

3 Kohlbergs *Moral Judgment Interview* – Eine Kritik

Die Überprüfung von Theorien zur moralischen Entwicklung stützt sich bislang weitgehend auf Untersuchungen mit dem *Moral Judgment Interview* (MJ) von Kohlberg und seinen Mitarbeitern. Mit dem MJ wird der Anspruch verbunden, die Struktur des moralischen Urteils theoretisch valide zu messen. Die Auswertungsmethode des MJ, das *Standard Issue Moral Judgment Interview and Scoring System*, bezeichnen die Autoren als "eine theoriegestützte Erfassungsmethode" (Colby, Kohlberg et al. 1987, S. 1). In diesem Kapitel soll untersucht werden, inwieweit das MJ diesen Anspruch theoretischer Validität erfüllt bzw. heute, nach vielfachen Revisionen besser als das ursprüngliche Verfahren erfüllt, und für die Klärung der Frage nach den Bedingungen der Moralentwicklung, unserem zentralen Thema, geeignet ist.

3.1 Zum Problem der Bestimmung der Validität eines Meßinstruments

Die Richtigkeit oder Falschheit einer wissenschaftlichen Hypothese läßt sich selten, wenn überhaupt, einfach an Daten ablesen; die Daten oder Beobachtungen, an denen eine Theorie überprüft werden kann, müssen, wie der Wissenschaftstheoretiker Mittelstraß (1992, S. 16 ff) feststellte, zumeist erst hergestellt werden. Das bedeutet nicht, daß die Daten genau passend zu den Vorhersagen der Theorie gemacht werden (dies wäre nur durch eine Fälschung der Daten möglich). Es heißt vielmehr, daß die Wahl der Meß- und Auswertungsmethode, also die situativen Bedingungen, unter denen eine Person antwortet, und die Vorgehensweise bei der Zusammenfassung und Bewertung ihrer Antworten zu der zu prüfenden Theorie passen muß und nicht in Widerspruch zu ihr stehen darf. Erst wenn diese Voraussetzung erfüllt ist, erlauben die Meßdaten eine adäquate Überprüfung der Theorie. Sind diese Voraussetzung jedoch nicht erfüllt, sind die Meßdaten nicht theoretisch valide. Werden invalide Meßdaten zur Bewertung einer Theorie herangezogen, kann dies zu krassen Fehleinschätzungen dieser Theorie – zur einer falschen Bestätigungen wie einer falschen Widerlegungen – führen. In der Psychologie wird die Frage nach der Validität eines Meßinstruments ("Mißt das Instrument wirklich das,

was es messen soll?") zumeist rein empirisch zu bestimmen versucht, wozu die Korrelation der Meßwerte mit einem oder mehreren Außenkriterien herangezogen wird. Ergibt sich eine hohe positive Korrelation mit Indikatoren für dasselbe Konstrukt (und ergeben sich gleichzeitig niedrige Korrelationen mit Indikatoren für andere Konstrukte), dann sieht man die Validität des Meßinstruments als gegeben an. Dieses Vorgehen empfiehlt sich jedoch nicht für unsere Untersuchung. Hätten wir bereits ein zweifelsfrei valides Instrument zur Messung des moralischen Urteils, das als Außenkriterium dienen könnte, würde sich hier jede Diskussion erübrigen. Unser Problem besteht ja gerade darin, daß moralische Urteilsfähigkeit kein einfaches Konstrukt ist, das sich an einfachen Beobachtungen ablesen läßt, sondern daß es sich hier um ein komplexes Konstrukt handelt, das sich nur in einer sorgfältig hergestellten Beobachtungssituation valide messen läßt. Nicht jedes Verhalten, das wir im Alltag als Ausdruck von Moralität ansehen mögen, gewährt einen zuverlässigen und eindeutigen Einblick in die moralischen Fähigkeiten einer Person. Wir kommen hierauf zurück (S. 70). Wie schwer es ist, eine experimentelle Situation so zu konstruieren, daß das (Antwort-) Verhalten der befragten Person eine valide Messung moralischer Urteilsstrukturen ermöglicht, zeigt die lange Entwicklungsgeschichte des MJI und seiner Vorgänger.

3.2 Von der Intuition zur Wissenschaft

Eines der ersten Instrumente zur Erfassung moralischer Urteilsfähigkeit ("sittliche Reife") stammt von Levy-Suhl (1912). Dieser hat in seinen Interviews mit straffälligen Jugendlichen gefunden, daß bestimmte moralische Werte immer wieder genannt wurden und daß diese Werte recht gut den sittlichen Entwicklungsstand des Jugendlichen widerspiegelten. Erst in jüngster Zeit sind diese Arbeiten wiederentdeckt worden. Heidbrink (1989) hat in einer Reanalyse der Daten von Levy-Suhl gezeigt, daß sich diese Werte in eine Entwicklungsskala ähnlich der Kohlbergschen bringen lassen. Offenbar ohne Kenntnis dieser Vorarbeit hat Piaget (1973a/1932) seine "klinische Methode" entwickelt, um das moralische Urteil beim Kinde zu studieren. Bei der klinischen Methode handelt sich um eine Kombination von klassischen Verhaltens-Experimenten in naturalistischen Kontexten mit der Befragung der handelnden Personen.

Kohlberg (1958) hat für die Untersuchung von Jugendlichen und Erwachsene Piagets Meßmethode zu einer reinen Befragungsmethode abgewandelt.

Im Kontext der Psychologie der 50er Jahre nahm sich Kohlbergs Meßmethode ebenso wie seine kognitiv-strukturelle Theorie des moralischen Urteils revolutionär aus. In dieser Zeit entwickelte sich die Psychologie "erratisch", wie Wolfle (1963) feststellte; sie war "erfüllt vom Streit darüber, was die wichtigsten Probleme und die stärksten Methoden" seien (S. v). Um die Ursache dieser Krise zu finden, setzte die *American Psychological Association* eine Forschergruppe ein, die als Kern des Problems jene Theorien identifizierte, die sich mit inneren psychologischen Prozessen oder intervenierenden Variablen befaßten (Koch 1963, S. 735). Dies kam einem Verdikt gegen Theorien wie der Piaget-Kohlberg-Theorie gleich. Verstärkt wurde die Neigung der Psychologie jener Zeit, sich von der Erforschung von psychischen Dispositionen ganz zurückzuziehen. Moralität wurde weitgehend gleichgesetzt mit moralischem Begriffswissen ("Weiß das Kind, daß Stehlen verboten ist?") und moralischen Einstellungen ("Wie stark lehnt ein Kind Stehlen ab?"); beides wurde vom Bereich des "kognitiven Verhaltens" gänzlich getrennt.¹ Sofern von der Struktur der Persönlichkeit gesprochen wurde, handelt es sich um eine (z.B. mit einer Faktorenanalyse gefundenen) Eigenschaft einer Personen-*gruppe*, aber nicht um die Disposition eines Individuums. Die strukturellen Aspekte des moralischen Verhaltens gerieten weitgehend aus dem Blick.

