ, in Entsprechung zur ersten Lösung, jedoch alle auch mit substantiellen Ladungen - im Gegensatz zur ersten Lösung. (Bei dieser luden dort nur die Items "ungezwungen" und "gewachsen" nennenswert hoch.) Die zweiten Faktoren aus beiden Messungen tragen ausschließlich positive Itemladungen und können daher im hier erläuterten Zusammenhang als "Set Faktoren" angesprochen werden. Auffällig an ihnen ist die extrem unterschiedliche Höhe der Ladungen von nahezu null ("gewachsen", beide Messungen) bis hin zu .72 ("erfreut", erste Messung) sowie die Tatsache, daß die Ladungen für zwei Items ("vergnügt", "erfreut") im Set Faktor vom Betrag her höher sind als im Content Faktor.

Die Abbildungen 4 und 5 veranschaulichen diese Zusammenhänge noch einmal graphisch. Die fünf dysphorischen Items liegen als Gruppe im rechten oberen Quadranten, die fünf euphorischen im linken oberen Quadranten. Dabei bilden die Items "ungezwungen" und "gewachsen" wiederum ein Zweiercluster, diesmal nahe an der Abszisse; die anderen drei Items bilden ein weiter darüber gelegenes Cluster. Ein Vergleich mit den Abbildungen 2 und 3 verdeutlicht, daß sich die Lage der Punkte (Items) in Relation zueinander nicht verändert hat.

Abb. 4 Trait-Faktorenanalyse (unrotiert)
Faktorraum für 1. Messung

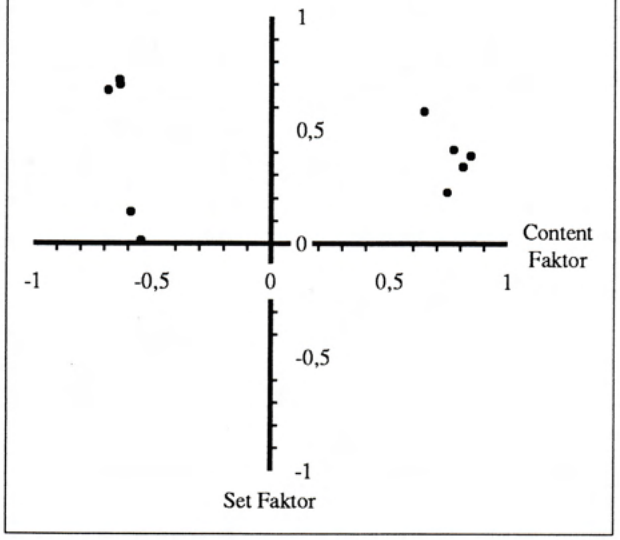
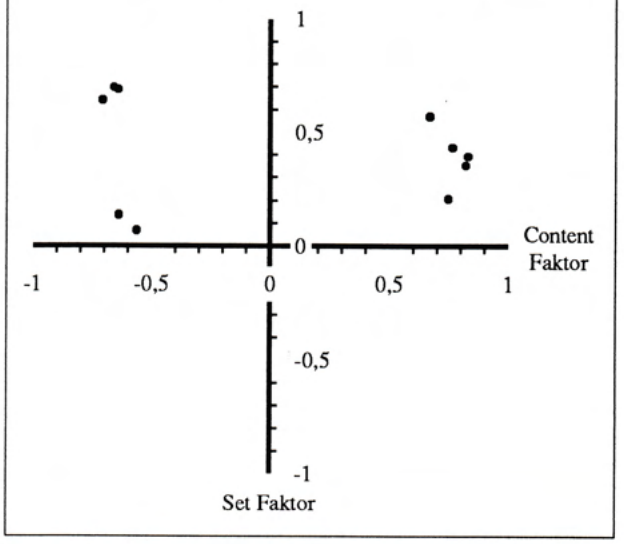


Abb. 5 Trait-Faktorenanalyse (unrotiert)
Faktorraum für 2. Messung



Festzuhalten bleibt, daß die Items beider Messungen in den nicht rotierten Ausgangslösungen substantiell und bipolar auf dem ersten, varianzstärksten Faktor laden und auf dem zweiten Faktor zwar unterschiedlich hoch, in jedem Falle aber positiv laden. Dies bedeutet einen klaren Hinweis auf

die Bipolarität des hier erfaßten Stimmungsraums sowohl für die erste wie für die zweite Messung, auch wenn die Items der euphorischen Stimmung in sich heterogen sind und in zwei Cluster zerfallen.

Man kann zum Zwecke einer Korrektur der ermittelten Varianz der Traitwerte s^2_t (vgl. dazu Kap. 3.4.2.1.) diejenigen ihrer Anteile, die durch den Response Set determiniert sind, rechnerisch so eliminieren, wie es von BUSE & PAWLIK (1991) durchgeführt wurde: Sie multiplizierten die Traitvarianz mit dem von eins subtrahierten Ladungsquadrat des Items i im Set Faktor (Faktor 2) und erhielten dadurch einen vom Set Einfluß bereinigten, korrigierten Traitvarianzwert $s^2_{t'}$.

(57) $s^2_{t'} = s^2_t * (1 - a^2_{i2})$ (ebd., S. 532)

Vorher rotierten sie die Ausgangslösung noch so, daß die Ladungen aller Items im Response Set Faktor möglichst gleich hoch und immer noch positiv waren.

Dieses Verfahren wurde zu Vergleichszwecken hier ebenfalls angewandt. Ausgangspunkt waren die Traitvarianzen aus Tabelle 1 sowie die Ladungen auf den Set Faktoren aus den Tabellen 7 und 8. Die Ergebnisse sind in Tabelle 9 dargestellt.

Tab. 9 Korrigierte Parameter auf der Basis des Modells von BUSE & PAWLIK (1991)

		s^2_t	$s^2_{t'}$
1. Messung	deprimiert	0,07	0,06
	sorgenvoll	0,08	0,07
	überfordert	0,08	0,07
	ängstlich	0,04	0,03
	betrübt	0,08	0,07
	ungezwungen	0,11	0,11
	gewachsen	0,21	0,21
	vergnügt	0,20	0,10
	froh	0,20	0,11
	erfreut	0,17	0,08
2. Messung	deprimiert	0,09	0,08
	sorgenvoll	0,11	0,09
	überfordert	0,09	0,08
	ängstlich	0,07	0,05
	betrübt	0,11	0,09
	ungezwungen	0,16	0,16
	gewachsen	0,22	0,22
	vergnügt	0,21	0,11
	froh	0,22	0,13
	erfreut	0,19	0,10

s^2_t Varianz der wahren Traitwerte

$s^2_{t'}$ korrigierte Varianz der wahren Traitwerte

3.5.1.3. Faktorenanalyse der Statewerte

Zur Faktorenanalyse der Statewerte wurden von den einzelnen Meßwerten aller wahrgenommenen Protokolltermine einer jeden Vp itemweise die dazugehörigen intraindividuellen Mittelwerte (als Schätzgrößen der Traitwerte) subtrahiert. Die so errechneten Differenzwerte stellen eine Schätzung der um die Traitkomponente bereinigten Statewerte dar, die einen für alle Vpn und alle Items gleichen intraindividuellen Mittelwert von null haben. Die einzelnen Matrizen mit den intraindividuellen Abweichungswerten von allen 74 Vpn (Protokolltermine in den Zeilen, Items in den Spalten) wurden "hintereinandergehängt". Die jeweils zehn Items aus den ersten und zweiten Messungen wurden getrennt über diese insgesamt 2669 gültigen Protokolltermine faktorisiert (Ketten-P-Technik). Es wurden wiederum Hauptkomponentenanalysen mit anschließender Varimax-Rotation durchgeführt. Der zuvor durchgeführte BARTLETT-Test auf Nicht-Sphärizität wird in beiden Fällen hochsignifikant, das KAISER-MEYER-OLKIN-Maß liegt knapp unter .9, womit auch für die Stateanalysen in dieser Hinsicht günstige Bedingungen vorliegen (vgl. Kap. 3.5.1.1.).

Vorweg kann bereits gesagt werden, daß die Ergebnisse mit denen aus der Analyse der Traitwerte vergleichbar sind. Es ergeben sich wiederum zwei Faktoren, die zusammen in beiden Messungen jeweils allerdings nur mehr gut 60% der Itemvarianz aufklären (Tabellen 10 und 11), was sich damit erklärt, daß die hier analysierten einzelnen Meßwerte unreliabler sind als die intraindividuellen Itemmittelwerte in den Traitanalysen.

Tab. 10 State-Faktorenanalyse / 1. Messung
Faktorladungsmatrix (varimax-rotiert)

	Dysph. Stimm.	Euphor. Stimm.
sorgenvoll	0,76521	-0,20429
ängstlich	0,73480	-0,04037
deprimiert	0,72117	-0,26873
betrübt	0,69910	-0,31704
überfordert	0,62342	-0,27496
erfreut	-0,18999	0,86860
froh	-0,20865	0,85933
vergnügt	-0,22700	0,84996
ungezwungen	-0,27310	0,56891
gewachsen	-0,51133	0,43105
Eigenwert	2,99	3,02
aufgeklärte s^2 (%)	29,9	30,2
Σ aufgeklärte s^2 (%)	60,1	

Tab. 11 State-Faktorenanalyse / 2. Messung
Faktorladungsmatrix (varimax-rotiert)

	Dysph. Stimm.	Euphor. Stimm.
sorgenvoll	0,76583	-0,24636
deprimiert	0,75645	-0,23900
ängstlich	0,75308	-0,06977
betrübt	0,72717	-0,31149
überfordert	0,62983	-0,31020
erfreut	-0,19648	0,88185
froh	-0,23322	0,86375
vergnügt	-0,22878	0,86252
ungezwungen	-0,36990	0,53487
gewachsen	-0,51452	0,45486
Eigenwert	3,20	3,08
aufgeklärte s^2 (%)	32,0	30,8
Σ aufgeklärte s^2 (%)	62,7	

Der erste Faktor ist bei beiden Messungen wieder charakterisiert durch hohe Ladungen der dysphorischen Items, der zweite Faktor trägt wieder die höchsten Ladungen der drei zum Konzept der Freude gehörigen Items. All diese Ladungen sind jedoch niedriger als bei den Traitanalysen. Wieder sind es dagegen die Items "ungezwungen" und "gewachsen", die aus dem Rahmen fallen. Für beide Messungen gilt hier, daß die Ladungen bei Analyse der Statewerte auf dem zweiten Faktor höher sind als bei der Analyse der Traitwerte, was sich im übrigen auch beim Betrag der Ladung von

"gewachsen" auf dem ersten Faktor zeigt. Die Ladung von "ungezwungen" auf dem ersten Faktor wird jedoch von der Trait- zur Stateanalyse vom Betrag her niedriger. Die Abbildungen 6 und 7 stellen dies noch einmal graphisch dar.

Abb. 6 State-Faktorenanalyse (varimax-rotiert) Faktorraum für 1. Messung

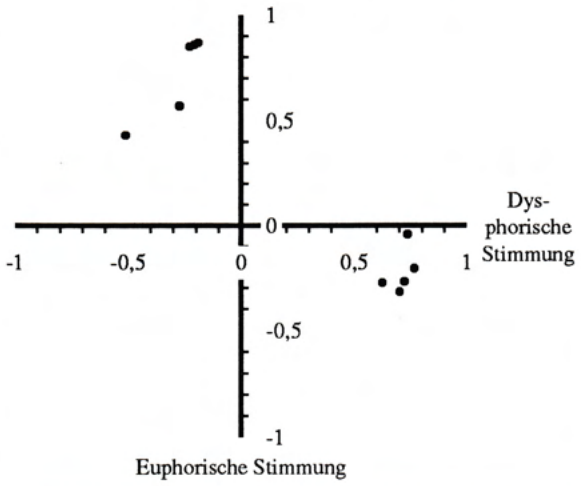
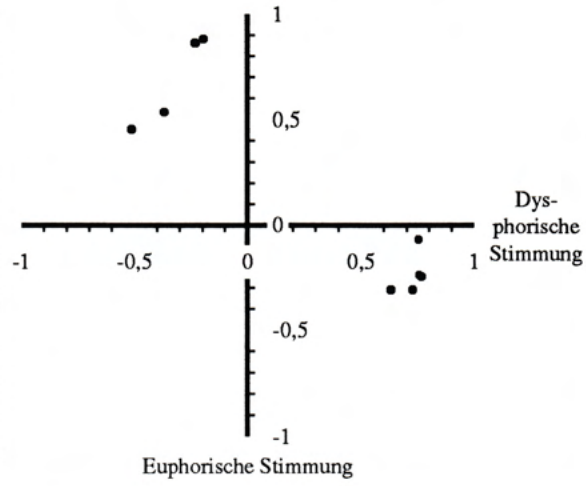


Abb. 7 State-Faktorenanalyse (varimax-rotiert) Faktorraum für 2. Messung



Trotz aller numerischen Unterschiede zwischen Trait- und Stateanalysen zeigt diese Darstellung recht gut die Ähnlichkeit zwischen beiden Lösungen. Wieder gruppieren sich die fünf dysphorischen Items in der Nähe der Abszisse in derem positiven Bereich, während die fünf euphorischen Items im oberen linken Quadranten liegen. Sowohl die dysphorischen Items als auch die drei Items zur freudigen Gestimmtheit haben sich dabei räumlich enger zueinander hin orientiert, während sich die beiden anderen Items ("ungezwungen" und "gewachsen") gegenüber den Traitlösungen räumlich voneinander entfernt haben und nun nicht mehr so leicht als Zweiergruppe zu identifizieren sind.

Beim Betrachten der entsprechenden Dreifaktorenlösungen (varimax-rotiert; nicht dokumentiert) fällt im Vergleich mit der Analyse der Traitwerte (vgl. Tabelle 5) auf, daß sich das dysphorische Item "Ich fühle mich überfordert" vom Faktor der dysphorischen Stimmung abgetrennt hat (Restladung hier jeweils pro Messung nur noch knapp .4) und nun substantiell negativ auf dem Faktor Kompetenz (mit den Items "gewachsen" und "ungezwungen") mit jeweils -.7 lädt.

Vergleicht man die Faktorenlösungen der ersten und zweiten Messung (jeweils auf der State- und der Traitebene) miteinander und außerdem jeweils innerhalb einer Messung die State- mit den Traitfaktoren, so bleibt der Eindruck bestehen, daß eine zur Konstruktion eines State-Trait-Instruments ausreichende Ähnlichkeit der vier verschiedenen Faktorenstrukturen besteht.

3.5.2. Skalenkonstruktion

3.5.2.1. Itemselektion

Um ein Meßinstrument ("Test") zu konstruieren, das zunächst für Forschungszwecke, später aber auch für individualdiagnostische Zwecke eingesetzt werden soll, ist es nötig und üblich, mehrere Items, die verschiedene Aspekte des gleichen Konstrukts messen sollen, zu einer Skala zusammenzufassen und die Meßwerte der einzelnen Items zu einem Summenwert zu aggregieren. Dieser Summenwert ist reliabler als die Einzelitems, da er meßfehlerreduziert ist (vgl. Kap. 2.3.1.).

Wie bereits erläutert (vgl. Kap. 3.5.1.), ist das wesentliche formale Kriterium zur Konstruktion eines änderungssensitiven Meßinstruments der Befindlichkeit das Ausmaß der Änderungssensitivität S der Items. Diese darf aber nicht maximal sein, denn die Items müssen zugleich eine klar von null verschiedene Traitcharakteristik T aufweisen, um auch zum Zwecke der Traitdiagnostik differenzierte Informationen liefern zu können. Betrachtet man nun die entsprechenden Itemparameter in Tabelle 1, so zeigt sich eine genügend hohe Traitcharakteristik für alle Items. Für die Statecharakteristik als der hier relevanten Änderungssensitivität wurde bereits oben festgestellt (vgl. Kap. 3.4.2.1.), daß sie zumindest für die erste Messung bei fast allen Items (Ausnahme: "gewachsen") hinreichend hoch ist. Anläßlich der zweiten Messung reduziert sie sich zwar (Ausnahme: "vergnügt"), doch liegt sie in allen Fällen noch über .5, was bedeutet, daß die Varianz der wahren Werte in allen Fällen zu mehr als 50% auf die Varianz der Statewerte zurückzuführen ist - auch bei dem Item "Ich fühle mich gewachsen", wenngleich die 50% dort nur knapp überschritten werden.