Um die moralische Struktur messen zu können, stützte sich Kohlberg anfänglich stark auf die "informierte psychologische Intuition" des Auswerters oder Schätzers (*rater*). Das größte Vertrauen hatte er in die Globaleinschätzung der Interview-Antworten anhand seines 6-Stufen-Schemas der Moralentwicklung (Kohlberg 1958; Oser 1988a). Dieses Vorgehen setzte voraus, daß die Schätzer mit Kohlbergs Theorie bestens vertraut waren. Da gemäß der Theorie die Schätzer, um die Antworten ihrer Probanden adäquat zu verstehen zu können, selbst eine mindestens ebenso hohe moralische Urteilsfähigkeit besitzen mußten wie diese, waren der Globaleinschätzung enge Grenzen gesetzt. Zudem war diese intuitive Meßmethode nur schwer in Einklang zu bringen mit der Grundforderung wissenschaftlicher Forschung nach vollständiger

1 Als Vertreter dieser Tendenz vgl. Hartshorne & May 1928, Krathwohl et al. 1964. Kritisch dazu Pittel & Mendelsohn 1966.

Wiederholbarkeit der Untersuchung² und allgemeiner Überprüfbarkeit der Ergebnisse: Die bei der Globaleinschätzung ablaufenden Deutungsprozesse könnten kaum exakt wiederholt werden; die Nachprüfung war nach der Theorie *a priori* auf die Leute beschränkt, die selbst nachgewiesen haben, daß sie mindestens eine ebenso hohe moralische Urteilsfähigkeit haben wie die Auswerter.

Kohlberg verteidigte diese "intuitive" Meßmethode mit dem Hinweis, daß die damals gängigen psychologischen Meßmethoden, vor allem die klassische Testtheorie, ungeeignet waren, um den kognitiv-strukturellen Aspekt der Moral adäquat zu erfassen (Kohlberg 1981). Dennoch erkannte er, daß, wie Colby, Kohlberg et al. (1987) schreiben, ein Wandel in "der Betrachtungsweise von Interpretation [von Daten] als eine Kunst zu der Betrachtungsweise von Interpretation als Wissenschaft" notwendig war (S. 40). Die Bemühung um eine mehr wissenschaftliche Dateninterpretation war aber – wohl bedingt durch das Fehlen von Alternativen und die heftige, um die Einhaltung der klassischen psychometrischen Regeln besorgte Kritik (Kurtines & Greif 1974) – hauptsächlich an jenen Regeln orientiert, die Kohlberg bis dahin als inkompatibel zu seiner Theorie ansah. Die Revision des MJI diene vor allem dem Ziel, "größere Objektivität und Reliabilität bei der Einstufung zu erreichen" (Colby, Kohlberg et al. 1987, S. 40; meine Übersetzung). Dies ist gut gelungen. Das modifizierte MJI ist, gemessen an den Standards der konventionellen Psychometrie, ein gutes Meßinstrument.³ Wie berichtet wird, erreichen die Schätzer eine fast perfekte Übereinstimmung innerhalb einer Drittstufe.⁴

2 Wenn die Ergebnisse nicht repliziert werden können, ist die geprüfte Theorie möglicherweise falsch, nicht aber das Experiment.

3 Die Reliabilität des MJI ist tatsächlich außerordentlich hoch. Innerhalb eines drei- bis sechswöchigen Zeitintervalls bleiben die Scores relativ stabil (die Test-Retest-Reliabilität liegt weit über 0,90). Die alternative Reliabilität wird mit 0,95 angegeben (S. 63-65). Die drei Reliabilitätsformen sind nicht völlig äquivalent; die Scores können sich bis zu einem Fünftel einer Stufe unterscheiden (S. 68). Der Meßfehler auf der Basis von Retestkorrelationen beträgt: 15,62 MMS-Punkte, berechnet nach der Formel (Colby, Kohlberg et al. 1987, S. 69):

$$\sigma_{mess} = \sigma * \sqrt{1 - r_{xy}}$$

4 Wegen der hohen Standardisierung der Auswertung benötigt die Ausbildung qualifizierter Schätzer nur kurze Zeit, trotz der nach wie vor geforderten

(Fortsetzung...)

Aber konnte durch diese Revision auch eine größere theoretische Validität des MJI, das heißt eine bessere Übereinstimmung zwischen Theorie und Methode erreicht werden?

3.3 Theorie: Ganzheitliche Betrachtungsweise des moralischen Urteilverhaltens

In Kohlbergs Theorie spielt die Struktur des moralischen Urteils eine zentrale Rolle.⁵ Kohlberg geht davon aus, daß das moralische Urteil nicht nur affektive Aspekte oder moralische Einstellungen umfaßt, sondern auch kognitiv-strukturelle Aspekte. Die Veränderung der kognitiv-strukturellen Aspekte der Moral bedingen in jeder Entwicklungsstufe unterschiedliche Denkweisen und nicht nur die Änderung von moralischen Einstellungen; die moralische Entwicklung ist also qualitativ-strukturell und nicht nur quantitativ. Die strukturell definierten Stufen moralischer Denkweisen bilden eine invariante Sequenz der moralischen Entwicklung. Schließlich, so Kohlberg, sind kognitiv-moralische Entwicklungsstufen hierarchisch integriert: Die Stufen bilden eine Anordnung zunehmend differenzierter und integrierter Strukturen. Bemerkenswerterweise spricht Kohlberg – zumindest in den theoretischen Teilen seiner Arbeiten – von moralischen Strukturen im Sinne von *manifestem Verhalten* und nicht im Sinne von latenten moralischen Vorstellungen: "Was uns wichtig ist, ist, wie moralische Urteile zustandekommen, wenn [ein moralisches Prinzip] wirklich auf Wertkonflikte angewendet werden. Moral ist eine Sache von Wahl und Entscheidung. Es ist nicht bloß eine Frage des Gebrauchs von abstrakten Prinzipien wie Gerechtigkeit". Entsprechend versuchen er und seine Kollegen "nicht die abstrakten philosophischen Positionen einer Person zu messen, sondern vielmehr das, was wirksam ist, wenn diese Person einer wirklichen moralischen Entscheidung gegenübersteht" (Colby, Kohlberg et al. 1987, S. 58). Ein struktureller Meßansatz muß also die im tatsächlichen Verhalten wirk-

4(...Fortsetzung)

Theoriekenntnis und Verstehensleistung. Bei den Schätzern handelte es sich um Universitätsstudenten mit psychologischem Vorwissen, die einige Zeit mit dem MJI gearbeitet oder an einem viertägigen Auswertungsseminar teilgenommen hatten.

5 Siehe Kohlberg 1984; auch Lind 1985b, Oser 1981a und Kapitel 2.

samen moralischen Dispositionen erfassen und nicht, wie etwa der phänomenologische Ansatz, die subjektive Wahrnehmung solcher Dispositionen durch die befragte Person. Im strukturellen Meßansatz wird die befragte Person zum Beispiel aufgefordert, ihre Meinung zu einem bestimmten moralischen Problem zu begründen; im phänomenologischen wird sie dagegen veranlaßt anzugeben, welche moralische Argumente sie in solchen Situationen typischerweise verwenden würde bzw. an welchen Werte sie sich gewöhnlich orientiert.⁶

Aus diesen Postulaten und Annahmen folgt, daß nicht jedes Verhalten oder jede Verhaltenseinheit geeignet ist für die Messung moralischer Dispositionen. Wenn eine Person mit einem moralischen Dilemma konfrontiert ist, sind drei Arten von Verhalten zu unterscheiden: die Meinung einer Person über die richtige Lösung eines Dilemmas, ihre Begründung für diese Meinung und die moralische Bedeutung dieser Begründung. Davon sind die ersten beiden Verhaltenseinheiten nicht für die Erfassung moralischer Dispositionen geeignet: 1. Vom Verhalten als *Meinung zu einer bestimmten Entscheidung* kann nicht auf moralische Urteilsstrukturen und Stufen geschlossen werden. Eine solche Entscheidung ist gewöhnlich von sehr vielen persönlichen und situativen Merkmalen bestimmt und daher ambivalent. Im bekannten "Heinz-Dilemma" beispielsweise soll der oder die Befragte sich entscheiden, ob Heinz, um das Leben seiner kranken Frau zu retten, eine sehr teure Medizin stehlen soll, die er sich finanziell nicht leisten kann. Diese Entscheidung ist stark situations-

6 Beide Meßansätze sind "subjektiv" insofern in beiden das Subjekt involviert wird. Aber es wäre falsch und irreführend deshalb beide als phänomenologisch zu bezeichnen, wie es Kohlberg unter dem Einfluß seiner Habermas-Rezeptionen zuletzt tat und viele seiner Schüler tun. Dieses Mißverständnis scheint auch viel zu der Verwirrung um das sogenannte Urteil-Handlungs-Problem beizutragen. Wird nämlich das moralische Urteilen als eine vom Verhalten unabhängige Entität begriffen, die nicht direkt gemessen, sondern nur aus den Berichten der betreffenden Person erschlossen werden kann, stellt sich in der Tat die Frage, inwieweit diese Berichte mit dem tatsächlichen Urteilsverhalten korrelieren. Will er aber diese Frage empirisch klären, kommt der radikal-phänomenologische Meßansatz in eine Zwickmühle. Wenn eine direkte Erfassung moralischer Dispositionen nicht möglich ist, läßt sich die Frage nämlich prinzipiell nicht empirisch beantworten; würde man zugeben, daß eine direkte Erfassung möglich ist, würde dieser Meßansatz sich selbst obsolet machen.

gebunden. Schon geringe Modifikationen der Situation (z.B., wenn die Medizin von der Krankenkasse bezahlt wird oder wenn der Befragte die Perspektive von Heinz oder seiner Frau einnimmt) können die befragte Person zu einer ganz anderen Antwort veranlassen, ohne daß sich ihr moralisches Urteilsniveau verändert hat. Zwischen konkreten, existentiellen Entscheidungen und einem allgemeinen moralischen Prinzip kann es kaum eine einfache Beziehung geben, weil oft mehr als dieses eine Prinzip beachtet werden muß, und weil die Abwägung der Folgen ein notwendiger Bestandteil einer moralischen Entscheidung ist.

2. Auch Verhalten als *Verwendung oder Bewertung moralischer Argumente* ist keine adäquate Basis für die Erschließung moralischer Dispositionen. Wenn eine Person Argumente benutzt, die linguistisch als moralisch klassifiziert werden können (z.B. Ehrlichkeit, Gerechtigkeit), dann kann davon ausgegangen werden, daß die Person diese moralischen Begriffe kennt und wertschätzt. Aber das bloße Äußern oder Bewerten moralischer Argumente zeigt noch keine bestimmte moralische Urteilsfähigkeit an. Sagte ein Befragter beispielsweise: "Es ist falsch, wenn Heinz die Medizin stiehlt, um das Leben seiner Frau zu retten, *da es gegen das Gesetz ist*", kann dieses, wie Kohlberg (1985) feststellt, nicht eindeutig einer Stufe der moralischen Argumentation zugeordnet werden: "Der Gebrauch der Norm des Gesetzes ist auf jeder Stufe zu finden" (S. xvi). Das bloße Erwähnen von moralischen Normen besagt noch nicht, ob eine Person gewillt und in der Lage ist, in konkreten Entscheidungssituationen moralische Normen und Werte adäquat anzuwenden. Wenn jemand zu einer Begründung gedrängt wird, wird er vielleicht nur deshalb moralische Prinzipien benutzen, um seine Wahl nachträglich zu rechtfertigen oder zu "rationalisieren", wie ein Psychoanalytiker das bezeichnen würde.

3. Nur die Beobachtung von *ganzheitlichen Verhaltensmuster* kann uns in die Lage versetzen, die Existenz und Ausprägung moralischer Dispositionen einer Person valide zu messen. Gemäß der kognitiven Entwicklungstheorie besteht die konstitutive Komponente des moralischen Verhaltens einer Person nicht bloß darin, daß sie moralische Prinzipien anwendet, sondern auch und besonders darin, wie konsistent und differenziert sie sich bei ihren Entscheidungen daran orientiert. Für Kohlberg und seine Kollegen muß sich die Messung des moralischen Urteils daher im Kern auf "die Analyse beobacht-

barer Denkmuster [konzentrieren], wie sie sich in den Antworten einer Person im Moralinterview offenbaren" (S. 4).

3.4 Methode: Zerlegung des Verhaltens

Das Hauptziel der Revision des *Moral Judgment Interviews* war eine Verbesserung im Sinne der klassischen Testtheorie, die davon ausgeht, daß in jedem Verhaltensakt die zu messende Disposition repräsentiert ist. Dies versuchten die Autoren zu erreichen, indem "klare und konkrete Stufenkriterien spezifiziert und die Entwicklungsabfolge eines jeden moralischen Begriffs innerhalb jeder Stufe und global, für die allgemeinen Stufen-Strukturen definiert wurde" (Colby & Kohlberg 1987, S. 40).

Wie ist das Vorgehen beim MJI? Ähnlich Piagets (1973a) Vorgehen, wird im MJI die befragte Person mit einem moralischen Dilemma konfrontiert, d.h. mit einer kurzen Geschichte, in der sich zwei oder mehrere moralische Prinzipien gegenüberstehen. Die Person wird gebeten, eine Wahl zwischen diesen Prinzipien zu treffen. Indem die Entscheidung zwischen zwei Lösungen auch eine zwischen verschiedenen moralischen Prinzipien ist, stimuliert das Dilemma die Person, sich mit der moralischen Seite statt nur mit der "technischen" Seite eines Problems zu beschäftigen. Durch gezieltes Nachfragen, durch "Warum-Fragen", wird die befragte Person angehalten, ihre Meinung mit Argumenten zu begründen, und durch die Konfrontation der Person mit variierenden Rahmenbedingungen für das Dilemma, mit neuen Dilemmas und mit gegnerischen Meinungen und Sichtweisen. Unter anderem wird gefragt: "Weshalb hast Du dich für die eine Handlungsalternative entschieden? Was sind deine Gründe? Was rechtfertigt diese Entscheidung?" etc. Die Qualität dieser Nachfragen ist, worauf Colby, Kohlberg et al. (1987) zu Recht hinweisen, "folgeschwer für die Erfassung der moralischen Urteilsstufe" (S. 151). Das gezielte Nachfragen ist notwendig, um eine gute Auswertbarkeit der Antworten zu erreichen, um die am weitesten entwickelte Stufe des moralischen Argumentierens hervorzulocken, zu der die Person fähig ist, und um überprüfen zu können, in welchem Maße die Person fähig ist, bei ihren Entscheidungen moralische Prinzipien anzuwenden (S. 5, 58, 61, 169).