Zusätzlich wird als Selektionskriterium eine annehmbar hohe durchschnittliche intraindividuelle Korrelation als State-Retest-Reliabilität $\bar{r}_{ii'}$ gefordert, die eine niedrige Fehlervarianz s^2_e impliziert. Die in Tabelle 1 angegebenen Koeffizienten weisen eine beachtliche Höhe auf, reflektieren sie doch mittlere Zusammenhänge auf Itemebene. Auf das Problem der Itemschwierigkeit wurde bereits eingegangen (vgl. Kap. 3.5.1.). Demnach erweisen sich zwar einige Items als recht schwierig ("deprimiert" und "ängstlich"), die meisten weisen jedoch eine annehmbare Schwierigkeit auf. Dieser Parameter soll hier lediglich zur Kenntnis genommen, aber nicht weiter analysiert werden. Schon bei der Vorauswahl der Items für diese Untersuchung ließ sich nicht in allen Fällen (insbesondere bei den Items der dysphorischen Stimmung) ein mittlerer Schwierigkeitsindex realisieren (vgl. Kap. 3.2.2.2.).

Bis hierher bleibt bei der Bewertung der Items hinsichtlich ihrer Eignung zur Skalenbildung festzustellen, daß sie alle, sowohl in der ersten wie in der zweiten Messung, als gut brauchbar angesehen werden können. Zwar erweisen sich die Statecharakteristiken nicht in allen Fällen als so stark ausgeprägt, wie es wünschenswert wäre, sie reduzieren sich zudem von der ersten zur zweiten Messung, doch sind sie alle vom Betrag her so groß (S>.5), daß nicht einzelne Items ausgesondert werden müßten. Zusätzlich sind die Traitcharakteristiken ebenso wie die intraindividuellen Korrelationen zwischen erster und zweiter Messung genügend hoch, so daß bei wohlwollender Betrachtung (unter Absehung der z.T. extremen Itemschwierigkeiten) alle Items für die

Konstruktion einer oder mehrerer änderungssensitiver State-Trait-Skalen zur Messung der Befindlichkeit in Frage kommen. Damit ist aber noch nicht geklärt, in welcher Weise die Items kombiniert werden sollen. Wichtig für die Kombination mehrerer Items zum Zwecke der Skalenbildung und der damit intendierten psychometrischen Erfassung eines durch die Skala zu messenden Merkmals ist ihre gemeinsame Kovariation, und zwar sowohl auf der State- wie auf der Traitebene. Dies wurde mit Hilfe der durchgeführten Faktorenanalysen überprüft (vgl. Kap. 3.5.1.1. bis 3.5.1.3.).

Bei deren Interpretation ist zunächst der Befund bedeutsam, daß die Faktorenstruktur "robust" ist, d.h. daß sie sich offensichtlich von der ersten zur zweiten Messung nicht grundlegend ändert. Dasselbe gilt auch für den Vergleich von State- und Traitfaktoren. Darin zeigt sich, daß die Items in beiden Messungen und auf beiden Ebenen im wesentlichen die gleichen Konstrukte messen, auch wenn sich die Mittelwerte von der ersten zur zweiten Messung signifikant verändern (Reaktivitätseffekt; vgl. Kap. 3.4.2.2.) und die Varianz der Items zumindest teilweise von einem Response Set überlagert zu sein scheint (vgl. Kap. 3.5.1.2.). Bei Auswahl der zehn Items für diese Untersuchung wurde erwartet, daß jeweils fünf von ihnen das Konstrukt "allgemeine gehobene Stimmung" bzw. "allgemeine gedrückte Stimmung" messen würden (vgl. Kap. 3.2.2.2.). Dabei wurde prinzipiell offen gelassen, ob sich diese Konstrukte als bipolare Struktur auf einem gemeinsamen Faktor darstellen lassen würden oder ob sie als unipolare Faktoren unabhängig voneinander wären. Geht man nun von den empirischen Ergebnissen der unrotierten, bipolaren Lösungen aus, so zeigt sich, daß die bei Rotation entstehende (relative) Unipolarität der Faktoren höchstwahrscheinlich durch Einflüsse eines Response Sets oder anderer, die Iteminterkorrelationen in positiver Richtung verschiebender Determinanten, zustande gekommen ist. Daß die Unipolarität der rotierten Lösungen als Artefakt angesehen werden kann, zeigt sich auch bei näherem Betrachten der entsprechenden Ladungsmatrizen (vgl. Tabellen 3, 4, 10, 11). Auf den ersten Blick fallen die z.T. extrem hohen positiven Ladungen derjenigen Items auf, die zu demselben Konstrukt gehören. Auf den zweiten Blick läßt sich jedoch erkennen, daß die Items des jeweils anderen Konstrukts zwar nur substanzlose Ladungen auf ihren antonymen Faktoren aufweisen, daß diese jedoch in fast allen Fällen negativ sind und dadurch die Bipolarität des gesamten untersuchten Befindlichkeitsbereichs untermauern.

Betrachtet man die graphischen Darstellungen der Faktorenanalysen auf der State- wie auf der Traitebene, in rotierter oder unrotierter Form (vgl. Abbildungen 2 bis 7), so fällt auf, daß sich die zehn Items in drei Clustern gruppieren, eines mit fünf Items (alle dysphorischen Items), eines mit drei Items ("froh", "erfreut", "vergnügt") und ein z.T. nicht ganz klares mit zwei Items ("gewachsen", "ungezwungen"), welches sich in einer Dreifaktorenlösung besser darstellen läßt als in der Zweifaktorenlösung. In allen Fällen haben die beiden euphorischen Cluster mit den drei bzw. zwei Items einen geringeren räumlichen Abstand voneinander als jedes von ihnen zu dem dysphorischen Fünfercluster. Dies spricht für eine gewisse Ähnlichkeit der beiden "kleinen" Itemgruppen, die man vielleicht als Zusammengehörigkeit zu einem Gesamtkonzept der euphorischen Stimmung auffassen kann. (Diese Vermutung wird bestärkt durch das Ladungsmuster der Dreifaktorenlösung in Tabelle 5: Die Items der Freude bzw. der Kompetenz laden in dem jeweiligen anderen Faktor nur geringfügig, aber ausnahmslos positiv.) Außerdem spricht dies auch hier für eine zumindest optisch einleuchtende Bipolarität des gesamten erfaßten Merkmalsraums, denn die beiden

"kleinen" Cluster stehen dem großen Fünfercluster räumlich deutlich gegenüber. Daß sich die euphorischen Items dabei nicht so eng wie die dysphorischen gruppieren, mag an dem sehr geringen Itempool liegen, vielleicht aber auch daran, daß die einen drei Items mit den anderen zwei auch schon inhaltlich relativ wenig zu tun haben.

Faßt man die bisherigen Ergebnisse zusammen, so kommen alle zehn Items für die Bildung einer bipolaren Stimmungsskala in Frage. Die heterogene Struktur der gehobenen Stimmungslage jedoch, die vermutete, aber nicht wirklich bewiesene Bipolarität des gesamten Stimmungsraums sowie der Erkundungs- bzw. Pretestcharakter der gesamten Untersuchung gestatten es nicht nur, sondern erfordern es geradezu, auf der Grundlage der bisher zur Verfügung stehenden Informationen nicht bloß eine einzige Skala zu bilden, sondern dabei mehrere Varianten durchzuspielen und deren psychometrische Eigenschaften auf der Ebene des zugrundeliegenden testtheoretischen Modells zu analysieren. Erst danach soll eine Empfehlung für weitere Konstruktionsschritte ausgesprochen werden. Neben bipolaren Konfigurationen sollen dabei auch, unter der Prämisse einer eventuellen Unipolarität und Eigenständigkeit der ermittelten Dimensionen, mögliche unipolare Skalen mit einbezogen werden.

Es wurden sechs verschiedene Skalen gebildet, drei bipolare, mit "B" bezeichnete, sowie drei unipolare, mit dem Kennbuchstaben "U". Im einzelnen handelt es sich um folgende Itemkonfigurationen (die Zahlen hinter den Kennbuchstaben bezeichnen die Anzahl der zugrundeliegenden Items):

B10 umfaßt alle zehn Items

B8 wie B10, jedoch ohne "gewachsen" und "ungezwungen"

B7 wie B10, jedoch ohne "erfreut", "vergnügt" und "froh"

U5 umfaßt alle fünf dysphorischen Items

U3 umfaßt die drei euphorischen Items "erfreut", "vergnügt" und "froh"

U2 umfaßt die zwei euphorischen Items "gewachsen" und "ungezwungen"

Ein Skalenwert wird pro Vp, Protokolltermin und Messung durch Addition der Einzelscores der die Skala bildenden Items errechnet. Im Falle der drei bipolaren Skalen wurden dazu die Scores der dysphorischen Items vor der Aggregierung noch in ihre Komplementärwerte umgepolt (umgepolter Wert = 5 - angekreuztem Wert), so daß für diese drei Skalen ein hoher Skalenwert eine hohe Ausprägung der euphorischen und gleichzeitig eine geringe Ausprägung der dysphorischen Stimmung indiziert. Auch die beiden unipolaren Skalen der gehobenen Stimmung (U3 und U2) zählen selbstverständlich in Richtung der Ausprägung dieser Stimmung (ein hoher Skalenwert bedeutet gehobene Stimmung). Lediglich die Skala U5 mit ihren fünf dysphorischen Items macht insofern eine Ausnahme, als hier infolge einfacher Addition der Itemscores ein hoher Skalenwert eine

gedrückte Stimmung bedeutet. Dies ist bei der Interpretation korrelativer Zusammenhänge zu beachten. Wegen der Heterogenität der fünf euphorischen Items wurde auf die Bildung einer entsprechenden, alle diese Items umfassenden Skala verzichtet.

3.5.2.2. Modellparameter auf Skalenebene

Für die so gebildeten möglichen Merkmalsskalen wurden die entsprechenden Skalenwerte für alle Vpn und für alle Protokolltermine errechnet. Anschließend wurden in Analogie zu der Berechnung der Modellparameter auf Itemebene (vgl. Kap. 3.4.2.1.) diese Kennwerte auch für die sechs verschiedenen Skalen bestimmt (Tabelle 12). Die Höhe der angegebenen mittleren intraindividuellen Mittelwerte muß in Relation zu den minimal und maximal möglichen Skalenwerten betrachtet werden. Pro Item, das zu einer Skala gehört, wird mindestens ein Punkt erzielt, höchstens jedoch vier Punkte. Die Tabelle 12 enthält diese Eckwerte in Klammern hinter den Skalenbezeichnungen.

Tab. 12 Parameter für einzelne Itemkombinationen (Skalen) auf der Basis des Modells von BUSE & PAWLIK (1991)

		$\bar{\bar{x}}$	$s^2_{\bar{x}}$	$\bar{s}^2_i$	$\bar{r}_{ii'}$	s^2_t	s^2_s	s^2_e	T	S	R
1. Messung	**B10** (min. 10 / max. 40 Punkte)	31,075	6,113	18,420	0,910	5,60	16,76	1,66	0,25	0,75	0,77
	B8 (min. 8 / max. 32 Punkte)	25,203	4,096	11,932	0,911	3,77	10,87	1,06	0,26	0,74	0,78
	B7 (min. 7 / max. 28 Punkte)	24,378	2,681	7,403	0,851	2,48	6,30	1,10	0,28	0,72	0,69
	U5 (min. 5 / max. 20 Punkte)	6,495	1,415	3,585	0,844	1,32	3,03	0,56	0,30	0,70	0,70
	U3 (min. 3 / max. 12 Punkte)	6,698	1,742	4,275	0,892	1,62	3,81	0,46	0,30	0,70	0,78
	U2 (min. 2 / max. 8 Punkte)	5,873	0,503	1,397	0,730	0,46	1,02	0,38	0,31	0,69	0,55
2. Messung	**B10** (min. 10 / max. 40 Punkte)	30,656	7,519	19,875	0,910	6,97	18,09	1,79	0,28	0,72	0,80
	B8 (min. 8 / max. 32 Punkte)	24,967	4,977	12,973	0,911	4,62	11,82	1,15	0,28	0,72	0,80
	B7 (min. 7 / max. 28 Punkte)	24,054	3,480	8,043	0,851	3,26	6,84	1,20	0,32	0,68	0,73
	U5 (min. 5 / max. 20 Punkte)	6,635	1,878	3,942	0,844	1,77	3,33	0,61	0,35	0,65	0,74
	U3 (min. 3 / max. 12 Punkte)	6,602	1,886	4,483	0,892	1,76	4,00	0,48	0,31	0,69	0,78
	U2 (min. 2 / max. 8 Punkte)	5,690	0,614	1,368	0,730	0,58	1,00	0,37	0,37	0,63	0,61

$\bar{\bar{x}}$	mittlerer intraindividueller Mittelwert	s^2_t	Varianz der wahren Traitwerte
$s^2_{\bar{x}}$	Varianz der intraindividuellen Mittelwerte	s^2_s	Varianz der wahren Statewerte
$\bar{s}^2_i$	mittlere intraindividuelle Varianz	s^2_e	Varianz der Fehlerwerte
$\bar{r}_{ii'}$	mittlere intraindividuelle Korrelation (zwischen 1. und 2. Messung)	T	Traitcharakteristik
		S	Statecharakteristik (Änderungssensitivität)
		R	Reliabilität der Traitwerte

Die Ergebnisse sind angesichts der Befunde auf Itemebene nicht überraschend, setzten sich die Skalen doch aus den bereits analysierten Einzelitems zusammen. So verändern sich die mittleren intraindividuellen Mittelwerte wiederum in symptomatischer Hinsicht, d.h. bei allen in euphorischer Richtung zählenden Skalen verringern sie sich von der ersten zur zweiten Messung, nur bei der Skala U5 vergrößern sie sich. All diese Unterschiede sind hochsignifikant (t-Test für korrelierte Stichproben), sowohl auf der Basis aller 2669 gültigen Einzelmessungen als auch auf der Basis der Verteilungen der intraindividuellen Mittelwerte. Die interindividuellen Varianzen der

intraindividuellen Mittelwerte nehmen genauso wie die Varianzen der Traitwerte und die Traitcharakteristiken von der ersten zur zweiten Messung ohne Ausnahme zu, während die Statecharakteristiken entsprechend abnehmen. Diese steigen jedoch innerhalb einer Messung mit der Anzahl der pro Skala zusammengefaßten Items an. Sie ist in beiden Messungen für Skala B10 am höchsten, für Skala U2 am niedrigsten; lediglich Skala U3 weicht in der zweiten Messung von dieser Rangreihe ab. In allen Fällen sind die Statecharakteristiken als Maßzahlen der Änderungssensitivität mit Werten ab S=.69 (S=.63) für die erste (zweite) Messung annehmbar hoch. Bei den mittleren intraindividuellen Varianzen, den Varianzen der State- und der Fehlerwerte zeigt sich ebenfalls eine Zunahme von der ersten zur zweiten Messung mit Ausnahme der Skala U2, bei der sich diese Parameter verringern. Auch das Trait-Reliabilitätsmaß R nimmt von der einen zur anderen Messung zu; hier macht jedoch die Skala U3 eine Ausnahme, bei der der Wert gleich bleibt. Die mittleren intraindividuellen Korrelationen als State-Retest-Reliabilitäten liegen auf Skalenebene im Schnitt höher als auf Itemebene und außerdem nehmen sie von der Tendenz her mit zunehmender Aggregatgröße (Skalenumfang) zu; beides sind Zeichen für die aufgrund der Aggregierung reduzierten Fehlerkomponenten.

Insgesamt ist die Befundlage auf Skalenebene der auf Itemebene sehr ähnlich: Die Mittelwerte auf der Traitebene verändern sich von einer Messung zur anderen symptomatisch in Richtung einer Abnahme der gehobenen und einer Zunahme der gedrückten Stimmung und außerdem ist die zweite Messung offensichtlich auch auf Skalenebene traitabhängiger als die erste Messung, was sich in einer Verringerung der Statecharakteristiken ausdrückt. Diese sind jedoch ebenso wie die intraindividuellen Korrelationen zwischen erster und zweiter Messung (Ausnahme hier: Skala U2) annehmbar hoch.

Da die Skalen wegen ihrer unterschiedlichen Itemanzahlen unterschiedlich große maximale Meßwertumfänge aufweisen und da die Betrachtung der Varianzen der intraindividuellen Mittelwerte im Zusammenhang mit den dazugehörigen mittleren Mittelwerten nur eine relativ abstrakte Vorstellung von der empirischen Ausnutzung des Skalenumfangs durch die Vpn gestattet, sind aus Gründen der besseren Anschaulichkeit in Tabelle 13 für die Ebene aller 2669 Einzelmessungen und für die Aggregatebene mit ihren 74 intraindividuellen Mittelwerten die empirisch höchsten und niedrigsten erreichten Werte mit dem dazugehörigen Range ausgewiesen.