Trotz der großen Bedeutung, denen die Autoren der Interviewsituation zu-messen, wird diese im Handbuch nicht systematisch behandelt. Sie basiert wei-

terhin größtenteils auf Intuition und Alltagsverständnis (S. 153-158). Die Instruktionen bleiben oft vage. Beispielsweise wird der Interviewer dazu angehalten, gezielt nachzufragen, solle jedoch nicht zu oft "warum"-Fragen stellen (S. 54). Die bei jedem Dilemma gestellten Standardfragen berühren systematisch die zwei Hauptthemen (Issues) des jeweiligen Dilemmas. Allerdings bleibt es dem Interviewer überlassen, die im Dilemma enthaltenen Normkonflikte auszuloten.⁷ Noch bedauerlicher ist, daß die Konfrontation mit Gegenargumenten im neuen MJI kaum mehr angesprochen wird. Während Kohlberg (1958) früher mit dieser Technik die Festigkeit der moralischen Überzeugungen der Befragten getestet hatte, dient sie heute nur noch dazu, ausreichende Mengen an linguistischem Material für die Auswertung zu gewinnen (S. 151). Argumente, die von den Probanden zugunsten der Gegenmeinung vorgebracht werden, erhalten – sofern sie überhaupt noch in die Einstufung der individuellen Urteilsstruktur einbezogen werden – im neuen Auswertungssystem eine geringere Gewichtung als die Argumente, die sie zur Begründung ihrer eigenen Meinung vorbringen.

Die Last der Identifikation der strukturellen Eigenschaft des individuellen Urteilsverhaltens liegt beim MJI somit ganz auf der Auswertungsseite, die daher einer eingehenden Betrachtung bedarf. Die Auswertung der Interviewurteile erfolgt in drei Schritten. Erster Schritt: Identifizierung der adäquaten Maßeinheit. Zweiter Schritt: Zuordnung eines Stufenscores. Dritter Schritt: Kalkulation eines globalen und durchschnittlichen Scores für das Individuum über alle Maßeinheiten hinweg. Wie werden die Maßeinheiten identifiziert? Man muß laut Handbuch zunächst bei einem Issue (z.B. dem Issue "Leben" im Heinz-Dilemma) anfangen und alle Antworten gemäß der moralischen *Norm* und dem moralischen *Element* klassifizieren.⁸ Die daraus resultierende Issue x Norm x Element-Kombination, von den Autoren auch als "Interview judgment" (IJ) bezeichnet, dient beim MJI als Grundlage für die hermeneutische Analyse. IJs werden nach den Issues klassifiziert, denen sie zugeordnet werden. Die "Rules for issue classification" (S. 163-165) liefern die tatsäch-

7 Colby, Kohlberg et al. 1987, Band 2, S. 4 und 56.

8 Das Handbuch erläutert eingehend, was unter Normen und Elementen verstanden wird, und gibt für beides Beispiele.

lichen Maßeinheiten. Sind die Maßeinheiten identifiziert, können sie durch das sogenannte "matching"-Verfahren einer Kohlberg-Stufe zugeordnet werden, indem man die Antworten der befragten Person jener Stufe zuordnet, für die im Handbuch die ähnlichsten Beispielantworten zu finden sind. Das Handbuch gibt für dieses *matching* fünf Entscheidungshilfen: 1. Für die vorläufige, globale Einschätzung aufgrund aller Antworten (Handbucharweisung: "Lies erneut alle Antworten durch, die unter einem Thema eingeordnet wurden, und denke darüber nach, was sie bedeuten"). 2. Für die Frage der Auswertbarkeit ("Liefere die *interview judgments* Argumente, hält sie der Befragte für valide und sind sie präskriptiver Natur?"). 3. Für die Einschätzung der Wertzuweisung und -transposition ("Rekonstruiere die Bedeutung von verstümmelten Worten und zerstückelten Sätzen"). 4. Für die Verteilung der IJs auf die *criterion judgments* (CJs) der richtigen Stufen. 5. Schließlich für die "strukturelle Bewertung" der vorgeschlagenen Zuordnungen, mit der "jene oberflächlichen Zuordnungen ausgemerzt oder mit einem 'Veto' versehen werden, für welche die Stufe des *criterion judgments* nicht die wahre strukturelle Bedeutung des Interviewmaterials widerspiegelt" (Colby, Kohlberg et al. 1987, S. 165-177).

Kohlberg und seine Koautoren werten die Interviewantworten im MJJ also primär nach den Vorgaben der klassischen Testtheorie aus, indem sie das gesamte Interview in Maßeinheiten zergliedern, diese Einheiten isoliert einstufen und sie dann zu einem Gesamtwert addieren. Diese zergliedernde Analyse versuchen sie, durch eine hermeneutische ("wahre strukturelle Bedeutung") abzusichern. Es ist aber fraglich, ob die hermeneutische Analyse dieses leisten kann, zumal in diesen Teil der Auswertung vielfältige theoretische Annahmen des Auswerters einfließen können, die die Neutralität der Daten bezüglich der zentralen Postulate der Kohlberg-Theorie in Frage stellen.

3.5 Das methodologische Dilemma des MJJ

Hierin besteht das methodologische Dilemma des revidierten MJJ. Auf der einen Seite soll das MJJ strukturelle Eigenschaften des individuellen Urteilsverhaltens (wie Inkonsistenz und Differenziertheit) messen, auf der anderen Seite werden (gemäß der klassischen Testtheorie, an der sich die Revision des MJJ orientiert hat) solche strukturellen Informationen zur