Tab. 13 Ranges auf der Ebene von Einzelmessungen und auf Aggregatebene

	Einzelmessungen			Aggregatebene		
	Range	min. Wert	max. Wert	Range	min. Wert	max. Wert
1. Messung						
B10 (min. 10 / max. 40 Punkte)	29	11	40	10,09	26,14	36,22
B8 (min. 8 / max. 32 Punkte)	24	8	32	8,54	20,74	29,28
B7 (min. 7 / max. 28 Punkte)	20	8	28	6,51	20,88	27,38
U5 (min. 5 / max. 20 Punkte)	15	5	20	4,54	5,00	9,54
U3 (min. 3 / max. 12 Punkte)	9	3	12	6,11	4,08	10,19
U2 (min. 2 / max. 8 Punkte)	6	2	8	3,50	3,97	7,47
2. Messung						
B10 (min. 10 / max. 40 Punkte)	29	11	40	10,83	25,70	36,53
B8 (min. 8 / max. 32 Punkte)	23	9	32	8,82	20,54	29,36
B7 (min. 7 / max. 28 Punkte)	20	8	28	7,93	19,63	27,56
U5 (min. 5 / max. 20 Punkte)	14	5	19	5,18	5,00	10,18
U3 (min. 3 / max. 12 Punkte)	9	3	12	6,05	4,22	10,27
U2 (min. 2 / max. 8 Punkte)	6	2	8	3,97	3,65	7,62

Wie sich aus der Analyse der Einzelmessungen zeigt, werden praktisch alle Skalen in vollem Umfang ausgenutzt; es gibt nur hier und da ein paar Extremwerte, die nicht vorgekommen sind. Auf der Aggregatebene sieht das aber ganz anders aus, denn die hier repräsentierten Werte sind Mittelwerte, aus denen die intraindividuell variierenden Komponenten herausgenommen sind. Sie verteilen sich deshalb in einem enger umgrenzten Skalenbereich.

3.5.2.3. Reaktivitätseffekte auf Skalenebene

Die Befunde im letzten Kapitel zur Veränderung der Mittelwerte auf Skalenebene untermauern die Vermutung, daß zwischen erster und zweiter Messung der bereits auf Itemebene beschriebene Reaktivitätseffekt eingetreten ist (vgl. Kap. 3.4.2.2.). Zur Stützung dieser Vermutung wurden auch auf Skalenebene noch einmal alle intraindividuellen Mittelwertunterschiede zwischen erster und zweiter Messung je Vp und Skala auf Signifikanz getestet (t-Test für korrelierte Stichproben). Von den 444 durchgeführten Prüfungen dürfen bei Gültigkeit der Nullhypothese nur maximal 5%, also 22 t-Tests, signifikant werden, je Skala etwa drei bis vier. Ein Blick auf Tabelle 14 zeigt jedoch, daß 127 dieser t-Tests (entsprechend 28.6%) signifikant werden, davon 112 in symptomatischer und nur 15 in gegenläufiger Richtung.

Tab. 14 Häufigkeit und Richtung signifikanter t-Tests auf Skalenebene zwischen 1. und 2. Messung (Basis: 74 t-Tests je Skala) (P≤.05)

	Gesamt	2. Mittelwert ist	
		größer	kleiner
B10	29	4	25
B8	17	4	13
B7	26	-	26
U5	17	14	3
U3	14	4	10
U2	24	-	24
SUMME	127		

Bis auf die Skala U5 verringern sich die signifikant verschiedenen Mittelwerte tendenziell in allen Skalen von der ersten zur zweiten Messung, was einer Abnahme der euphorischen Gestimmtheit entspricht. Für die Skala U5 gilt umgekehrt das gleiche: Die Mittelwerte erhöhen sich in 14 Fällen zwischen den Messungen, was einer stärkeren Ausprägung der dysphorischen Stimmung anläßlich der zweiten Messung entspricht.

Auch ein möglicher Effekt der Veränderung in der mittleren Stimmungsausprägung über den gesamten Beobachtungszeitraum hinweg wurde noch einmal auf Skalenebene überprüft. Dazu wurden wiederum je Skala und Vp (getrennt für jede Messung) zwei intraindividuelle Mittelwerte gebildet, einer für die ersten und einer für die letzten 21 Protokolltermine. Die 74 intraindividuellen Mittelwerte aus der ersten Hälfte wurden je Skala gegen diejenigen aus der zweiten Hälfte getestet (t-Test für korrelierte Stichproben). Wie schon auf Itemebene weisen die empirischen Mittelwerte aus beiden Hälften auch auf Skalenebene nur so geringe Differenzen auf, daß sie mit hoher Wahrscheinlichkeit zufällig zustande gekommen sind.

Die Ergebnisse aus Kapitel 3.4.2.2. können also nur unterstrichen werden. Auf der Grundlage der hier angewandten Auswertungsverfahren läßt sich zwischen der ersten und der zweiten Messung ein deutlicher Reaktivitätseffekt nachweisen. Ein Reaktivitätseffekt als Mittelwertveränderung über die gesamte Dauer des Beobachtungszeitraums kann jedoch ausgeschlossen werden.

3.6. Testgütekriterien

Zusätzlich zu den Skalencharakteristiken, die sich aus dem zugrundegelegten testtheoretischen Modell ergeben, soll auch eine Abschätzung der wichtigsten klassischen Testgütekriterien (vgl. Kap. 2.1.1.2. und Kap. 2.2.3.) vorgenommen werden. Die konstruierten Skalen wurden daher auch im Hinblick auf ihre Objektivität, ihre Reliabilität und ihre (mögliche) Validität untersucht.

3.6.1. Objektivität

Die Frage nach der Objektivität läßt sich nicht differenziert im Hinblick auf die verschiedenen Itemkombinationen klären, sondern kann nur für die gesamte Untersuchung und damit für alle gebildeten Skalen zusammen beantwortet werden.

Die Bestimmung der Durchführungsobjektivität dieser Untersuchung, so wie sie angelegt wurde, ist nur schwer möglich. Denn nach LIENERT (1989, S. 13) betrifft sie "den Grad der Unabhängigkeit der Testergebnisse durch zufällige oder systematische Verhaltensvariationen des Untersuchers während der Testdurchführung, die ihrerseits zu Verhaltensvariationen des Pbn führt und dessen Ergebnis beeinflußt", womit sich diese Begriffsbestimmung mehr auf klassische Labortests bezieht, die unter Aufsicht des Untersuchungsleiters bearbeitet werden. Sie erscheint daher nicht so gut geeignet, die Verhältnisse in einer ESM-Studie wiederzugeben. Bezogen auf die hier durchgeführte Untersuchung läßt sie sich lediglich für die Testeinweisung und Vorbefragung sowie für die Nachbefragung anwenden, nicht aber für die eigentliche Felduntersuchung. Hinsichtlich der

Testeinweisung und Vorbefragung und der Nachbefragung muß dabei festgestellt werden, daß sie zwar immer nach dem gleichen Schema und daher praktisch standardisiert abliefen (vgl. Kap. 3.2.3.), insbesondere wurden alle Anweisungen zum Bearbeiten der Frage- und Protokollbögen in immer gleicher Form vorgelesen, daß aber andererseits die Instruktionen z.B. nicht durch das Abspielen eines vorher besprochenen Tonbands dargeboten wurden, wodurch infolge verschiedener Betonungen, Satzmelodien usw. bei jedem neuen Vorlesen stets auch neue semantische Aspekte in die Instruktion mit eingeflossen sein dürften. Bereits hier ist die Durchführungsobjektivität also als beeinträchtigt anzusehen.

Viel bedeutsamer ist aber der Umstand, daß die Vpn zwar alle gleichermaßen während der Testeinweisung in die Bearbeitung der Protokollbögen eingewiesen wurden, daß sie aber während des Beobachtungszeitraums ganz auf sich alleine gestellt waren und deshalb nicht sichergestellt werden konnte, ob sie die Instruktionen auch richtig verstanden hatten oder sie richtig umsetzen konnten. Es wurde in Kapitel 3.4.3.1. der Versuch einer Überprüfung der Qualität der Protokolldaten unternommen, doch fußen die Ergebnisse dort auf Selbstauskünften der Vpn zur Untersuchung, die ex post erhoben wurden, und nicht auf einer tatsächlichen, simultanen Kontrolle der Erhebungsbedingungen. Nur im Falle von auftretenden Schwierigkeiten oder Pannen sollten sich die Vpn an den Untersuchungsleiter wenden, der sich seinerseits während des Beobachtungszeitraums lediglich ein- bis zweimal bei den Vpn telefonisch meldete, vor allem, um nötigenfalls die Compliance anzumahnen oder die Motivation aufrecht zu erhalten, und nicht so sehr, um sich explizit von der Korrektheit der Protokollierungen zu überzeugen. Denn es liegt ja gerade im Wesen von ESM-Untersuchungen, daß sie eine alltagsnahe und insbesondere auch eine in eher (für die untersuchten Vpn) privaten Situationen stattfindende Datenerhebung ermöglichen sollen, in denen eine Kontrolle der korrekten Durchführung der Untersuchung nicht möglich ist. Und ob eine Schwierigkeit oder Panne während der Feldphase eingetreten war, die eine Konsultation des Untersuchungsleiters erforderlich gemacht hätte, oder nicht, entschieden auch allein die Vpn. Die einzige Möglichkeit, eine hohe Durchführungobjektivität zu gewährleisten, besteht in einer präzisen Instruierung der Vpn und in einem intensiven Vpn-Training (wie z.B. bei PAWLIK & BUSE, 1982). Lediglich zu ihrer nachträglichen Abschätzung kann, wie hier geschehen, eine Nachbefragung der Vpn dienen. Das Training beschränkte sich in diesem Falle aber auf das Bearbeiten von zwei Protokollbögen während der Testeinweisung und Vorbefragung sowie auf eine Einweisung in die Bedienung des Signalgebers.

Die Auswertungsobjektivität ist, was die Protokollbögen anbelangt, praktisch völlig verwirklicht. Das Testverhalten der Vpn bestand darin, Kreuze in den Protokollbögen zu machen. Diese konnten anschließend während der Auswertung eindeutig in Zahlenwerte umgewandelt und so erfaßt und verrechnet werden. War diese Eindeutigkeit nicht gegeben (z.B. wenn ein Item zwei Kreuze erhalten hatte), so wurde die Angabe als fehlender Wert behandelt.

Bei den hier vorgestellten Skalen handelt es sich um erste Vorschläge auf der Basis eines recht kleinen Itempools. Die Frage nach der Interpretationsobjektivität kann daher zu diesem frühen Zeitpunkt in der Konstruktionsphase noch nicht sinnvoll gestellt werden, weil noch gar keine Regeln zur Interpretation der aus der Administration des Meßinstruments gewonnenen Daten aufgestellt

werden können. Denn noch sind Fragen etwa nach dem Vertrauensintervall der Meßwerte oder nach Möglichkeiten zur Normierung eines solchen Instruments nicht geklärt und vor allem ist noch völlig unklar, ob die hier entwickelten Skalen für irgendwelche Konstrukte valide sind, ob der Itempool vielleicht vergrößert und noch andere Items in die Skalen mit aufgenommen werden müssen usw. Grundsätzlich bleibt jedoch festzustellen, daß Ergebnisse aus einem für Zwecke der Feldpsychodiagnostik entwickelten Meßinstrument wie dem hier entworfenen durchaus eine hohe Interpretationsobjektivität erreichen können, sofern u.a. die genannten Grundbedingungen (Vertrauensintervalle, Normierung, Validität usw.) geklärt sind.

3.6.2. Reliabilität

Zur Bestimmung der Reliabilität der Skalen kann zunächst auf zwei bereits vorgestellte modellbezogene Parameter zurückgegriffen werden: Die Tabelle 12 weist zum einen die mittlere intraindividuelle Korrelation zwischen erster und zweiter Messung $\bar{r}_{ii'}$ aus, die schon mehrfach als intraindividuelle State-Retest-Reliabilität angesprochen worden ist. Zum anderen kann aus ihr die traitbezogene Reliabilität R entnommen werden, die nach (55) das Verhältnis von wahrer Traitvarianz zur Summe aus wahrer Traitvarianz plus Fehlervarianz unter Eliminierung der wahren Statevarianz darstellt. Zusätzlich zu diesen Maßen wurden auf der State- und auf der Traitebene noch je ein weiterer Reliabilitätskennwert sowie die Standardmeßfehler und die Vertrauensintervalle der Testwerte bestimmt.

Die mittlere intraindividuelle Korrelation der Skalenwerte zwischen erster und zweiter Messung wurde zusammen mit ihrer dazugehörigen mittleren intraindividuellen Varianz zur Berechnung der Vertrauensintervalle der einzelnen Meßwerte herangezogen. Zunächst mußte dafür entsprechend (32) durch Radizieren von (52) der Standardmeßfehler der Einzelmeßwerte SMF_s bestimmt werden:

$$(58) \quad SMF_s = s_e = \bar{s}_i * \sqrt{1 - \bar{r}_{ii'}}$$

Gemäß (33) wurde dann das dazugehörige Vertrauensintervall CL errechnet. Dieses Vertrauensintervall gibt an, in welchem Bereich um den tatsächlich erhobenen einzelnen Meßwert herum der wahre Statewert mit einer Irrtumswahrscheinlichkeit von 5% zu finden ist. Zu beachten ist wiederum, daß diese Größen auf den über alle Vpn gemittelten Korrelationen und Varianzen basieren und daher auch nur mittlere Schätzungen der Standardmeßfehler und der Vertrauensintervalle ermöglichen. Es ist aber ohne weiteres möglich, diese Parameter auch für eine einzelne Vp zu berechnen, wenn statt der mittleren Korrelationen und Varianzen die jeweils entsprechenden individuellen Werte in (58) eingesetzt werden. Zusätzlich wurde als eine weitere Maßzahl der Statereliabilität gemäß (24b) je Skala und Vp ein intraindividuelles Konsistenzmaß α gebildet.

Zur Reliabilitätsprüfung auf der Traitebene wurde zunächst die Split-Half-Reliabilität der intraindividuellen Mittelwerte nach der Odd-Even-Methode bestimmt. Dafür wurden für jede Skala

und jede Vp zwei intraindividuelle Mittelwerte errechnet, einer basierend auf allen von ihr wahrgenommenen ungeradzahligen Protokollterminen und einer basierend auf allen geradzahligen Terminen. Diese beiden intraindividuellen Mittelwerte wurden je Skala über alle Vpn miteinander korreliert. Die Odd-Mittelwerte basierten dabei für alle Vpn zusammen auf insgesamt 1356 Protokollterminen, die Even-Mittelwerte auf 1313 Protokollterminen. Der so erhaltene interindividuelle Korrelationskoeffizient wurde entsprechend (24a) nach der SPEARMAN-BROWN-Formel aufgewertet. Ausgehend von dieser Split-Half-Reliabilität R_{tt} und der Varianz der intraindividuellen Mittelwerte $s^2_{\bar{x}}$ wurde der Standardmeßfehler der Traitwerte SMF_t ermittelt und das Vertrauensintervall der intraindividuellen Mittelwerte bestimmt. Der Standardmeßfehler errechnet sich dabei wie folgt:

$$(59) \quad SMF_t = s_{\bar{x}} * \sqrt{1 - R_{tt}}$$

Die Berechnung des Vertrauensintervalls erfolgt wieder nach (33).

Die Tabelle 15 listet die hier besprochenen Kennwerte für alle Skalen auf. Neben den bereits aus Tabelle 12 bekannten modellbezogenen Koeffizienten $\bar{r}_{ii'}$ und R werden auf der State- und der Traitebene die errechneten Standardmeßfehler mit dazugehörigen Vertrauensintervallen aufgeführt. Die als Vertrauensintervalle angegebenen Werte geben dabei den Meßwertbereich auf jeder der beiden Seiten des empirischen Meßwerts an, innerhalb dessen der wahre Wert mit 95%iger Sicherheit zu finden ist. Auf der Stateebene sind zusätzlich die interindividuell bestimmten Mediane der 74 je Skala ermittelten intraindividuellen Konsistenzkoeffizienten angegeben und auf der Traitseite außerdem noch die Split-Half-Reliabilitäten.

Tab. 15 Reliabilitätskennwerte für einzelne Itemkombinationen (Skalen)

	State				Trait			
	$\bar{r}_{ii'}$	SMF_s	CL (P≤.05)	Mdn α	R	(odd / even) R_{tt}	SMF_t	CL (P≤.05)
1. Messung								
B10 (min. 10 / max. 40 Punkte)	0,910	1,288	±2,524	0,838	0,77	0,946	0,575	±1,126
B8 (min. 8 / max. 32 Punkte)	0,911	1,031	±2,020	0,820	0,78	0,947	0,466	±0,913
B7 (min. 7 / max. 28 Punkte)	0,851	1,050	±2,059	0,752	0,69	0,949	0,370	±0,725
U5 (min. 5 / max. 20 Punkte)	0,844	0,748	±1,466	0,726	0,70	0,954	0,255	±0,500
U3 (min. 3 / max. 12 Punkte)	0,892	0,679	±1,332	0,885	0,78	0,950	0,295	±0,578
U2 (min. 2 / max. 8 Punkte)	0,730	0,614	±1,204	0,566	0,55	0,934	0,182	±0,357
2. Messung								
B10 (min. 10 / max. 40 Punkte)	0,910	1,337	±2,621	0,859	0,80	0,960	0,548	±1,075
B8 (min. 8 / max. 32 Punkte)	0,911	1,075	±2,106	0,835	0,80	0,957	0,463	±0,907
B7 (min. 7 / max. 28 Punkte)	0,851	1,095	±2,146	0,784	0,73	0,968	0,334	±0,654
U5 (min. 5 / max. 20 Punkte)	0,844	0,784	±1,537	0,752	0,74	0,967	0,249	±0,488
U3 (min. 3 / max. 12 Punkte)	0,892	0,696	±1,364	0,899	0,78	0,957	0,285	±0,558
U2 (min. 2 / max. 8 Punkte)	0,730	0,608	±1,191	0,611	0,61	0,966	0,144	±0,283

State		Trait	
$\bar{r}_{ii'}$	mittlere intraindividuelle Korrelation	R	Reliabilität der Traitwerte
SMF_s	Standardmeßfehler (Basis: $\bar{r}_{ii'}$)	(odd/even) R_{tt}	Split-Half-Reliabilität der Testwerte
CL (P≤.05)	95% - Vertrauensintervall	SMF_t	Standardmeßfehler (Basis: R_{tt})
Mdn α	Median der (N = 74) intraindividuellen Konsistenzkoeffizienten	CL (P≤.05)	95% - Vertrauensintervall

Betrachtet man auf der Stateebene zunächst die Standardmeßfehler, so zeigen sie die gleichen Eigenheiten wie die Varianzen der Fehlerwerte s^2_e in Tabelle 12, aus denen sie durch Radizieren gewonnen wurden: Sie nehmen kontinuierlich innerhalb einer Messung mit sinkendem Skalenumfang ab (Ausnahme: B7) und steigen von der ersten zur zweiten Messung an (Ausnahme: U2). Für die Vertrauensintervalle gilt genau das gleiche, denn sie sind durch lineare Transformation aus den Standardmeßfehlern erzeugt worden. Die Vertrauensintervalle sind alle verhältnismäßig klein, bedenkt man, daß die Meßwerte in den dazugehörigen Skalen fast über die ganze Skalenbreite streuen (vgl. Tabelle 13). Vertrauensintervalle sollten generell möglichst klein sein, damit der wahre Wert einer Vp bei Kenntnis ihres Meßwertes so genau wie möglich und vor allem in Abgrenzung von den wahren Werten anderer Vpn geschätzt werden kann. Anläßlich der zweiten Messung ist deshalb durch die vergrößerten Vertrauensintervalle nur eine ungenauere Schätzung der wahren Werte im Vergleich zur ersten Messung gegeben (Ausnahme: U2). Hinsichtlich der durchschnittlichen intraindividuellen Konsistenzen der Skalen (Mediane) sind drei Charakteristika dieser Parameter hervorzuheben. Zum einen sind sie niedriger als die dazugehörigen intraindividuellen Korrelationen (State-Retest-Reliabilitäten) (Ausnahme: U3, zweite Messung), zum anderen nehmen sie mit abnehmendem Skalenumfang ab (Ausnahme: U3). Schließlich nehmen sie von der ersten zur zweiten Messung zu, was bedeutet, daß die Items anläßlich der zweiten Messung höher miteinander korrelieren als bei der ersten Messung.

Eine Gegenüberstellung der statefreien Reliabilität R und der Split-Half-Reliabilität R_{tt} auf der Traitseite zeigt, daß die Split-Half-Koeffizienten mit Werten bis zu R_{tt}=.968 erheblich höher liegen

als die modellbezogenen Parameter R. Letztere stellen relativ abstrakte Schätzgrößen für die statefreie Reliabilitätsbestimmung der Traitwerte zu einem einzelnen Meßzeitpunkt dar, wohingegen die Split-Half-Koeffizienten durch Korrelation von aggregierten Meßwerten errechnet und nach SPEARMAN-BROWN aufgewertet wurden; diese Split-Half-Koeffizienten nehmen alle von der ersten zur zweiten Messung zu. Die Standardmeßfehler auf der Traitebene werden dagegen zwischen den Messungen geringer und nehmen innerhalb der Messungen mit geringer werdendem Skalenumfang ebenfalls ab (Ausnahme: U3). Entsprechendes gilt auch hier für die dazugehörigen Vertrauensintervalle. Diese Trait-Vertrauensintervalle sind (z.T. erheblich) kleiner als die auf der Stateebene, d.h. der wahre Traitwert einer Vp ist auf der Skala in einem kleineren Streubereich um ihren intraindividuellen Mittelwert herum zu suchen als der wahre Statewert um den Einzelmeßwert herum. Dieses Vertrauensintervall muß allerdings in Relation zum Range der Traitwerte gesehen werden (vgl. Tabelle 13). Die Meßwertbereiche umfassen zwar je Skala auf der State- wie auf der Traitebene gleich viele Punktwerte, doch ist der Range der Traitwerte wesentlich kleiner als der der Statewerte. Es zeigt sich, daß die Vertrauensintervalle auf der Traitebene ungefähr 10% der dazugehörigen Variationsweite der Mittelwerte umfassen, was deutlich mehr ist als auf der Stateebene. Andererseits muß bei Betrachtung des Ranges der Einzelmeßwerte beachtet werden, daß sie sich auf das Gesamt aller Protokolltermine und aller Vpn beziehen. Der mittlere intraindividuelle Range (der der hier zur Spannweitenbeschreibung geeignetere Parameter wäre) dürfte erheblich kleiner sein. Deutlich wird in jedem Falle, daß die Vertrauensintervalle auf der Stateebene von der ersten zur zweiten Messung in fast allen Skalen größer werden, mithin eine ungenauere Schätzung des wahren Statewertes anläßlich der zweiten Messung gestatten, während die Vertrauensintervalle auf der Traitebene zugleich kleiner werden mit dem umgekehrten Effekt bezüglich der Schätzgenauigkeit. Die erste Messung gestattet also genauere Angaben zur Höhe der Statewerte, die zweite Messung sagt die Traitwerte besser voraus.

3.6.3. Validität

Das Problem der Validität der hier konstruierten Skalen kann im Rahmen dieser Arbeit nur gestreift werden, denn es geht bei ihrer Entwicklung zunächst um die Bereitstellung eines Verfahrens zur Informationsgewinnung über ein oder mehrere psychologische Konstrukte im Sinne eines Forschungsinstruments und nicht primär um die Entwicklung eines Tests, aufgrund dessen vielleicht für einzelne Vpn wichtige Entscheidungen getroffen werden könnten und für dessen Validierung ein unter pragmatischen Gesichtspunkten relevant erscheinendes Kriterium herangezogen werden könnte. Ein Instrument zur Stimmungsmessung im Feld muß seine Eignung also vor allem im Sinne einer Konstruktvalidierung beweisen, die sowohl theoretisch wie untersuchungstechnisch höchste Ansprüche stellt und zudem die Vorhersage eines konkret interessierenden einzelnen Kriteriums bereits beinhalten müßte (vgl. Kap. 2.1.1.2.). Da eine Konstruktvalidierung an dieser Stelle auch nicht ansatzweise durchgeführt werden kann und kein aus konkreten Zusammenhängen begründbarer diagnostischer Informations- oder Handlungsbedarf besteht, beschränkt sich die Validierung der Skalen auf ihren korrelativen Abgleich mit eher akzidentiell angefallenen Daten aus anderen, etablierten Meßinstrumenten, nämlich der PRF, dem STAI und dem SVF. Es soll hier auch gar nicht der Versuch unternommen werden, auf theoretischer Ebene einen vertieften Bezug zwischen den mit diesen Verfahren gemessenen Konstrukten und den Konstrukten der gehobenen oder gedrückten

Stimmung, wie sie durch die Protokollbögen erfaßt werden sollten, herzustellen.

Festgehalten werden kann dennoch, daß es zwischen dem STAI-Stateteil (STAI-S18), der zur Übereinstimmungsvalidierung der Statewerte herangezogen werden soll, und den zehn relevanten Items gewisse Bezüge auf formaler wie auf semantischer Ebene gibt. So haben einige Items des STAI-S18 ähnliche Inhalte wie die relevanten Items (z.B. STAI-S18 Items "bekümmert" und "besorgt" können in semantische Nähe zu "sorgenvoll" gebracht werden oder STAI-S18 "selbstsicher" kann so etwas ähnliches bedeuten wie "gewachsen"); weiter soll der STAI-Stateteil die momentane Ausprägung von Zustandsangst messen und eines der zehn relevanten Items lautet "Ich bin ängstlich"; schließlich tauchen zwei Items des originalen STAI-Stateteils ("froh" und "vergnügt") in der Liste der relevanten Items und damit in einigen der hier konstruierten Skalen auf (allerdings nicht gleichzeitig auch im STAI-S18). Darüber hinaus erhebt das STAI-S18 Daten ähnlicher Qualität wie die relevanten Items, nämlich solche zum Aktualerleben der Vpn, d.h. zu den momentanen Ausprägungen ihrer Befindlichkeiten, zudem noch mit derselben Antwortskala, im selben Layout und mit derselben Instruktion. Es darf also zwischen Meßwerten aus beiden Bereichen mit einer hohen methodisch bedingten Kovarianz gerechnet werden. Andererseits ist zu vermuten, daß die konstruierten Skalen und das STAI-S18 tatsächlich auch verwandte Konstrukte messen.

Wenn die Items des STAI-S18 und die der Skalen verwandte Konstrukte messen, so könnte es auch auf der Traitebene zwischen den dort erhaltenen Skalenmittelwerten und den STAI-Traitwerten eine Beziehung geben. Wenn die über den Beobachtungszeitraum variierende Stateangst der Vpn mit Gefühlen der Besorgtheit und der Befürchtung (und deshalb vielleicht auch mit einer gedrückten Stimmungslage oder einer Dämpfung der euphorischen Stimmung) einhergeht und diese Gefühle durch solche Reizkonstellationen in den Protokollsituationen ausgelöst werden, die die Vpn subjektiv als bedrohlich erleben und die dadurch in ihnen Angst als Reaktion auf diesen erlebten Streßcharakter der Situation auslösen (LAUX, 1983, SPIELBERGER, 1972), so könnten Beziehungen zwischen den Traitwerten aus den Protokollbögen und Traitmaßen der Streßbewältigung bestehen, wie sie mit dem SVF erhoben wurden. Bei der Analyse der so ermittelten Korrelationen muß jedoch berücksichtigt werden, daß die Traitmaße auf der Basis der Protokollbögen und die auf der Basis der Persönlichkeitsfragebögen (einschließlich PRF) unterschiedliche Datenqualitäten repräsentieren, dem Umstand, daß beides Papier- und Bleistiftverfahren sind, zum Trotz. Es wird hier also keine so hohe gemeinsame Methodenvarianz zum Tragen kommen, denn die Traitdaten der Protokollbögen sind durch Mittelung von Aktualerleben, d.h. zahlreicher, tatsächlich protokollierter Zustandsbefindlichkeiten, errechnet worden, während die Traitmaße aus den Fragebögen die durchschnittlichen Verhaltens- und Erlebensweisen der Vpn so widerspiegeln, wie sie in den Vpn mental repräsentiert sind.

Als erstes sollen die Beziehungen zwischen STAI-S18 und den Skalenwerten auf der Stateebene untersucht werden. Trotz der oben vorgebrachten Bedenken hinsichtlich einer gemeinsamen Methodenvarianz beider Maße stellt diese Übereinstimmungsvalidierung der Statewerte den substanzreichsten Ansatz dar. Dazu wurden für alle Vpn und alle Protokolltermine die STAI-S18 Summenwerte errechnet (nach Umpolung der in Richtung Angstfreiheit formulierten Items) und für jede der zwölf Skalen alle 74 intraindividuellen Korrelationskoeffizienten r_{tc} mit den STAI-S18

Summenwerten bestimmt und nach (56) in FISHER's Z-Werte transformiert. Diese Werte wurden pro Skala gemittelt und in mittlere Korrelationskoeffizienten zurückgerechnet. Zur Bestimmung einer Validität im engeren Sinne als minderungskorrigierte Korrelation zwischen den Skalenwerten und den wahren STAI-S18 Werten (vgl. Kap. 2.1.1.2.) wurde zusätzlich die Reliabilität des Kriteriums R_C , d.h. des STAI-S18 Wertes, entsprechend (38) berücksichtigt. Sie wurde als mittlere intraindividuelle Split-Half-Reliabilität nach der Odd-Even-Methode bestimmt, die zusätzlich nach SPEARMAN-BROWN gemäß (24a) aufgewertet wurde. Konkret wurde dabei für jede Vp ein intra-individueller STAI-S18 Split-Half-Koeffizient errechnet, der in einen FISHER's Z-Wert umgerechnet wurde. Der Mittelwert dieser 74 intraindividuellen Z-Werte wurde dann in einen Korrelationskoeffizienten zurückgerechnet, der wiederum nach SPEARMAN-BROWN aufgewertet wurde (mittlere intraindividuelle Split-Half-Reliabilität der STAI-S18 Werte nach Aufwertung: .920). Anschließend wurde die Wurzel dieses Split-Half-Koeffizienten zur Minderungskorrektur der Validitätskoeffizienten verwendet. Tabelle 16 listet neben den mittleren empirischen Korrelationen $\bar{r}_{tc}$ die so gewonnenen, aufgewerteten mittleren Korrelationen $\bar{r}_{tc'}$ zwischen den Skalen und den STAI-S18 Werten auf.

Tab. 16 Mittlere Korrelationen zwischen einzelnen Skalen und den STAI-S18 - Werten (Stateebene)

		$\bar{r}_{tc}$	$\bar{r}_{tc'}$
1. Messung	**B10** (min. 10 / max. 40 Punkte)	-0,777	-0,810
	B8 (min. 8 / max. 32 Punkte)	-0,738	-0,769
	B7 (min. 7 / max. 28 Punkte)	-0,771	-0,804
	U5 (min. 5 / max. 20 Punkte)	0,700	0,730
	U3 (min. 3 / max. 12 Punkte)	-0,613	-0,639
	U2 (min. 2 / max. 8 Punkte)	-0,653	-0,681
2. Messung	**B10** (min. 10 / max. 40 Punkte)	-0,806	-0,840
	B8 (min. 8 / max. 32 Punkte)	-0,771	-0,804
	B7 (min. 7 / max. 28 Punkte)	-0,794	-0,828
	U5 (min. 5 / max. 20 Punkte)	0,723	0,754
	U3 (min. 3 / max. 12 Punkte)	-0,655	-0,683
	U2 (min. 2 / max. 8 Punkte)	-0,695	-0,725

$\bar{r}_{tc}$ mittlere empirische Korrelation

$\bar{r}_{tc'}$ mittlere minderungskorrigierte Korrelation

Die Koeffizienten sind alle negativ, bis auf den der Skala U5. Dies macht Sinn, denn alle Skalen mit Ausnahme von U5 zählen in Richtung gehobene Stimmung und bei einem Anstieg der Stateangst ist ein Abfall in der gehobenen Stimmung bzw. ein Zuwachs der Ausprägung der gedrückten Stimmung zu erwarten. Die Koeffizienten steigen alle vom Betrag her von der ersten zur zweiten Messung an, was ebenfalls plausibel erscheint, wenn man den Reaktivitätseffekt bedenkt, der zwischen erster und zweiter Messung, d.h. während des Bearbeitens der STAI-S18 Items, auftritt (vgl. Kap. 3.4.2.1. und Kap. 3.4.2.2.). Wie bereits erwähnt, scheint es nämlich so zu sein, daß die erste Messung eher die Stimmung in der Alltagssituation widerspiegelt, die zweite Messung dagegen mehr die Stimmung während der Erhebungssituation. Die erstere variiert dabei stärker mit dem Alltagsgeschehen, die letztere dagegen mehr mit den Protokollsituationen, die sich aber alle untereinander stark ähneln. Die

Stimmungslage während des Protokollierens ist dann auch diejenige, während derer sowohl das STAI-S18 als auch der zweite Itemblock bearbeitet werden. Das bedeutet, daß zur Validierung der ersten Messung besser ein Kriterium herangezogen werden sollte, das der Stimmung unmittelbar vor Beginn der Protokollierung zu entsprechen hätte, und daß das STAI-S18 im Grunde nur zur Validierung der zweiten Messung verwendet werden darf.

Auf der Traitebene wurden die intraindividuellen Mittelwerte der konstruierten Befindlichkeitsskalen mit den Merkmalsskalen von PRF, STAI-Traitteil und SVF korreliert, wobei im Falle von PRF und STAI mit den nicht-normierten Rohwerten gearbeitet wurde. (Der SVF ist ohnehin nicht normiert.) Tabelle 17 zeigt die Korrelationskoeffizienten für die PRF und den STAI-Traitteil; angegeben sind nur solche Koeffizienten, die sich in der Grundgesamtheit hochsignifikant von null unterscheiden.