Bewertung des Meßinstruments benötigt, so daß sie nicht mehr für die Einstufung der Urteilsstruktur zur Verfügung stehen. Es wird anerkannt, daß es zwischen den Antwortmustern von Befragten strukturelle Unterschiede gibt, aber diese Unterschiede werden nicht als Basis für eine strukturelle Einschätzung der individuellen Urteilsfähigkeit herangezogen. So wird zum Beispiel die Tatsache, daß einige Personen für nicht gewählte Issues weniger Argumente oder niedrigere Stufenwerte produzieren als für die gewählten (S. 161), durch die "Issue classification rules" von der Analyse ausgeschlossen. Unterschiede innerhalb eines Issues (Antwortinkonsistenzen) werden durch die "Lower-stage inclusion rules" weitgehend nivelliert. Die Konstrukteure gehen zwar von der Annahme aus, daß das moralische Urteilen eine strukturelle Einheit bildet, also keine wesentliche Variation zwischen den Urteilen einer Person auftritt. Aber in einer Studie, so berichten Colby, Kohlberg et al. (1987), konnten mit dem MJI-Verfahren nur 29 von 188 Personen einer "reinen" Stufe der moralischen Urteilskompetenz zugeordnet werden (S. 136-139).⁹ Mischstufen stellen ein weiteres Problem für den Versuch dar, klassische Testtheorie und kognitive Entwicklungstheorie miteinander zu verbinden. Auf der einen Seite, könnte man argumentieren, sind gemischte Stufenklassifikationen durch unzureichende Reliabilität („Meßfehler“) bedingt. Wenn Stufen der moralischen Urteilskompetenz strukturelle Einheiten darstellen, dann sollte eine Person für jedes Issue oder jedes Dilemma den gleichen Score erhalten. Die Annahme der unzureichenden Reliabilität würde zwar die Hypothese der strukturellen Einheit retten, doch würden damit die häufig berichteten Teilstufen (und die Unterschiede, die kleiner als eine Stufe sind) bedeutungslos. Auf der anderen Seite können wir Mischstufen als Stufenübergänge begreifen. In diesem Falle wäre aber Kohlbergs Annahme der strukturellen Ganzheit nicht mehr haltbar. Die Idee der strukturellen Ganzheit verträgt sich nicht mit dem Konzept von Mischstufen. Das bestätigt wieder einmal, daß man den sprichwörtlichen Kuchen nicht gleichzeitig essen und behalten kann. Die beobachtete Konsistenz (oder Varianz) der Forschungsdaten läßt sich nicht gleichzeitig dem Individuum und dem Meßinstrument zu-

9 Meine Berechnung. Da keine Angaben über die Berechnungsgrundlagen der "Stage Usage Percentage" gemacht werden, habe ich die "Issue-Einheit" verwendet.

ordnen. Man kann nicht versuchen, die Struktur des individuellen Urteilsverhaltens mit einer Methodologie zu erfassen, die diese Struktur als Fehler des Meßinstruments betrachtet. Entweder nimmt man an, die intra-individuelle Verhaltensvariation sei ein Ausdruck des Meßfehlers des Instruments, oder sie sei ein Ausdruck bestimmter kognitiver Strukturen des Befragten. Die klassische Testtheorie entschied sich für die erste Annahme (Lumsden 1976). Die kognitive Entwicklungstheorie legt eher die zweite Betrachtungsweise nahe, da sonst keine Möglichkeit bestünde, Urteilsstrukturen anhand von objektiven Beobachtungsdaten nachzuweisen.

3.6 Resümee

Das *Moral Judgment Interview* (Colby, Kohlberg et al. 1987) ist eine einzigartige Interview-Methode, und dennoch ist es vom Standpunkt der kognitiv-strukturellen Theorie der Moralentwicklung zu kritisieren. Seine theoretischen Grundlagen und seine Auswertungsregeln sind sehr transparent, was diese Methode vielen anderen Interview-Methoden überlegen macht, die allein auf globalen Einschätzungen beruhen. Erst diese Transparenz der MJI-Methode ermöglicht eine differenzierte Kritik des Verfahrens, aus der sich konstruktive Hinweise für eine Verbesserung ergeben. Auch ist das MJI gemessen an den Kriterien der klassischen Testtheorie hoch reliabel und daher geeignet für psychologische Forschungs- und Diagnostikfragen, die eine hohe Reliabilität im Sinne dieser Testtheorie erfordern. Aber gerade wegen der Ausrichtung des MJI an der klassischen Testtheorie ist die Validität des MJI eingeschränkt. Die auf Zergliederung des individuellen Verhaltens angelegte Auswertung des MJI macht die Struktur der individuellen Urteilmuster für den Forscher *unsichtbar*.

Da das Auswertungsverfahren des MJT auf der Klassischen Testtheorie basiert, basiert es (implizit) auch auf der Annahme, daß die Probanden sich bei der Beurteilungen eines moralischen Dilemmas *alle gleichermaßen* an moralischen Normen (wenn auch auf verschiedenen Stufen) orientieren. Gleichzeitig erlaubt das Verfahren es nicht, diese Annahme empirisch zu prüfen. Dennoch gibt es, wie wir in den nachfolgenden Kapitel zeigen werden, vielfältige Hinweise dafür, daß diese Annahme falsch ist, das heißt, daß sich die Menschen sehr stark unterscheiden können hinsichtlich der Frage, ob und wie

stark moralische Dispositionen ihr Urteilsverhalten determinieren. Art und Grad dieser Unterschiede lassen sich sogar aufgrund theoretischer Überlegungen vorhersagen. Es scheint, daß die Auswertung des MJI auf Annahmen beruht, die viel mehr mit klassischen Testtheorie als mit der kognitiv-strukturellen Theorie der Moralentwicklung in Einklang stehen.

Für die Überprüfung solcher Annahmen benötigen wir also ein Instrument, das ohne diese Annahmen auskommt und die valide Messung von Inhalt *und* Struktur des moralischen Urteilsverhaltens ermöglicht. Man muß damit vor allem messen können, *ob* und wie stark eine befragte Person sich in der Diskussion über Probleme an moralischen Prinzipien orientiert. Das *Ausmaß der Determination des individuellen Urteilsverhaltens durch moralische Orientierungen* ist der entscheidende Gradmesser für die individuelle moralische Urteilsfähigkeit. Die Erfassung dieser Fähigkeit ist die denn auch die zentrale Meßabsicht des *Moralisches Urteil-Tests*, der im nächsten Kapitel vorgestellt wird.

Colby, Kohlberg et al. (1987) sehen die Validität des MJI durch die Tatsache bestätigt, daß die MJI-Daten die Hypothesen der invarianten Entwicklungssequenz und der strukturellen Ganzheit bestätigen. Wie wir gesehen haben, wirft diese Feststellung einige schwer lösbare Probleme auf, da das MJI aufgrund seines Auswertungsdesigns *voreingenommen* (biased) gegenüber der Prüfung dieser Hypothesen ist. Die Auswertungsanleitung des MJI setzt an vielen Stellen die Richtigkeit der Annahme einer invarianten Entwicklungsabfolge und strukturellen Ganzheit voraus. Die Einstufung der Probanden mit dem MJI soll laut Kohlberg (1976) so erfolgen, "daß das Kriterium der aufeinanderfolgenden Bewegung eingehalten wird" (S. 47). Somit begünstigen die mit dem MJI gewonnenen Daten in einem schwer abschätzbaren Ausmaß bereits *a priori* die Bestätigung beider Postulate. Aber auch wenn man annehmen könnte, daß die Messung des moralischen Urteilens mit dem MJI unverzerrt und unvoreingenommen ist, verbleibt das Problem, daß die Erfüllung der beiden von Kohlberg genannten Kriterien (invariante Sequenz und strukturelle Ganzheit) keine hinreichende Bedingung für die Validität dieses Meßverfahrens darstellen. Es gibt viele Maße (z.B. Gewicht und Höhe, mathematische Fähigkeiten), die beide Kriterien erfüllen, aber dennoch keine validen Maße des moralischen Urteils sind. Damit ein Meßinstrument die *moralische* Urteilsfähigkeit valide erfassen kann, muß die darin gestellte Aufgabe für die

Befragten tatsächlich eine *moralische* Aufgabe sein. Die Kohlberg-Gruppe hält dies im Falle des MJI offensichtlich für gegeben und widmet sich dieser Frage daher nicht explizit.

Implizit lassen sich beim MJI die darin enthaltene *moralische Aufgabe* und die für ihre Bewältigung notwendige Fähigkeit rekonstruieren. Bei der Diskussion und Lösung eines moralischen Dilemmas müssen die Befragten vielfältige Erwartungen und Werte koordinieren. Sie müssen sich für eine von zwei Handlungsalternativen entscheiden, die beide moralische Probleme aufwerfen: Hatte Heinz, der Protagonist im gleichnamigen Dilemma im Kohlberg-Interview, die Medizin stehlen dürfen, um seine todkranke Frau zu retten, oder durfte er keinen Diebstahl begehen? Hätte er es auch für einen anderen Menschen tun müssen? Ist diese Rechtfertigung generalisierbar? Was wäre, die befragte Person selbst die Bestohlene wäre? Offenbar sind diese Fragen nicht leicht zu beantworten. Moralisch wenig entwickelte Personen kommen wohl kaum über die ersten Fragen des Interviews hinaus. Viele finden es schwierig, wenn nicht gar unmöglich, sich zum Beispiel auf eine rationale Diskussion über Ansichten einzulassen, die ihrer eigenen Meinung widersprechen. In diesem Sinn stellt das *Moral Judgment Interview* von Kohlberg offenbar eine Aufgabe für die befragte Person dar. Aber im Auswertungshandbuch wird dieses für die Validität wichtige Thema an keiner Stelle angesprochen. So bleiben viele Fragen unbeantwortet. Welche Überlegungen haben zur Auswahl der Dilemmas geführt, die das MJI enthält? Erfordern alle MJI-Dilemmas eine Begründung auf der im Sinne von Kohlberg höchsten Stufe des moralischen Urteils, oder lassen sich einige der dort vorgelegten moralischen Probleme auch einer niedrigeren Stufe adäquat lösen bzw. wäre es adäquater, sie auf einer tieferen Stufe zu lösen zu versuchen? Einige Dilemmas im MJI scheinen durchaus auf einer tieferen Stufe als auf Stufe 6 adäquat lösbar. Schließlich die Frage: Worin besteht moralische Urteilsfähigkeit? In der Wahl einer möglichst hohen Stufe zur Rechtfertigung einer einmal gefaßten Meinung? Oder besteht sie nicht eher in der Fähigkeit, die eigene Meinung abzuändern, wenn man erkennt, daß nicht in Einklang steht mit den eigenen moralischen Überzeugungen?

4 Der *Moralisches Urteil-Test* - Ein experimenteller Fragebogen

Um die empirische Gültigkeit der verschiedenen Theorien der Moralentwicklung prüfen zu können, benötigen wir valide Beobachtungen der moralischen Urteilsfähigkeit von Menschen. Damit es valide ist, muß das Beobachtungs- oder Meßinstrument theorie-kompatibel, objektiv (replizierbar) und leistungsabfordernd sein. Es muß *theorie-kompatibel* sein, damit die Meßwerte überhaupt als Kriterium für die Richtigkeit oder Falschheit von Vorhersagen dienen können. Daten von einem invaliden Meßinstrument besitzen für die Prüfung einer Theorie keinen Aussagewert. Im Rahmen unserer Untersuchung heißt die Forderung nach Validität vor allem, daß das Konstruktions- und Auswertungsdesign des Meßinstruments die Erfassung der moralischen Urteilsfähigkeit genau in der Weise garantiert, wie sie definiert ist. Wie wir in Kapitel 3 gesehen haben, ist diese Forderung insbesondere in bezug auf die kognitiv-strukturellen Eigenschaften des moralischen Urteilsverhaltens offenbar kaum oder gar nicht einzulösen, wenn man sich auf klassische psychologische Meßmethoden stützt.

Objektiv muß das Meßinstrument sein, damit die hier berichteten Ergebnisse auch von anderen Auswertern und bei anderen Befragten überprüft werden können. Damit die individuelle moralische Urteilsfähigkeit auf eine intersubjektiv überprüfbare Weise festgestellt werden kann, sollte das Instrument Meßwerte liefern, ohne daß dafür intuitive Einschätzungen durch die befragte Person selbst oder durch die Auswerter notwendig sind. Die Beteiligung von Schätzern ist kein prinzipieller Einwand gegen die Validität eines Meßinstruments; selbst in den Naturwissenschaften werden Meßwerte häufig durch geschulte Schätzer ermittelt. Aber das Angewiesensein auf Schätzer stellt hohe Anforderungen an die Klarheit der Auswertungsregeln und an die Kompetenz der Schätzer und sie begrenzt aus ökonomischen Gründen sehr stark die Überprüfungsmöglichkeiten durch die Forschung. Die Beteiligung der Befragten bei der Einschätzung ihres eigenen Verhaltens hingegen stellt eine prinzipielle Einschränkung der Meßvalidität dar, da deren Selbstwahrnehmungsstil mit der zu messenden moralischen Urteilsfähigkeit mehr oder minder hoch korrelieren

kann, und wir daher nicht wissen, in welchem Ausmaß der Meßwert diese Fähigkeit widerspiegelt oder die Art der Selbstwahrnehmung. Auf den MJI trifft nur der erste Einwand zur Objektivität des Verfahrens zu. Der zweite Einwand trifft auf Verfahren zu, die von den Befragten verlangen, daß er oder sie ihr eigenes moralisches Urteil einschätzen sollen ("Wie würden *Sie* selbst in einer solchen Situation urteilen?"). Die hier verlangte Selbsteinschätzung ist, wie wir aus der psychologischen Forschung wissen, in erheblichem Ausmaß durch "response sets", soziale Erwünschtheit und "Halo-Effekte" beeinflusst.

Mit dem Kriterium "*leistungsabfordernd*" soll sichergestellt werden, daß das Meßinstrument nicht nur irgendeine kognitive Orientierung erfaßt, sondern eine Fähigkeit, nämlich moralische Urteilsfähigkeit. Damit die Bearbeitung eines Tests eine Leistung darstellt, muß dieser eine Aufgabe enthalten, deren Bewältigung eine Fähigkeit erfordert. Im Unterschied zu gewöhnlichen Leistungstests, bei denen die Richtigkeit der Lösung zumeist aufgrund objektiver Merkmale von außen festgelegt werden kann, haben wir es bei moralischen Dilemmas mit Aufgaben zu tun, für die es keine "objektive", von außen festlegbare Lösung gibt (Pittel & Mendelsohn 1966). Kohlberg (1971) hat daher vorgeschlagen, die Güte des individuellen Urteils nicht an der gewählten Entscheidung zu messen, sondern an der Qualität der gegebenen Argumente, und hat dafür gewichtige philosophische Gründe angeführt. Im MJI wird daher festgelegt, daß je höher das Argumentationsniveau einer Person ist, desto größer ist ihre moralische Urteilsfähigkeit. Diese Festlegung ist aber auch von außen, der befragten Person also external, und sie ist zu rigide, um kulturelle Unterschiede bei der Definition einer adäquaten Argumentationsweise und um Unterschiede in der Anforderungsstruktur verschiedener Dilemmatypen gerecht werden zu können. Dilemmas, bei denen in unserem Kulturkreis eine moralische Diskussion auf Stufe 6 des Kohlberg-Schemas erwartet wird, damit als adäquat behandelt gelten können, stellen in einer anderen Kultur vielleicht gar keine Stufe-6-Dilemmas dar, und umgekehrt (siehe Kapitel 7). Auch ist zu erwarten, daß manche der zur Messung verwendeten Dilemmas auf tieferen Moral-Stufen adäquat diskutiert werden können.