Tab. 17 Korrelationen zwischen den gebildeten Skalen (intraindividuelle Mittelwerte) und den PRF-Merkmalsskalen (Rohwerte) sowie dem STAI-Traitteil (Rohwert)
(angegeben sind nur Koeffizienten mit P≤.01) (Traitebene)

	1. Messung						2. Messung					
	B10	B8	B7	U5	U3	U2	B10	B8	B7	U5	U3	U2
PRF												
Leistungsstreben												
Geselligkeit												
Aggressivität	-0,300		-0,477	0,443		-0,358	-0,346	-0,285	-0,484	0,430		-0,401
Dominanzstreben												
Ausdauer												
Bedürfnis nach Beachtung												
Risikomeidung												
Impulsivität			-0,301	0,291					-0,296	0,278		
Hilfsbereitschaft												
Ordnungsstreben												
Spielerische Grundhaltung												
Soziales Anerkennungsbedürfnis												
Anlehnungsbedürfnis			-0,312	0,295					-0,283			
Allgemeine Interessiertheit												
STAI-Trait												
STAI-Traitwert	-0,422	-0,405	-0,498	0,499		-0,314	-0,440	-0,420	-0,502	0,486		-0,346

Von den Merkmalsskalen der PRF zeigt die Aggressivität den deutlichsten Zusammenhang zu den mittleren Befindlichkeitsskalenwerten aus den Protokollbögen. Dieses Merkmal weist zu den meisten Skalen signifikante Beziehungen auf, die außerdem mit Varianzaufklärungen von z.T. über 20% auch noch recht hoch sind (im Falle von B7). Daneben gibt es Korrelationen mit den PRF-Merkmalen Impulsivität und Anlehnungsbedürfnis, die aber eher als unbedeutend einzuordnen sind. Zusammenhänge bis zu r=-.502 (für B7, zweite Messung) finden sich aber zwischen dem STAI-

Traitteil und allen Skalen mit Ausnahme von U3 (beide Messungen), was dafür spricht, daß es zwischen den Konstrukten der gehobenen bzw. gedrückten Stimmung und der Traitangst eine substantielle Beziehung gibt.

Ein Blick auf die Tabelle 18 zeigt die signifikanten Korrelationen zwischen den Befindlichkeitsskalen und den Streßbewältigungsstrategien, wie sie durch den SVF operationalisiert sind.

Tab. 18 Korrelationen zwischen den gebildeten Skalen (intraindividuelle Mittelwerte) und den SVF-Merkmalsskalen (Rohwerte) (angegeben sind nur Koeffizienten mit P≤.01) (Traitebene)

	1. Messung						2. Messung					
SVF	B10	B8	B7	U5	U3	U2	B10	B8	B7	U5	U3	U2
Bagatellisierung												
Herunterspielen durch Vergleich mit anderen			0,320	-0,277		0,274	0,299		0,336			0,356
Schuldabwehr												
Ablenkung von Situationen												
Ersatz-befriedigung												
Suche nach Selbstbestätigung					0,304						0,314	
Situationskontroll-versuche												
Reaktionskontroll-versuche						0,279						0,298
Positive Selbstinstruktion												
Bedürfnis nach sozialer Unter-stützung												
Vermeidungs-tendenz												
Fluchttendenz												
Soziale Abkapselung												
Gedankliche Weiterbe-schäftigung				0,313					-0,292	0,350		
Resignation			-0,309	0,343					-0,335	0,356		
Selbst-bemitleidung				0,276					-0,279	0,308		
Selbst-beschuldigung												
Aggression			-0,280						-0,307	0,276		
Pharmaka-einnahme												

Die hier auftretenden Korrelationen sind vom Betrag her eher unerheblich, auch wenn sie höchstwahrscheinlich von null verschieden sind, denn die höchsten von ihnen erklären gerade knapp 13% gegenseitig determinierter Varianz. Es wird aber deutlich, daß es die Befindlichkeitsskalen B7

und U5 sind, durch die sich die meisten Zusammenhänge zwischen den relevanten Items und den erhobenen Traitmaßen herstellen lassen, denn nicht nur hier, sondern auch in Tabelle 17 weisen sie zahlenmäßig die meisten Korrelationen und hinsichtlich des STAI-Traitmaßes (Tabelle 17) auch die höchsten Koeffizienten auf.

4. Diskussion

In diesem Abschnitt sollen die in den Kapiteln 3.4. bis 3.6. referierten Untersuchungsergebnisse und daraus abgeleitete Maßnahmen zusammenfassend analysiert und bewertet werden. Dies beinhaltet im Rahmen einer explorativen Arbeit auch das Aufzeigen von Verbesserungsmöglichkeiten und einen Ausblick auf mögliche weitere Schritte zur Entwicklung einer Endform des hier konzipierten Meßinstruments.

Zunächst einmal kann festgestellt werden, daß die eingangs formulierten Zielvorgaben dieser Arbeit (vgl. Kap. 1.2.) erfüllt worden sind. So konnten die Modellparameter für das State-Trait-Modell von BUSE & PAWLIK (1991) erfolgreich und im Sinne einer grundsätzlichen Stützung dieses Modells empirisch ermittelt und dokumentiert werden. Auf der Basis dieser Parameter sowie aufgrund von Faktorenanalysen konnten die relevanten Befindlichkeitsitems aus den Protokollbögen in verschiedenen Kombinationen zu änderungssensitiven Skalen zusammengefaßt werden. Eine Abschätzung der klassischen Hauptgütekriterien für diese Skalen war in dem hier leistbaren Umfang z.T. recht genau möglich. Eine Analyse der Untersuchungsmethode unter den Aspekten der Sicherung der Datenqualität, der Abschätzung der ökopsychologischen Gütekriterien und der Methodenakzeptanz konnte ebenfalls durchgeführt werden. Auf die Einsatzmöglichkeiten eines so konstruierten Instruments und auf einige andere Details wird im folgenden jedoch noch näher einzugehen sein.

4.1. Bewertung der Ergebnisse

4.1.1. Reaktivitätseffekte

Eines der wichtigsten Ergebnisse der vorliegenden Untersuchung ist der Umstand, daß es während der Protokollierung offenkundig zu messungsbedingten Veränderungen in den erhobenen Befindlichkeitsmerkmalen gekommen ist, die sich auf Item- wie auf Skalenebene mitteilen. Dadurch muß die Merkmalskonstanz von der ersten zur zweiten Messung, welche eine zentrale Grundannahme dieser Untersuchung darstellt, als nicht völlig gegeben angesehen werden. Eine Verletzung der angenommenen Merkmalskonstanz hat Bedeutung sowohl für das State-Trait-Modell als auch im Hinblick auf den ESM-Charakter der Untersuchung.

Die im Rahmen des State-Trait-Modells errechneten intraindividuellen Korrelationen zwischen erster und zweiter Messung müssen von einer Konstanz in den wahren Merkmalen ausgehen, um als Reliabilitäten eingestuft werden zu können. Die Annahme eines bei allen Vpn gleichermaßen wirkenden Reaktivitätseffekts (im Sinne einer linearen Veränderung der zweiten Meßwerte), der die

Auswirkungen auf die intraindividuelle State-Retest-Reliabilität abschwächen könnte, ist zum einen unrealistisch, zum anderen durch das Datenmaterial nicht belegbar. Signifikante Mittelwertveränderungen treten nämlich nur bei manchen Vpn auf, nicht bei allen. Zudem sind bei einigen wenigen Vpn sogar Mittelwertveränderungen in gegenläufiger, nicht-symptomatischer Richtung aufgetreten. Der Reaktivitätseffekt hat sich also nicht bei allen, sondern nur bei gerade so vielen Vpn bemerkbar gemacht, daß ein zufälliges Auftreten solcher Mittelwertdifferenzen ausgeschlossen werden kann. Wenn überhaupt, dann tritt er allerdings vornehmlich in symptomatischer, d.h. in sehr plausibler, Richtung auf. Der Reaktivitätseffekt ist also nicht als eine gleichgeschaltete Einflußgröße im Sinne einer quasi-kontrollierten Störquelle anzusehen; über die Bedingungen seines konkreten Wirksamwerdens und seines Einflusses auf die intraindividuellen Korrelationen ist nichts bekannt. Die Beeinträchtigung der korrelativen Zusammenhänge ist aus der Zahl, Richtung oder Höhe aufgetretener Mittelwertunterschiede nicht abschätzbar. Da die intraindividuelle State-Retest-Reliabilität ein zentraler Baustein für die Bestimmung der Modellparameter ist, darf die Bedeutung ihrer möglichen Verzerrung durch Änderungen in den wahren Merkmalsausprägungen infolge eines Reaktivitätseffekts nicht unterschätzt werden. Zur Zeit dürfte diese spezielle Art von kurzfristig auftretendem Reaktivitätseffekt jedoch noch kaum beobachtet oder beschrieben worden sein, denn bisher sind wahrscheinlich nur wenige Untersuchungen durchgeführt worden, die eine doppelte Itemvorgabe oder Messung zu demselben Erhebungszeitpunkt erforderlich gemacht hätten.

Zugleich bedeutet das Auftreten eines solchen Reaktivitätseffekts auch eine Minderung des ESM-Charakters der Untersuchung, denn es kommt zu einer messungsbedingten Beeinflussung des zu untersuchenden Erlebens während des Protokollierens, was ausdrücklich unerwünscht ist. Jedoch muß hier gefragt werden, wie gering ein zweifellos stets auftretender Effekt der Messung selbst auf das zu messende Verhalten oder Erleben denn gehalten werden kann. Es ist naiv anzunehmen, daß ein Datenerhebungsverfahren, das eine aktive Beteiligung der Vpn erforderlich macht, ohne ein reaktives Moment stattfinden könnte (vgl. GACHOWETZ, 1987).

Methodenkritisch muß an dieser Stelle angemerkt werden, daß der Reaktivitätseffekt mit Hilfe eines t-Tests für korrelierte Stichproben unter Verwendung der Rohdaten untersucht wurde. Die Meßdaten in beiden Wertereihen (erste und zweite Messung) sind aber zuvor nicht unter Anwendung zeitreihenanalytischer Verfahren auf autokorrelative Abhängigkeiten oder gar auf Trends oder zyklische Komponenten untersucht worden. Die Eliminierung autokorrelativer Zusammenhänge zwischen den Meßwerten einer intraindividuellen Wertereihe ist jedoch erforderlich, um innerhalb dieser Reihe voneinander unabhängige Meßwerte zu erhalten ("White Noise") (vgl. dazu SCHMITZ, 1989, WEST & HEPWORTH, 1991). In der Mißachtung einer möglichen seriellen Abhängigkeit von Daten aus Zeitreihen bei der Anwendung von Signifikanztests steht diese Arbeit allerdings bei weitem nicht alleine da: "The most common approach in psychology has been to ignore the problem, effectively assuming that there is no serial dependency in the data." (WEST & HEPWORTH, 1991, S. 626)

Positiv bleibt bezüglich der Reaktivitätseffekte anzumerken, daß keine Veränderung der Meßwerte hinsichtlich ihrer zentralen Tendenz über die Beobachtungszeiträume stattgefunden hat, weder auf

Item- noch auf Skalenebene. Effekte wie der hier vermutete, die sich über den ganzen Erhebungszeitraum vom ersten bis zum letzten Protokolltermin auswirken, sind ansonsten in ESM-Studien oder anderen Untersuchungen, bei denen denselben Vpn das gleiche Stimulusmaterial zu verschiedenen Zeitpunkten wiederholt vorgegeben wurde, gelegentlich beobachtet oder befürchtet worden (vgl. z.B. CATTELL, 1967, CSIKSZENTMIHALYI & LARSON, 1987, PERREZ & REICHERTS, 1989, STONE, KESSLER & HAYTHORNTHWAITE, 1991, TENNEN, SULS & AFFLECK, 1991).

4.1.2. Modellparameter

Die Höhe der mittleren intraindividuellen Korrelationskoeffizienten zwischen erster und zweiter Messung ist für die hier untersuchten zehn Befindlichkeitsitems als zufriedenstellend einzustufen. Sie liegt im Mittel über alle zehn Items bei $\bar{r}_{ii'}$=.7 und erreicht damit eine Größenordnung, die für Ratingskalen als typisch bezeichnet werden kann (MURPHY & DAVIDSHOFER, 1991, S. 103). BUSE & PAWLIK (1991) berichten dagegen wesentlich niedrigere mittlere intraindividuelle Korrelationen für die von ihnen untersuchten Befindlichkeitsitems (dort: $\bar{r}_{ii'}$=.52). (Von den 17 dort vorgegebenen Items gehören vier auch zu dem Pool der hier administrierten Items.) Gleiches gilt für die State-Charakteristiken als den Änderungssensitivitäten dieser Items, die in der vorliegenden Untersuchung im Mittel bei $\bar{S}$=.7, bei BUSE & PAWLIK (1991) jedoch (auch nach Korrektur des Response Sets) nur bei $\bar{S}$=.47 liegen. Entsprechend hoch liegen dort die Trait-Charakteristiken. Die niedrigste Änderungssensitivität weist in beiden Untersuchungen übereinstimmend das Item "gewachsen" auf. Es muß allerdings berücksichtigt werden, daß BUSE & PAWLIK (1991) das von ihnen entwickelte State-Trait-Modell auf einen bereits vorhandenen, aus einer Datenrecordererhebung stammenden Felddatensatz angewandt haben, bei dem keine Retest-Reliabilitätsbestimmung mit denselben Items möglich war, weil jedes Item nur einmal vorgegeben worden war. Die Schätzung der intraindividuellen Korrelationen erfolgte dort mit Hilfe paralleler Items. Auch im Antwortformat, in der Zusammensetzung der Stichprobe und in den Instruktionen der Vpn unterscheiden sich die Untersuchungen von BUSE & PAWLIK (1991) und die hier durchgeführte.

Zusätzlich zu den Befindlichkeitsmessungen erhoben BUSE & PAWLIK (1991) noch einige andere Merkmale, u.a. die psychophysiologischen Parameter Pulsfrequenz und Körpertemperatur. Für diese wurden nicht nur zwei, sondern drei Messungen je Protokolltermin vorgenommen, was die exakte Berechnung von State- und Trait-Charakteristiken für jede dieser drei Messungen ermöglichte. Dabei stellt sich zunächst heraus, daß beide psychophysiologischen Merkmale in hohem Maße änderungssensitiv sind. Betrachtet man weiter die Veränderung der Modellparameter über alle drei Messungen, so zeigt sich bei der Pulsfrequenz, daß die State-Charakteristik von der ersten über die zweite bis zur dritten Messung abnimmt, und bei der Körpertemperatur, daß sie von der ersten zur zweiten Messung abnimmt, von der zweiten zur dritten Messung aber gleich bleibt. Dies bestätigt die in dieser Arbeit auf Item- wie auf Skalenebene gefundene Verringerung der State-Charakteristiken von der ersten zur zweiten Messung.

4.1.3. Skalenkonstruktion

Für die Präferenz einer der konstruierten Befindlichkeitsskalen können die gleichen formalen Kriterien wie für die Itemselektion herangezogen werden: Ausgehend vom zugrundegelegten State-Trait-Modell ist diejenige Itemkombination als geeignet auszuwählen, die die höchste mittlere intraindividuelle Korrelation zwischen erster und zweiter Messung sowie die höchste Änderungssensitivität in beiden Messungen aufweist, wobei die Trait-Charakteristik deutlich von null verschieden sein soll. Darüber hinaus sollen die klassischen Testgütekriterien in einer dem Standard entsprechenden Höhe erfüllt sein. Bedenkt man dann noch die sowohl vor dem Hintergrund der gesichteten Literatur wie auch im Hinblick auf das hier vorliegende Datenmaterial bereits ausgiebig geführte Diskussion zur Uni- oder Bipolarität von Befindlichkeitsdimensionen, so ist, bei Gleichheit der übrigen Parameter, einer bipolaren Skala prinzipiell der Vorzug zu geben.

Die höchsten durchschnittlichen intraindividuellen Korrelationen und zugleich die höchsten Änderungssensitivitäten weisen die Skalen B10 und B8 auf. Zugleich haben sie die niedrigsten Traitcharakteristiken, die aber doch deutlich von null verschieden sind. Auch die Skala U3 weist noch akzeptable Kennwerte auf, während die Skalen B7, U5 und besonders U2 demgegenüber abfallen.