Um diesen Einwänden gerecht zu werden, definieren wir im MUT die moralische Urteilsfähigkeit einer Person ohne Berücksichtigung der von ihr gewählten Stufe der Argumentation als das Ausmaß, mit dem sie sich bei ihren Begründungen (oder bei der Bewertung von Begründungen) *konsistent* an

moralischen Prinzipien - anstatt zum Beispiel an Meinungskonformität - auf einer beliebigen Stufe orientiert. Bei dieser Definition ist das Stufenkriterium nicht mehr extern bestimmt - die befragte Person legt selbst fest, welche Stufe des moralischen Urteils sie zur Lösung des Dilemmas heranzieht. Extern definiert ist allein noch das Kriterium der Urteilskonsistenz, das allein in der psychologischen Meßabsicht der Untersuchung begründet ist, aber nicht in philosophischen oder kulturellen Festlegungen. Diese Definition stellt eine Präzisierung von Kohlbergs (1964, S. 425) Definition der moralischen Urteilsfähigkeit dar, der sie als das "Vermögen" bestimmte, "*moralische Entscheidungen und Urteile zu treffen, die moralisch sind (also auf moralischen Prinzipien gründen), und in Übereinstimmung mit diesen Urteilen zu handeln.*"

4.1 Konsistenz: Meßfehler oder Strukturmerkmal?

Die Konstruktion eines validen Instruments zur Messung der moralischen Urteilsfähigkeit setzt, wie in Kapitel 3 deutlich wurde, die Lösung eines methodologischen Dilemmas voraus (siehe oben, S. 74). Zur Lösung dieses Dilemmas ist einerseits die Frage zu beantworten, wie relationale Eigenschaften der individuellen Urteilsstruktur zu interpretieren sind: Ist Urteilskonsistenz oder -inkonsistenz, wie die klassische Testtheorie voraussetzt, als ein Merkmal des Meßinstruments (als "Reliabilität" oder "Meßfehler") zu interpretieren oder, wie aufgrund der kognitiven Entwicklungstheorie angenommen werden müßte, als ein Merkmal der befragten Person? Zum anderen muß, falls die zweite Interpretation zutreffend ist, eine Forschungsmethode gefunden werden, die eine Messung von relationalen Eigenschaften des moralischen Urteilsverhaltens ermöglicht. Wenden wir uns zunächst der ersten Frage zu.

Wie bereits in Kapitel 3 argumentiert wurde, besteht für die Interpretation von Konsistenz des moralischen Urteilsverhaltens als Eigenschaft des Meßinstruments aus kognitions-psychologischer Sicht wenig Plausibilität. Die Piaget-Kohlberg-Theorie legt vielmehr eine Deutung formaler Antwortmerkmale als Manifestationen der Organisation oder Struktur des moralischen Urteils nahe. Wir brauchen uns bei dieser Frage aber nicht mit Plausibilitätsüberlegungen zufrieden geben. Aus der kognitiv-strukturellen Sichtweise folgen testbare Hypothesen, die, sofern sie bestätigt werden können, die impliziten psycholo-

gischen Voraussetzungen der klassischen Testtheorie auch empirisch widerlegen würden.

Erste Hypothese: Menschen variieren nicht nur bezüglich ihrer moralischen Einstellungen, sondern auch hinsichtlich des Ausmaßes, mit dem diese ihr Urteilsverhalten bestimmen. Das heißt, das Urteilsverhalten von Menschen unterscheidet sich nicht nur hinsichtlich der – positiven oder negativen – Einstellung zu der moralischen Orientierung, die gemessen werden soll, es unterscheidet sich auch danach, wie stark diese Einstellung überhaupt das Urteilsverhalten determiniert und welche anderen moralischen oder außermoralischen Einstellungen es determinieren. Das Urteilsverhalten unterscheidet sich demnach auch hinsichtlich formaler Eigenschaften wie Konsistenz und Inkonsistenz, Integriertheit und Differenziertheit. Es bestehen zwischen ihnen nicht nur, wie die klassische Meßtheorie annimmt, Richtungs- und Ausprägungsunterschiede, sondern auch Strukturunterschiede, die sich empirisch nachweisen lassen.

Zweite Hypothese: Insofern es solche Strukturunterschiede gibt, stehen sie in einem engen Zusammenhang mit der Komplexität der anzuwendenden moralischen Orientierung und damit zur Schwierigkeit des zu lösenden moralischen Dilemmas (siehe Kapitel 2).

Dritte Hypothese: Diese Strukturunterschiede spiegeln strukturelle Entwicklungsunterschiede wieder, d.h. es sind Unterschiede, die in der kognitiv-moralischen Struktur des Individuums begründet liegen und durch geeignete Maßnahmen, wie beispielsweise durch schulische Allgemeinbildung, beeinflusst werden können.

Wir haben diese Hypothesen mit dem *Moralisches Urteil-Test* umfassend überprüft.¹⁰ Ohne diese Befunde im einzelnen zu diskutieren, kann man feststellen, daß sie die Interpretation der kognitiven Entwicklungstheorie in allen Punkten bestätigen. Unsere Studien zeigen erstens eine hohe Varianz der individuellen Urteilsvarianz (Lind 1985a), zweitens einen systematischen Zusammenhang zwischen der individuellen Urteilsvarianz und der Höhe der sogenannten "Test-Reliabilität" oder "internen Konsistenz", einen starken systematischen Zusammenhang zwischen der Höhe der Moralstufe im Sinne der

10 Siehe Lind 1978, 1985a; Lind & Wakenhut 1985; dazu auch Kapitel 11.

Kohlbergschen Moralstufen und der Urteilskonsistenz (s. Kapitel 2) und drittens einen systematischen und in allen Untersuchungen bestätigten Zusammenhang zwischen der Höhe der Urteilskonsistenz und dem schulischen Bildungsniveau (Kapitel 11). Mit anderen Worten: diese Befunde widersprechen durchweg der Annahme, daß Konsistenz und Varianz des Urteilsverhaltens ein Merkmal der Messung (also bloß ein Ausdruck des Meßfehlers) sind. Vielmehr bestätigen sie die Theorie, daß die individuelle Urteilsvarianz ein Strukturmerkmal der untersuchten Person widerspiegelt. Damit ist auch die Annahme bekräftigt, wonach moralische Dispositionen keineswegs bloß hypothetisch oder überflüssig sind, sondern konkret und empirisch überprüfbar.