Hinsichtlich der Objektivität unterscheiden sich die einzelnen Skalen nicht voneinander. Die bereits diskutierten Aspekte der Durchführungs-, Auswertungs- und Interpretationsobjektivität gelten für alle Skalen gleichermaßen, so daß diese nicht dazu herangezogen werden können, eine Auswahl unter ihnen zu treffen. Grundsätzlich kann ihnen jedoch eine hinlängliche Durchführungsobjektivität unterstellt werden, während die Auswertungsobjektivität sichergestellt ist. Die Reliabilitätskennwerte sollten möglichst hoch sein. Dabei sollen Werte von R=.6 überschritten werden, um eine akzeptable Reliabilität zu gewährleisten (MURPHY & DAVIDSHOFER, 1991, S. 103), wenngleich zur Beurteilung von Gruppendifferenzen laut LIENERT (1989, S. 309) auch schon Koeffizienten von R≥.5 ausreichend sind. Da die hier zur Diskussion stehenden Befindlichkeitsskalen aber grundsätzlich auch zur Beurteilung (intra- wie inter-) individueller Differenzen verwendbar sein sollen, müssen (auf State- wie auf Traitebene) die Koeffizienten eine Höhe von R≥.8 (ebd.) erreichen oder sogar in der Nähe von R=.9 liegen, was als hohe Reliabilität bezeichnet werden kann (MURPHY & DAVIDSHOFER, 1991, S. 103). Zugleich verlangt LIENERT (1989, S. 309), daß Koeffizienten der inneren Konsistenz eine Höhe von R≥.9 aufweisen sollen. Betrachtet man die ermittelten Reliabilitätskennwerte, so zeigt sich, daß auf der Traitebene die Split-Half-Koeffizienten für alle Skalen Werte zwischen min. .934 und max. .968 annehmen und damit zwar zufriedenstellend hoch sind, jedoch praktisch keine Differenzierung zwischen den Skalen ermöglichen. Die modellbestimmten, statefreien (Trait-) Reliabilitätskoeffizienten liegen niedriger und erreichen für die Skalen B10, B8 und U3 Werte von R=.77 bis R=.80 . Auf der Stateseite sind die intraindividuellen State-Retest-Reliabilitäten, wie schon erwähnt, für die Skalen B10 und B8 am höchsten, gefolgt von U3. Die Skala U3 weist die höchsten Mediane der Konsistenzkoeffizienten auf, liegt dabei aber mit Mdn α=.885 (erste Messung) bzw. Mdn α=.899 (zweite Messung) noch knapp unter der geforderten Mindesthöhe. Bezüglich der Validität soll hier nur die Übereinstimmungsvalidierung der Statewerte an den STAI-S18 Werten herangezogen werden. Die

Beträge der mittleren, nicht minderungskorrigierten Validitätskoeffizienten (Test-Kriteriumskorrelationen) erreichen dabei für alle Skalen bis auf U3 und U2 Werte von $\bar{r}_{tc} \geq .7$, was nach LIENERT (1989, S. 310) hinreichend ist. Die höchsten Koeffizienten weist Skala B10 auf, gefolgt von B7.

Zusammengenommen sprechen diese Befunde für eine Bevorzugung der Skala B10, die die höchsten Änderungssensitivitäten, die zweithöchste mittlere intraindividuelle Korrelation, sehr hohe (Trait-) Split-Half-Reliabilitäten, mit die höchsten modellbezogenen Reliabilitätskennwerte und die höchsten Validitätskoeffizienten auf Stateebene aufweist. Zudem hat sie nach Skala U3 die zweithöchsten mittleren Konsistenzkoeffizienten (Mediane), die allerdings insgesamt für alle Skalen recht niedrig liegen. Darüber hinaus ist sie bipolar und deckt damit die beiden zu messenden Konstrukte der euphorischen und dysphorischen Stimmung ab. Betrachtet man jedoch die diesen Überlegungen zugrundeliegenden Maßzahlen aus den Tabellen 12, 15 und 16 im Überblick, so erkennt man, daß der hier vorgenommenen Bevorzugung der Skala B10 eine gewisse Beliebigkeit anhaftet, denn sie besitzt, wie geschildert, durchaus nicht in allen Parametern die höchsten Werte. Oft sind die in den Tabellen ausgewiesenen Unterschiede zwischen den Skalen in diesen Parametern nämlich vom Betrag her so gering, daß ihnen auf der Interpretationsebene praktisch keine Bedeutung zukommt. Deshalb kann lediglich die Skala U2 als deutlich ungeeignet ausgesondert werden. Soll statt der beiden breiten Dimensionen, deren Messung hier intendiert ist, nur das Konstrukt "Freude" gemessen werden, so kann durchaus auf die (noch validierungsbedürftige) Skala U3 zurückgegriffen werden, die mit ihren drei Items ebenfalls sehr hohe Kennwerte erreicht.

4.1.4. Stichprobe und Untersuchungsmethode

Die untersuchte Gesamtstichprobe weist mit N=74 Vpn einen akzeptablen Umfang auf. Ein Blick in die Literatur zeigt, daß viele Untersucher auch mit geringeren Stichprobengrößen gearbeitet haben (z.B. BOHNER, HORMUTH & SCHWARZ, 1991, CANTOR et al., 1991, DIENER & LARSEN, 1984, McADAMS & CONSTANTIAN, 1983), wobei eine Anzahl von N=40 Vpn eine untere Grenze zu bilden scheint. Stichprobenumfänge von mehr als einhundert Vpn werden allerdings auch nur selten berichtet (z.B. CSIKSZENTMIHALYI & FIGURSKI, 1982, PAWLIK & BUSE, 1992, WONG & CSIKSZENTMIHALYI, 1991). Der sieben Tage umfassende Beobachtungszeitraum kann als für ESM-Studien typisch angesehen werden (CSIKSZENTMIHALYI & LARSON, 1987), nur selten gab es in anderen Untersuchungen bedeutend längere Beobachtungszeiträume (wie z.B. sechs Wochen bei DIENER, LARSEN & EMMONS, 1984). Mit sechs Protokollterminen pro Beobachtungstag liegt diese Untersuchung etwas unterhalb dessen, was aus anderen Studien berichtet wird. Üblich sind dort sieben bis zehn Termine (CSIKSZENTMIHALYI & LARSON, 1987; außerdem z.B. BOHNER, HORMUTH & SCHWARZ, 1991, BUSE & PAWLIK, 1984, McADAMS & CONSTANTIAN, 1983). Letztendlich hängt eine Entscheidung über die Anzahl täglicher Protokolltermine von der gewünschten zeitlichen Auflösungsfähigkeit der Untersuchung, damit vom untersuchten Merkmalsbereich sowie von der Belastbarkeit der Vpn ab.

Daß die hier untersuchten Psychologiestudenten und die Angehörigen meines Bekanntenkreises hinsichtlich der erhobenen, normierten Persönlichkeitsmerkmale für die Bevölkerung der Bundesrepublik nicht repräsentativ sind, ist bedauerlich, aber nicht überraschend. Immerhin konnte das Geschlechterverhältnis repräsentativ gehalten werden.

Die untersuchten Frauen unterscheiden sich in mehreren erhobenen Persönlichkeitsmerkmalen von den Männern. Hinsichtlich der SVF-Skalen muß bei der Interpretation dieser Unterschiede jedoch beachtet werden, daß bereits die Autoren des SVF in ihrer Handanweisung signifikante Geschlechtsunterschiede für mehrere Subskalen berichten, u.a. für Bedürfnis nach sozialer Unterstützung, Resignation und Aggression (JANKE, ERDMANN & BOUCSEIN, 1985, S. 28), in denen sich auch die Frauen der vorliegenden Untersuchung durch höhere Mittelwerte von den Männern unterscheiden. Auffällig ist dennoch, daß die Frauen nicht nur im SVF-Merkmal Aggression höher scoren, sondern ebenfalls in den normierten PRF-Skalen Aggressivität und Dominanzstreben. Es ist zu begrüßen, daß sich die Psychologiestudenten von den Nicht-Psychologiestudenten nur in einigen wenigen der erhobenen Persönlichkeitsmerkmale unterscheiden.

Die Qualität der erhobenen Protokolldaten kann als gut bis sehr gut eingestuft werden. Die Undurchschaubarkeit des Erhebungsinstruments und die Verschleierung des Untersuchungsziels konnten gewährleistet werden. Die Vpn haben sich alles in allem in hinreichendem Maße an die Instruktionen gehalten; das gilt im besonderen auch für die Einhaltung der vorgezogenen Rückgabepflicht der Protokollbögen durch die Psychologiestudenten, einer Maßnahme zur Sicherung der Compliance, wie sie auch schon von anderen Untersuchern getroffen worden ist (z.B. DIENER & LARSEN, 1984). Daß eine nicht unerhebliche Anzahl der Vpn Teile der Untersuchung als anstrengend oder störend empfanden oder sonst an ihr etwas auszusetzen hatten, ist grundsätzlich ein Indiz für eine mögliche Minderung der Datenqualität. Doch gibt es andererseits keine Anzeichen für das Wirksamwerden wirklich gravierender Störeffekte oder für eine umfangreiche Verletzung der Compliance, die die Verwertbarkeit der Protokolldaten ernsthaft in Frage stellen könnten.

In dieser Untersuchung können die ökopsychologischen Gütekriterien als verwirklicht angesehen werden. Für fast alle Vpn kann angenommen werden, daß sie Protokollierungen unter Alltagsbedingungen vorgenommen haben, wodurch das Kriterium der ökologischen Validität erfüllt ist. Wichtigste Indikatoren für diese Feststellung sind einerseits die Selbstauskünfte der Vpn auf die entsprechende Frage 1 des Nachbefragungsbogens sowie der Umstand, daß 84% aller Protokollierungen in gewohnter Umgebung erfolgten. Weiterhin waren die Vpn in 65% aller Protokollsituationen nicht allein, was sich mit Befunden von PAWLIK & BUSE (1992) deckt (dort: 62%). Die ökologische Repräsentativität der gezogenen Situationsstichproben kann ebenfalls als gegeben angesehen werden. Die mittlere Reaktionsrate von 85.9% aller vorgesehenen Protokolltermine liegt über dem in der Literatur berichteten Schnitt (vgl. Übersichten bei CSIKSZENTMIHALYI & LARSON, 1987, PAWLIK & BUSE, 1992; außerdem z.B. BOHNER, HORMUTH & SCHWARZ, 1991, CANTOR et al., 1991, CSIKSZENTMIHALYI & FIGURSKI, 1982, McADAMS & CONSTANTIAN, 1983, WONG & CSIKSZENTMIHALYI, 1991). Es trat kein Stichprobenschwund auf und die niedrigste Anzahl der je Vp wahrgenommenen Protokolltermine liegt mit 23 Terminen entsprechend 54.8% über dem, was in anderen

Untersuchungen noch als Minimum akzeptiert worden ist (z.B. McADAMS & CONSTANTIAN, 1983, WONG & CSIKSZENTMIHALYI, 1991). Die Häufigkeitsverteilung der Latenzzeiten macht einen sehr günstigen Eindruck. Sie ist den bei CSIKSZENTMIHALYI & LARSON (1987) oder bei HORMUTH (1986) mitgeteilten sehr ähnlich und scheint damit für ESM-Studien typisch zu sein.

Die Akzeptanz der Untersuchungsmethode im Alltag ist recht groß. Mehr als drei Viertel aller Vpn würden sich erneut für so eine Untersuchung zur Verfügung stellen, nur 18% lehnten dies klar ab, was sich mit dem Befund von HORMUTH (1986) deckt (dort: 75% Ja-Sager). Einige der von den Vpn vorgebrachten Veränderungswünsche hinsichtlich der Durchführung der Untersuchung und der Gestaltung der Protokollbögen können prinzipiell berücksichtigt werden. Auch die Akzeptanz des Signalgebers war trotz aller Kritikpunkte sehr hoch; zudem gab es fast keine Schwierigkeiten mit seiner technischen Handhabung im Feld, im Gegensatz zu den in der Untersuchung von CANTOR und Kollegen (CANTOR et al., 1991) eingesetzten Uhren oder auch im Gegensatz zu den von WONG & CSIKSZENTMIHALYI (1991) eingesetzten elektronischen Funksignalempfängern. In der hier durchgeführten Form ist diese Methode der Informationsgewinnung über das Stimmungserleben unter Alltagsbedingungen in psychologiestudentischen und nicht-psychologiestudentischen Populationen gut einsetzbar (eine Optimierung der Befindlichkeitsskalen und ihre weitere psychometrische Überprüfung vorausgesetzt). Die Nicht-Psychologiestudenten waren zum größten Teil allerdings auch Studenten (anderer Fächer), Hochschulabsolventen oder zumindest Abiturienten, darüber hinaus durch persönlichen Kontakt mit dem Untersuchungsleiter wahrscheinlich überdurchschnittlich hoch zur Teilnahme an der Untersuchung motiviert, so daß die Befunde zur Methodenakzeptanz nicht einfach auf alle Nicht-Psychologiestudenten generalisiert werden können. Es darf aber eine hohe Akzeptanz in solchen Fällen erwartet werden, in denen die Vpn selber ein großes Eigeninteresse an zuverlässigen Ergebnissen haben, wie z.B. im klinischen Bereich. So können Leidensdruck und Therapiemotivation das Engagement der Protokollierung mit Hilfe eines tragbaren Minicomputers bei Alkoholikern und Bulimikerinnen erhöhen (PERREZ & REICHERTS, 1989). Die Einsatzmöglichkeiten des hier vorgestellten Erhebungsverfahrens müssen daher nicht auf die Normalpopulation beschränkt bleiben.

Die Untersuchungsresultate sind insgesamt so ermutigend, daß dem weiteren Einsatz eines solchen Datenerhebungsinstruments zu Forschungszwecken und später auch für individualdiagnostische Fragestellungen grundsätzlich nichts im Wege steht. Dabei mag der Aufwand in der Durchführung im Verhältnis zum Nutzen zunächst als etwas zu groß erscheinen, doch wenn man sich vor Augen hält, daß das wirklich Bedeutsame am Gegenstandsbereich der Psychologie der ganze Mensch ist und nicht ein paar seiner isoliert betrachteten Teilfunktionen, so relativiert sich dieser Aufwand wieder. Mit dem hier vorgestellten Instrument können zwar auch nur einige wenige Merkmale erhoben werden, aber immerhin dort, wo der Mensch wirklich lebt - in seiner Alltagsumgebung!

4.2. Ausblick

Eine Weiterentwicklung des Instruments erfordert im nächsten Untersuchungsschritt (wieder als ESM-Studie mit Protokollbögen und elektronischen Signalgebern durchzuführen) eine Verbesserung und d.h. vor allem eine Vergrößerung des Itempools, um so auf faktorenanalytischem Wege eindeutigere Stimmungsdimensionen zu erhalten. Die fünf Items der dysphorischen Stimmung bedürfen dabei am wenigsten einer Ergänzung oder Modifikation, wichtiger ist eine Optimierung des euphorischen Bereichs. Dort kann das Item "Ich fühle mich gewachsen" aufgrund seiner geringen Änderungssensitivität gestrichen werden. Die Eliminierung eines der drei zum Konstrukt Freude gehörigen Items ("erfreut", "vergnügt" oder "froh") kann ebenfalls erwogen werden, denn sie kovariieren in sehr hohem Maße miteinander. Da der Itemumfang einer Skala besonders in einem Instrument für ESM-Studien nicht beliebig groß sein kann, muß eine besonders sorgfältige Abwägung des "Bandwidth-Fidelity Dilemma" (MURPHY & DAVIDSHOFER, 1991, S. 104) erfolgen, demzufolge bei konstanter Itemanzahl eine Verbreiterung des zu messenden Merkmalsbereichs nur mit einem Verlust an Reliabilität zu erreichen ist. Das Konstrukt euphorische Stimmung ist so breit angelegt, daß noch mehr seiner Aspekte erfaßt werden müssen, als es durch die hier vorgestellten fünf Items möglich ist. Eine Eliminierung eines dieser drei genannten und im Grunde redundanten Items bedeutet daher zwar vermutlich einen Reliabilitätsverlust für die Skala, doch vergrößert sich bei gleichzeitiger Hinzunahme anderer Items der euphorischen Stimmung die Bandbreite der von der Skala erfaßten Informationen. Der Itempool kann dabei erheblich aufgestockt werden, eine Gesamtzahl von 55 bis 60 Befindlichkeitsitems (einschließlich der Pufferitems) ist nicht unrealistisch, insbesondere auch nicht im Hinblick auf die "Schallmauer" von 20 Vp-Stunden bei den Psychologiestudenten. Dabei sollte jedoch nach einer Eingewöhnungszeit eine für ESM-Studien typische Bearbeitungsdauer von zwei Minuten je Protokollbogen (CSIKSZENTMIHALYI & LARSON, 1987) nicht wesentlich überschritten werden. Eine Erweiterung des Itempools sollte unter dem Gesichtspunkt einer Optimierung der bipolaren Skala B10 erfolgen.

Auf der methodischen Seite muß als nächstes der zwischen der ersten und zweiten Messung auftretende Reaktivitätseffekt näher untersucht und möglichst beseitigt werden. Eine weitere Untersuchung dieses Effekts am vorliegenden Datenmaterial müßte zum einen den Einsatz zeitreihenanalytischer Verfahren beinhalten, zum anderen sollten die Daten nicht nur auf Mittelwertveränderungen, sondern auch auf Varianzveränderungen hin analysiert werden. Im Rahmen eines zeitreihenanalytischen Ansatzes müssen die Daten auf mögliche Trends und (circadiane) zyklische Veränderungen überprüft werden.

Eine Kontrolle des Reaktivitätseffekts könnte im Rahmen einer Datenrecordererhebung gelingen: Dort brauchen die doppelt vorgegebenen Items nicht durch eine große Anzahl von nacheinander vorgegebenen Pufferitems in zwei Blöcke getrennt zu werden, sondern sie können paarweise, minimal nur durch gerade so viele Pufferitems getrennt vorgegeben werden, wie nötig sind, um die retroaktive Hemmung der Pufferitems zur Wirkung kommen zu lassen. Die Abstände zwischen erster und zweiter Itemvorgabe könnten dabei zwischen den Protokollterminen ebenso zufällig variiert werden, wie die Positionen der Itempaare in der Abfolge der Items. Dadurch würden die beiden doppelt vorgegebenen Items manchmal beide kurz nacheinander zu Beginn der Protokollierung,

manchmal erst am Ende der Protokollierung, mal mit einem größeren (zeitlichen) Abstand voneinander und mal mit einem kleineren vorgegeben werden. Ein alleine zeitkorreliert und itemunabhängig während der Protokollierung auftretender Reaktivitätseffekt würde sich so über den gesamten Beobachtungszeitraum randomisiert auf alle Items auswirken, und zwar sowohl auf die Items der ersten wie der zweiten Messung. Als Pufferitems brauchen dabei nicht nur neutrale (einmal vorgegebene) Items verwendet zu werden, es können in der (minimalen) Lücke zwischen der ersten und zweiten Messung des Items A durchaus auch die Erstvorgabe des Items B und die Zweitvorgabe des Items C positioniert werden.

Ist eine Neutralisierung des Reaktivitätseffekts und die Herstellung einer Äquivalenz der ersten und zweiten Messung gelungen, so können pro Item, Vp und Protokolltermin die mittleren Meßwerte aus beiden Messungen für die weitere Datenverarbeitung verwendet werden. Ein Ermitteln von getrennten Parametern für die erste und zweite Messung erübrigt sich dann. Nach einer Optimierung und einer befriedigenden (zumindest Kriterien-) Validierung der Skala könnten dann interindividuelle Normierungen des Instruments auf der Traitebene vorgenommen werden. Intraindividuelle Normierungen auf der Stateebene könnten bereits jetzt für jede Vp erfolgen. Dies würde etwa dann Sinn machen, wenn die Vpn das Instrument nicht in Alltagssituationen bearbeiten, sondern während eines längeren Treatments, z.B. während eines mehrtägigen Therapieworkshops. Dort könnten dann gezielt die situativen Auswirkungen auf die Befindlichkeit als Abweichungen vom intraindividuellen Mittelwert dargestellt werden.

Weitere Einsatzmöglichkeiten des Instruments können z.B. in Langzeituntersuchungen experimenteller, differentiell- oder ökopsychologischer Art (etwa zur transsituativen Konsistenz des Verhaltens und Erlebens), sowohl unter Feld- als auch unter Laborbedingungen liegen. So ein Instrument kann zur Verlaufskontrolle vielfältiger Prozesse, z.B. therapeutischer oder anderer Interventionen, zum Erfassen des zeitlichen Verlaufs nicht beeinflußter Krankheitsbilder, aber auch zum Abbilden periodischer Veränderungen von Befindlichkeiten unter Alltagsbedingungen bei normalen Vpn, zu Forschungs- oder zu diagnostischen Zwecken nützlich sein. Es eignet sich ausschließlich zur Erhebung solcher Informationen, von deren Bekanntwerden die Vpn zumindest keine Nachteile befürchten müssen, denn dieser methodische Ansatz schließt eine Kontrolle der Erhebungsbedingungen durch den Untersuchungsleiter aus und erfordert somit eine sehr hohe Compliance der Vpn. Wenn seine Konstruktvalidität abgeschätzt werden kann, kann dieses Instrument auch zur Validierung herkömmlicher Traitinstrumente dienen.

Für eine denkbare Folgestudie können eine Reihe von Anregungen für methodische Verbesserungen gegeben werden. So kann die Ratingskala der Protokollbögen anders formuliert und vor allem symmetrisch gestaltet werden, möglicherweise kann auch die Zahl der Antwortstufen vergrößert werden. (Soll jedoch das Konstrukt Angst gemessen werden, so empfiehlt sich die Beibehaltung der hier verwendeten Grundstruktur des Instruments einschließlich der STAI-Stateitems.) Der Itempool sollte vergrößert und verändert werden (s.o.), die Items zur Tätigkeit der Vpn, zum Setting usw. können bei Bedarf erheblich differenzierter gestaltet werden. Items, die den Vpn zur Beschreibung ihrer tatsächlich erlebten Gefühle gefehlt haben (z.B. Ärger, Traurigkeit), sollten in die Itemsammlung mit aufgenommen werden, auch wenn sie für die Fragestellung irrelevant sind. Es ist

wichtig, daß die Vpn das Gefühl bekommen, nicht nur als Merkmalsträger benutzt, sondern auch als Menschen wirklich gefragt zu werden. Die Anzahl der täglichen Protokolltermine und die Dauer des gesamten Beobachtungszeitraums können variiert werden; dabei müssen die Belastbarkeit der Vpn und die meßbedingten Beeinflussungen des Verhaltensstroms gegen die Notwendigkeit, zur Abbildung von Prozessen und periodischen Veränderungen eine möglichst hohe Dichte der Protokolltermine zu gewährleisten, abgewogen werden. Theoretische Kenntnisse über die zu erwartenden Änderungsverläufe müssen dabei berücksichtigt werden. In den Nachbefragungsbogen muß die Frage danach, ob den Vpn die Struktur der Protokollbögen deutlich geworden ist, aufgenommen werden; einige andere Fragen können modifiziert oder gestrichen werden. Das gleiche gilt für einige Aussagen in der Itembatterie der Frage 6, deren Antwortskala von der Richtung her umgepolt werden sollte, um einer möglichen Akquieszenz entgegenzuwirken. In der Testeinweisung und Vorbefragung sollte der SVF gestrichen werden (es sei denn, es geht konkret um Fragestellungen im Zusammenhang mit Maßnahmen zur Streßbewältigung). Aus Gründen des mangelnden Bekanntheitsgrades sollte die PRF als Persönlichkeitsfragebogen beibehalten werden, denkbar ist aber auch eine versuchsweise Verwendung des (bei Psychologiestudenten bekannteren) FPI (FAHRENBERG, HAMPEL & SELG, 1984). Unter Validierungsgesichtspunkten sollten aber auch spezielle Traitmaße aus dem Stimmungsbereich erhoben werden; zur Angstmessung kann der STAI-Traitteil weiter verwendet werden, zur Erhebung von Meßwerten zur depressiven Verstimmung, besonders bei klinischen Fragestellungen, kann auf einige der Instrumente zurückgegriffen werden, die bei FÄHNDRICH, HELMCHEN & LINDEN (1986) aufgeführt werden. Sinnvoll ist auch eine Vorgabe der Befindlichkeitsitems aus den Protokollbögen in der Testeinweisung und Vorbefragung unter Verwendung derselben Ratingskala, aber mit der Instruktion zur Einschätzung der durchschnittlichen (Trait-) Stimmung. Auf ähnliche Art könnte mit veränderten Instruktionen auch eine mittlere Einschätzung der jeweiligen Tagesstimmung am Abend eines jeden Erhebungstages gewonnen werden. Ratings zur eigenen Stimmungsvariabilität könnten ebenfalls in der Testeinweisung und Vorbefragung erhoben und mit der tatsächlich ermittelten Variabilität verglichen werden.

5. Zusammenfassung

In dieser Arbeit werden die ersten Schritte zur Entwicklung eines änderungssensitiven Instruments zur Messung der Befindlichkeit (Stimmungen) unter Feldbedingungen beschrieben. Es werden die für seine Konstruktion notwendigen testtheoretischen, ökopsychologischen und differentiellpsychologischen Grundlagen erläutert. Dabei wird das von BUSE & PAWLIK (1991) vorgestellte State-Trait-Modell, das die testtheoretische Basis des entwickelten Meßinstruments bildet, aus der klassischen Testtheorie hergeleitet. Dieses Modell erfordert je Versuchsperson und je Meßzeitpunkt die doppelte Vorgabe derselben Items und gestattet bei Mehrzeitpunkterhebungen die Ermittlung eines Kennwertes zur Bestimmung der Änderungssensitivität der Items für Zustandsänderungen von einem Meßzeitpunkt zum anderen. Die im Rahmen dieser Arbeit durchgeführte Untersuchung wurde als Feldstudie im Sinne einer "experience sampling method" konzipiert, in der mit Hilfe eines randomisierten Zeitstichprobenplans unter Einsatz eines elektronischen Signalgebers und eines als Stimmungsprotokoll gestalteten (Papier- und Bleistift-) Erhebungsinstruments ökologisch repräsentative Stichproben aus dem alltäglichen Verhaltensstrom

von 74 Versuchspersonen gezogen wurden. Auf der Basis der erhobenen Daten wird das State-Trait-Modell empirisch überprüft und es werden Vorschläge zu einer Kombination einzelner Befindlichkeitsitems im Sinne von änderungssensitiven Skalen gemacht. Diese Skalen werden anschließend im Hinblick auf ihre psychometrischen Eigenschaften untersucht. Parallel dazu wird die "experience sampling method" als eine noch relativ neue Untersuchungsmethode einer kritischen Bewertung unterzogen.

Literatur

ALLPORT, G.W. (1966). Traits revisited. *American Psychologist, 21,* 1-10.

AMELANG, M. & BARTUSSEK, D. (1981). *Differentielle Psychologie und Persönlichkeitsforschung.* Stuttgart: Kohlhammer.

AMELANG, M. & BORKENAU, P. (1986). The trait concept: Current theoretical considerations, empirical facts, and implications for personality inventory construction. In: ANGLEITNER, A. & WIGGINS, J.S. (Eds.): *Personality assessment via questionnaires.* Heidelberg: Springer, 7-34.

ATTESLANDER, P. & KOPP, M. (1987). Befragung. In: ROTH, E. (Hrsg.): *Sozialwissenschaftliche Methoden.* München: Oldenbourg, 144-172.

BACKHAUS, K., ERICHSON, B., PLINKE, W. & WEIBER, R. (1990). *Multivariate Analysemethoden.* Berlin: Springer.

BARKER, R.G. & WRIGHT, H.F. (1951). *One boy's day.* New York: Harper & Brothers.

BASTINE, R.H.E. (1990). *Klinische Psychologie, Band 1.* Stuttgart: Kohlhammer.

BEM, D.J. & ALLEN, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506-520.

BEREITER, C. (1967). Some persisting dilemmas in measurement of change. In: HARRIS, C.W. (Ed.): *Problems in measuring change.* Madison: University of Wisconsin Press, 3-20.

BOHNER, G., HORMUTH, S.E. & SCHWARZ, N. (1991). Die Stimmungs-Skala: Vorstellung und Validierung einer deutschen Version des "Mood Survey". *Diagnostica, 37,* 135-148.

BORTZ, J. (1984). *Lehrbuch der empirischen Forschung für Sozialwissenschaftler.* Berlin: Springer.

BORTZ, J. (1989). *Statistik für Sozialwissenschaftler.* Berlin: Springer.

BORTZ, J., LIENERT, G.A. & BOEHNKE, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

BOTTENBERG, E.H. (1970). Stimmung: Dimensionierte Messung, Situations- und Persönlichkeitsabhängigkeit. *Psychologie und Praxis, 14,* 18-37.

BOWERS, K.S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review, 80,* 307-336.

BRANDSTÄTTER, H. (1983). Emotional responses to other persons in everyday life situations. *Journal of Personality and Social Psychology, 45,* 871-883.

BROSIUS, G. (1988). *SPSS/PC+ Basics und Graphics.* Hamburg: McGraw-Hill.

BROSIUS, G. (1989). *SPSS/PC+ Advanced Statistics und Tables.* Hamburg: McGraw-Hill.

BÜNNING, J. (1984). *Eine empirische Untersuchung zur Struktur von Stimmungsfaktoren bei Anwendung einer Adjektiv-Checkliste unter besonderer Berücksichtigung ihrer Skalierung.* Diplomarbeit (unveröffentl.), Hamburg.

BUSE, L. & PAWLIK, K. (1984). Inter-Setting-Korrelationen und Setting-Persönlichkeit-Wechselwirkungen: Ergebnisse einer Felduntersuchung zur Konsistenz von Verhalten und Erleben. *Zeitschrift für Sozialpsychologie, 15,* 44-59.

BUSE, L. & PAWLIK, K. (1990). Verhaltenseinschätzung. In: KRUSE, L., GRAUMANN, C.-F. & LANTERMANN, E.-D. (Hrsg.): *Ökologische Psychologie.* München: PVU, 213-217.

BUSE, L. & PAWLIK, K. (1991). Zur State-Trait-Charakteristik verschiedener Meßvariablen der psychophysiologischen Aktivierung, der kognitiven Leistung und der Stimmung in Alltagssituationen. *Zeitschrift für experimentelle und angewandte Psychologie, 38,* 521-538.

CAMPBELL, J.D., CHEW, B. & SCRATCHLEY, L.S. (1991). Cognitive and emotional reactions to daily events: The effects of self-esteem and self-complexity. *Journal of Personality, 59,* 473-505.

CANTOR, N., NOREM, J., LANGSTON, C., ZIRKEL, S., FLEESON, W. & COOK-FLANNAGAN, C. (1991). Life tasks and daily life experience. *Journal of Personality, 59,* 425-451.

CATTELL, R.B. (1967). The structuring of change by P-technique and incremental R-technique. In: HARRIS, C.W. (Ed.): *Problems in measuring change.* Madison: University of Wisconsin Press, 167-198.

CATTELL, R.B. (1973). *Personality and mood by questionnaire.* San Francisco: Jossey-Bass.

CLAUSS, G. & EBNER, H. (1985). *Statistik.* Thun: Harri Deutsch.

CONRAD, W. (1992). Diagnostik als Messung. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik.* Weinheim: PVU, 245-256.

CRONBACH, L.J. & MEEHL, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

CSIKSZENTMIHALYI, M. & FIGURSKI, T.J. (1982). Self-awareness and aversive experience in everyday life. *Journal of Personality, 50,* 15-28.

CSIKSZENTMIHALYI, M. & LARSON, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease, 175,* 526-536.

DIENER, E. & LARSEN, R.J. (1984). Temporal stability and cross-situational consistency of affective, behavioral, and cognitive responses. *Journal of Personality and Social Psychology, 47,* 871-883.

DIENER, E., LARSEN, R.J. & EMMONS, R.A. (1984). Person x situation interactions: Choice of situations and congruence response models. *Journal of Personality and Social Psychology, 47,* 580-592.

DORSCH, F. (1987). *Psychologisches Wörterbuch.* DORSCH, F., HÄCKER, H. & STAPF, K.-H. (Hrsg.). Bern: Huber.

EPSTEIN, S. (1977). Traits are alive and well. In: MAGNUSSON, D. & ENDLER, N.S. (Eds.): *Personality at the crossroads: Current issues in interactional psychology.* New York: Wiley, 83-98.

EPSTEIN, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

EPSTEIN, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790-806.

EPSTEIN, S. & O'BRIEN, E.J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin, 98,* 513-537.

EWERT, O. (1983). Ergebnisse und Probleme der Emotionsforschung. In: THOMAE, H. (Hrsg.): *Theorien und Formen der Motivation (Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Band 1, Serie IV).* Göttingen: Hogrefe, 397-452.

EYSENCK, H.J. (1980). *Kriminalität und Persönlichkeit.* Frankfurt a.M.: Ullstein.

FÄHNDRICH, E., HELMCHEN, H. & LINDEN, M. (1986). Standardized instruments used in the assessment of depression in German-speaking countries. In: SARTORIUS, N. & BAN, T.A. (Eds.): *Assessment of depression.* Berlin: Springer, 1-8.

FAHRENBERG, J., HAMPEL, R. & SELG, H. (1984). *Das Freiburger Persönlichkeitsinventar FPI.* Göttingen: Hogrefe.

FAHRENBERG, J. & HEGER, R. (1991). Differentielle Psychophysiologie von Befinden, Blutdruck und Herzfrequenz im Labor-Feld-Vergleich. *Zeitschrift für Differentielle und Diagnostische Psychologie, 12,* 1-25.

FISCHER, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

FRÖHLICH, W.D. (1983). Perspektiven der Angstforschung. In: THOMAE, H. (Hrsg.): *Psychologie der Motive (Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Band 2, Serie IV).* Göttingen: Hogrefe, 110-320.

GACHOWETZ, H. (1987). Feldforschung. In: ROTH, E. (Hrsg.): *Sozialwissenschaftliche Methoden.* München: Oldenbourg, 255-276.

GIBBS, J.C. (1979). The meaning of ecologically oriented inquiry in contemporary psychology. *American Psychologist, 34,* 127-140.

GNIECH, G. (1976). *Störeffekte in psychologischen Experimenten.* Stuttgart: Kohlhammer.

GRUBITZSCH, S. (1978a). Sozialökonomische Grundlagen des Testens und Messens. In: GRUBITZSCH, S. & REXILIUS, G. (Hrsg.): *Testtheorie - Testpraxis.* Reinbek: Rowohlt, 40-51.

GRUBITZSCH, S. (1978b). Konstruktion psychologischer Tests. In: GRUBITZSCH, S. & REXILIUS, G. (Hrsg.): *Testtheorie - Testpraxis.* Reinbek: Rowohlt, 75-111.

GULLIKSEN, H. (1950). *Theory of mental tests.* New York: Wiley.

HAMPEL, R. (1977). Adjektiv-Skalen zur Einschätzung der Stimmung (SES). *Diagnostica, 23,* 43-60.

HECHELTJEN, K.-G. & MERTESDORF, F. (1973). Entwicklung eines mehrdimensionalen Stimmungsfragebogens (MSF). *Gruppendynamik, 4,* 110-122.

HEDGES, S.M., JANDORF, L. & STONE, A.M. (1985). Meaning of daily mood assessments. *Journal of Personality and Social Psychology, 48,* 428-434.

HEIDENREICH, K. (1987). Entwicklung von Skalen. In: ROTH, E. (Hrsg.): *Sozialwissenschaftliche Methoden.* München: Oldenbourg, 417-449.

HJELLE, L.A. & ZIEGLER, D.J. (1981). *Personality theories.* New York: McGraw-Hill.

HÖRMANN, H. (1982). Theoretische Grundlagen der projektiven Verfahren. In: GROFFMANN, K.-J. & MICHEL, L. (Hrsg.): *Grundlagen psychologischer Diagnostik (Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Band 1, Serie II).* Göttingen: Hogrefe, 173-228.

HORMUTH, S.E. (1986). The sampling of experiences in situ. *Journal of Personality, 54,* 262-293.

JÄGER, R.S. (1992). Statusdiagnostik. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik.* Weinheim: PVU, 200-202.

JÄGER, R.S. & SCHEURER, H. (1992). Prozeßdiagnostik. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik.* Weinheim: PVU, 202-208.

JANKE, W. & DEBUS, G. (1978). *Die Eigenschaftswörterliste EWL.* Göttingen: Hogrefe.

JANKE, W., ERDMANN, G. & BOUCSEIN, W. (1985). *Streßverarbeitungsfragebogen.* JANKE, W., ERDMANN, G. & KALLUS, W. (Hrsg.). Göttingen: Hogrefe.

JOERGES, B. (1990). Person und dingliche Umwelt. In: HOYOS, C. Graf, KROEBER-RIEL, W., ROSENSTIEL, L. v. & STRÜMPEL, B. (Hrsg.): *Wirtschaftspsychologie in Grundbegriffen.* München: PVU, 446-459.

KAMINSKI, G. & BELLOWS, S. (1982). Feldforschung in der ökologischen Psychologie. In: PATRY, J.-L. (Hrsg.): *Feldforschung.* Bern: Huber, 87-116.

KROHNE, H.W. & KOHLMANN, C.-W. (1990). Persönlichkeit und Emotion. In: SCHERER, K.R. (Hrsg.): *Psychologie der Emotion (Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Band 3, Serie IV).* Göttingen: Hogrefe, 485-559.

LANGER, I. (1969). *Methodische und methodisch-statistische Probleme in der Psychotherapieforschung.* Diplomarbeit (unveröffentl.), Hamburg.

LARSEN, R.J. & KASIMATIS, M. (1991). Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality, 59,* 387-423.

LAUX, L. (1983). Psychologische Streßkonzeptionen. In: THOMAE, H. (Hrsg.): *Theorien und Formen der Motivation (Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Band 1, Serie IV).* Göttingen: Hogrefe 453-535.

LAUX, L., GLANZMANN, P., SCHAFFNER, P. & SPIELBERGER, C.D. (1981). *Das State-Trait-Angstinventar.* Weinheim: Beltz.

LIENERT, G.A. (1989). *Testaufbau und Testanalyse.* München: PVU.

LÖSEL, F. (1992). Persönlichkeitsdaten (Tests). In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik.* Weinheim: PVU, 362-380.

LORD, F.M. (1967). Elementary models for measuring change. In: HARRIS, C.W. (Ed.): *Problems in measuring change.* Madison: University of Wisconsin Press, 21-38.

LORR, M. & SHEA, T.M. (1979). Are mood states bipolar? *Journal of Personality Assessment, 43,* 468-472.

MAJCEN, A.-M., STEYER, R. & SCHWENKMEZGER, P. (1988). Konsistenz und Spezifität bei Eigenschafts- und Zustandsangst. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9,* 105-120.

McADAMS, D.P. & CONSTANTIAN, C.A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology, 45,* 851-861.

MEDDIS, R. (1972). Bipolar factors in mood adjective checklists. *British Journal of Social and Clinical Psychology, 11,* 178-184.

MICHEL, L. & CONRAD, W. (1982). Theoretische Grundlagen psychometrischer Tests. In: GROFFMANN, K.-J. & MICHEL, L. (Hrsg.): *Grundlagen psychologischer Diagnostik (Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Band 1, Serie II).* Göttingen: Hogrefe, 1-129.

MISCHEL, W. (1968). *Personality and assessment.* New York: Wiley.

MOOSBRUGGER, H. (1992). Testtheorie: Klassische Ansätze. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik.* Weinheim: PVU, 310-322.

MURPHY, K.R. & DAVIDSHOFER, C.O. (1991). *Psychological testing: Principles and applications.* Englewood Cliffs: Prentice-Hall.

NOWLIS, D.P. & COHEN, A.Y. (1968). Mood-reports and the college natural setting: A day in the lives of three roommates under academic pressure. *Psychological Reports, 23,* 551-566.

NOWLIS, V. (1965). Research with the mood adjective check list. In: TOMKINS, S.S. & IZARD, C.E. (Eds.): *Affect, cognition, and personality.* New York: Springer, 352-389.

ORLIK, P. (1979). Sozialpsychologische Feldforschung. In: HEIGL-EVERS, A. & STREEK, U. (Hrsg.): *Die Psychologie des 20. Jahrhunderts, Band 8.* Zürich: Kindler, 110-116.

PALUS, C.J., NASBY, W. & EASTON, R.D. (1990). Executive identity and the hero's story: The voyage of Dodge Morgan and the American Promise. *Journal of Applied Behavioral Science, 26,* 501-527.

PASSINI, F.T. & NORMAN, W.T. (1966). A universal conception of personality structure? *Journal of Personality and Social Psychology, 4,* 44-49.

PATRICK, A.W., ZUCKERMAN, M. & MASTERSON, F.A. (1974). An extension of the trait-state distinction from affects to motive measures. *Psychological Reports, 34,* 1251-1258.

PATRY, J.-L. (1982). *Feldforschung.* Bern: Huber.

PAWLIK, K. (1978). Umwelt und Persönlichkeit: Zum Verhältnis von ökologischer und differentieller Psychologie. In: GRAUMANN, C.F. (Hrsg.): *Ökologische Perspektiven in der Psychologie.* Bern: Huber, 112-134.

PAWLIK, K. (1982). Modell- und Praxisdimensionen psychologischer Diagnostik. In: PAWLIK, K. (Hrsg.): *Diagnose der Diagnostik.* Stuttgart: Klett-Cotta, 13-43.

PAWLIK, K. (1988). "Naturalistische" Daten für Psychodiagnostik: Zur Methodik psychodiagnostischer Felderhebungen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9,* 169-181.

PAWLIK, K. & BUSE, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3,* 101-118.

PAWLIK, K. & BUSE, L. (1985). Verhalten in situ: Felduntersuchungen zur Umwelt-Verhaltens-Kovariation an männlichen Jugendlichen. In: ALBERT, D. (Hrsg.): *Bericht über den 34. Kongreß der Deutschen Gesellschaft für Psychologie in Wien 1984, Band 2*. Göttingen: Hogrefe, 843-846.

PAWLIK, K. & BUSE, L. (1992). Felduntersuchungen zur transsituativen Konsistenz individueller Unterschiede im Erleben und Verhalten. In: PAWLIK, K. & STAPF, K. (Hrsg.): *Umwelt und Verhalten*. Bern: Huber, 25-69.

PERREZ, M. & REICHERTS, M. (1989). Belastungsverarbeitung: Computerunterstützte Selbstbeobachtung im Feld. *Zeitschrift für Differentielle und Diagnostische Psychologie, 10,* 129-139.

PETERMANN, F. (1978). *Veränderungsmessung*. Stuttgart: Kohlhammer.

PETERMANN, F. (1986). Probleme und neuere Entwicklungen der Veränderungsmessung - Ein Überblick. *Diagnostica, 32,* 4-16.

PETERMANN, F. (1992). Situationsbezogene Diagnostik. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik*. Weinheim: PVU, 268-273.

PLUTCHIK, R. & CONTE, H.R. (1989). Measuring emotions and their derivatives: Personality traits, ego defenses, and coping styles. In: WETZLER, S. & KATZ, M.M. (Eds.): *Contemporary approaches to psychological assessment*. New York: Brunner / Mazel, 239-269.

RAUCHFLEISCH, U. (1989). *Testpsychologie*. Göttingen: Vandenhoeck & Ruprecht.

RENN, H. (1973). *Die Messung von Sozialisierungswirkungen*. München: Oldenbourg.

RETTLER, H. (1992). Verschränkung von Methode und Theorie. In: JÄGER, R.S. & PETERMANN, F. (Hrsg.): *Psychologische Diagnostik*. Weinheim: PVU, 277-286.

REXILIUS, G. (1978). Grenzen der Testerei. In: GRUBITZSCH, S. & REXILIUS, G. (Hrsg.): *Testtheorie - Testpraxis*. Reinbek: Rowohlt, 112-167.

ROLLETT, B. (1982). Kriterienorientierte Prozeßdiagnostik im Behandlungskontext. In: PAWLIK, K. (Hrsg.): *Diagnose der Diagnostik*. Stuttgart: Klett-Cotta, 131-148.

RORSCHACH, H. (1972). *Psychodiagnostik*. MORGENTHALER, W. (Hrsg.). Bern: Huber.

RUDINGER, G. (1985). Prozeß-Analysen. In: HERRMANN, T. & LANTERMANN, E.-D. (Hrsg.): *Persönlichkeitspsychologie*. München: Urban & Schwarzenberg, 202-214.

RUSSELL, J.A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37,* 345-356.

SACHS, L. (1974). *Angewandte Statistik.* Berlin: Springer.

SCHELTEN, A. (1980). *Grundlagen der Testbeurteilung und Testerstellung.* Heidelberg: Quelle und Meyer.

SCHMITZ, B. (1989). *Einführung in die Zeitreihenanalyse.* Bern: Huber.

SCHUBÖ, W., UEHLINGER, H.-M., PERLETH, C., SCHRÖGER, E. & SIERWALD, W. (1991). *SPSS Handbuch der Programmversionen 4.0 und SPSS-X 3.0* . Stuttgart: Fischer.

SCHULZ, T., MUTHIG, K.-P. & KOEPPLER, K. (1981). *Theorie, Experiment und Versuchsplanung in der Psychologie.* Stuttgart: Kohlhammer.

SCHWENKMEZGER, P. (1984). Kann durch das Prinzip der Aggregation von Daten die Konsistenzannahme von Eigenschaften beibehalten werden? *Zeitschrift für Differentielle und Diagnostische Psychologie, 5,* 251-272.

SCHWENKMEZGER, P. (1985). Angst. In: HERRMANN, T. & LANTERMANN, E.-D. (Hrsg.): *Persönlichkeitspsychologie.* München: Urban & Schwarzenberg, 331-338.

SINGH, A.K. (1986). *Tests, measurements and research methods in behavioural sciences.* New Delhi: Tata McGraw-Hill.

SMITH, B.D. (1988). Personality - Multivariate systems theory and research. In: NESSELROADE, J.R. & CATTELL, R.B. (Eds.): *Handbook of multivariate experimental psychology.* New York: Plenum Press, 687-736.

SPIELBERGER, C.D. (1966). Theory and research on anxiety. In: SPIELBERGER, C.D. (Ed.): *Anxiety and behavior.* New York: Academic Press, 3-20.

SPIELBERGER, C.D. (1972). Anxiety as an emotional state. In: SPIELBERGER, C.D. (Ed.): *Anxiety - current trends in theory and research, vol. 1.* New York: Academic Press, 23-49.

SPIELBERGER, C.D. (1977). State-trait anxiety and interactional psychology. In: MAGNUSSON, D. & ENDLER, N.S. (Eds.): *Personality at the crossroads: Current issues in interactional psychology.* New York: Wiley, 173-183.

SPIELBERGER, C.D., LUSHENE, R.E. & McADOO, W.G. (1977). Theory and measurement of anxiety states. In: CATTELL, R.B. & DREGER, R.M. (Eds.): *Handbook of modern personality theory.* New York: Wiley, 239-253.

Statistisches Bundesamt (Hrsg.) (1992). *Statistisches Jahrbuch 1992 für die Bundesrepublik Deutschland.* Stuttgart: Metzler-Poeschel.

STERN, E. (1986). *Reaktivitätseffekte in Untersuchungen zur Selbstprotokollierung des Verhaltens im Feld.* Frankfurt a.M.: Lang.

STEYER, R. (1987). Konsistenz und Spezifität: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 8,* 245-258.

STONE, A.A., KESSLER, R.C. & HAYTHORNTHWAITE, J.A. (1991). Measuring daily events and experiences: Decisions for the researcher. *Journal of Personality, 59,* 575-607.

STUMPF, H., ANGLEITNER, A., WIECK, T., JACKSON, D.N. & BELOCH-TILL, H. (1985). *Deutsche Personality Research Form (PRF).* Göttingen: Hogrefe.

SUPPRIAN, U. (1976). *Eppendorfer Stimmungs-Antriebs-Skala - ESTA III.* Weinheim: Beltz.

TACK, W.H. (1986a). Reliabilitäts- und Effektfunktionen - Ein Ansatz zur Zuverlässigkeit von Meßwertänderungen. *Diagnostica, 32,* 48-63.

TACK, W.H. (1986b). Veränderungsmessung - Ein Vorwort. *Diagnostica, 32,* 1-3.

TENNEN, H., SULS, J. & AFFLECK, G. (1991). Personality and daily experience: The promise and the challenge. *Journal of Personality, 59,* 313-337.

UEHLINGER, H.-M. (1988). *SPSS/PC+ Benutzerhandbuch, Band 1.* Stuttgart: Fischer.

ULLRICH DE MUYNCK, R. & ULLRICH, R. (1981). *Das Emotionalitätsinventar als Befindlichkeitsmaß.* München: Pfeiffer.

UNDERWOOD, B., FROMING, W.J. & MOORE, B.S. (1980). Mood and personality: A search for the causal relationship. *Journal of Personality, 48,* 15-23.

UPMEYER, A. (1985). *Soziale Urteilsbildung.* Stuttgart: Kohlhammer.

WALTER, P. (1978). Meß- und testtheoretische Grundlagen psychologischen Testens. In: GRUBITZSCH, S. & REXILIUS, G. (Hrsg.): *Testtheorie - Testpraxis.* Reinbek: Rowohlt, 52-74.

WEST, S.G. & HEPWORTH, J.T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 609-662.

WHEELER, L. & REIS, H.T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59,* 339-354.

WICKER, A.W. (1979). Ecological psychology (some recent and prospective developments). *American Psychologist, 34,* 755-765.

WONG, M.M. & CSIKSZENTMIHALYI, M. (1991). Motivation and academic achievement: The effects of personality traits and the quality of experience. *Journal of Personality, 59,* 539-574.

ZERSSEN, D. v. (1976). *Die Befindlichkeits-Skala.* Weinheim: Beltz.

ZEVON, M.A. & TELLEGEN, A. (1982). The structure of mood change: An idiographic / nomothetic analysis. *Journal of Personality and Social Psychology, 43,* 111-122.

ZIELKE, M. (1979). *Kieler Änderungssensitive Symptomliste - KASSL.* Weinheim: Beltz.

ZUCKERMAN, M. (1977). Development of a situation-specific trait-state test for the prediction and measurement of affective responses. *Journal of Consulting and Clinical Psychology, 45,* 513-523.