4.2 Die Meßintention des MUT

Die Meßintention des MUT basiert auf der Definition der moralischen Urteilsfähigkeit als *das Vermögen, moralische Entscheidungen auf der Grundlage von (selbstgewählten) moralischen Prinzipien zu treffen (statt z.B. auf Vorurteilen oder Konformität), und in bezug auf diese Prinzipien konsistent und unparteiisch zu handeln*. Mit dem MUT sollen die beiden in dieser Definition angesprochenen Aspekten gemessen werden: erstens an welchen Prinzipien bzw. Moralstufen eine Person ihr Urteil orientiert (=affektiver Aspekt) und zweitens wie stark ihre Argumentationen bzw. Bewertungen von Argumenten von den selbstgewählten Prinzipien determiniert sind (=kognitiver Aspekt). Die *affektiven Aspekte* des moralischen Urteils sind die individuellen moralischen Einstellungen, das heißt die moralischen Prinzipien, an denen der einzelne sein Urteilsverhalten ausrichtet. Wie beim MJI konzentriert sich auch beim MUT die Meßabsicht auf die sechs Stufen moralischer Orientierungen oder Perspektiven, die Kohlberg vorgeschlagen hat und die wir oben, im Kapitel 2 ausführlich dargestellt haben.¹¹ Bezüglich der kognitiven Aspekte, die mit dem MUT gemessen werden können (Lind 1978), steht das Ausmaß

11 Kohlberg hat zwar in der revidierten Fassung des MJI darauf verzichtet, die Stufe 6 im Auswertungshandbuch (Colby, Kohlberg et al. 1987) aufzuführen, aber nicht darauf, sich mit ihr als eine realistische Perspektive für die Lösung eines moralischen Dilemmas auseinanderzusetzen (Kohlberg et al. 1986). Auch wenn es nicht viele Menschen geben sollte, die alle präsentierten Dilemmas auf Stufe 6 adäquat lösen können, so gibt es schon lange bekannt, daß viele eine positive Einstellung zu dieser Stufe des moralischen Argumentierens haben (Rest 1973).

der Determination des Urteilsverhaltens durch moralische Prinzipien im Mittelpunkt. Die Determination kommt in der Konsistenz des Verhaltens *in bezug* auf diese Prinzipien zum Ausdruck: wenn sich eine Person bei der Beurteilung von Argumenten *konsistent* an der moralischen Qualität dieser Argumente orientiert, dann sagen wir, sie zeigt eine hohe moralische Urteilsfähigkeit. Wenn sie sich statt dessen bei der Bewertung von vorgetragenen Argumenten konsistent an der Übereinstimmung dieser Argumente mit ihrer eigenen (vorgefaßten) Meinung orientiert, dann erreicht sie nur einen niedrigen Urteilsfähigkeitswert im MUT.

Der MUT mißt beide Aspekte *gleichzeitig*, ohne sie jedoch zu vermengen, wie es das MJI und alle anderen, bisher bekannten Verfahren (der DIT von Rest, das SRM von Gibbs u.a.m.) tun. Der affektive und der kognitive Aspekt des moralischen Urteilens sind logisch unterscheidbar, aber nicht dinglich trennbar. Das heißt, sie können nicht an verschiedenen Verhaltensweisen in verschiedenen Situationen beobachtet werden; sie können nur gleichzeitig, an ein und demselben Verhaltens- oder Urteilmuster erfaßt werden. Sie sind eben zwei *Aspekte* oder Eigenschaften des individuellen Urteilens und keine *Komponenten* (s. Kapitel 2).

Für den kognitiven Aspekt allein reservieren wir den Begriff der *moralischen Urteilsfähigkeit*, da es wohl ohne Anstrengung oder Erfahrung möglich ist, hohe Stufen des moralischen Urteilens zu präferieren oder zu benutzen (weshalb der affektive Aspekt *keine* Fähigkeit darstellt), aber wohl nicht, sich konsistent an moralischen Prinzipien zu orientieren. Allerdings ist die Konsistenz des Urteils allein kein hinreichender Indikator für moralische Urteilsfähigkeit. Zum einen muß das Urteilsverhalten konsistent *in bezug* auf moralische Prinzipien sein, um als Ausdruck *moralischer* Urteilsfähigkeit gelten zu können. Einem Urteilsverhalten, das konsistent in bezug auf andere Dispositionen ist wie zum Beispiel in bezug auf die eigene Meinung zur Lösung eines Dilemmas, kann ein das Prädikat *moralisch* offensichtlich nicht zugesprochen werden. Zum anderen muß das Meßexperiment aus einer Aufgabe bestehen, da wir sonst nicht in der Lage sind, Konsistenz als Ausdruck einer *Fähigkeit* zu werten. Wie beim MJI und anderen Verfahren wird die oder der Befragte mit mehreren (im Standard-MUT mit zwei) Dilemmas konfrontiert. Hierdurch kann festgestellt werden, ob eine Person moralisch konsistent urteilt, das heißt

immer wieder die Argumente einer bestimmten Stufe gebraucht oder akzeptiert, um ihre Meinung zu begründen bzw. abzustützen. Aber diese Konsistenz kann auch Ausdruck von Gewohnheit, Habitus, Rigidität oder ähnlichen sein. Damit sich beim Bearbeiten des MUT auch die Fähigkeit zum moralischen Urteilen erweisen kann, wird bei diesem Meßverfahren die befragte Person mit Argumenten konfrontiert, die zu ihrer eigenen Meinung konträr sind, d.h. die ihrer Meinung widersprechen.

Wenn eine Person eine moralische Position akzeptiert oder übernimmt, die ihrer eigenen vorgefaßten Meinung widerspricht, und diese Meinung im Lichte dieser Position neu bewertet und möglicherweise sogar revidiert, so hat sie zumeist unsere persönliche Hochschätzung. Das ist so, weil, wie der Philosoph Wellmer (1986), es formuliert, wir in der Art, wie Personen zu Argumenten stehen, die gegen ihre (vorgefaßte) Meinung sprechen, ein Zeichen für den Grad ihrer *Vernunftgeleitetheit* sehen. Es entspricht offenbar unserer allgemeinen Erfahrung, daß Menschen gewöhnlich dazu tendieren, an ihrer vorgefaßten Meinung festzuhalten und möglichst viele Gründe dafür zu suchen, die ein solches Festhalten rechtfertigen können, daß sie nur wenig fähig sind, sich mit Einwänden auseinanderzusetzen, die diese Meinung in Frage stellen, und noch weniger, sich von einer Gegenmeinung überzeugen zu lassen. Wer dagegen dazu in der Lage ist, den schätzen wir hoch, weil solche Personen damit zeigen, daß sie moralische Argumente nicht benutzen, um vorgefaßte Meinungen nachträglich zu begründen, sondern daß sie ihr Urteil wirklich an moralischen Prinzipien auszurichten vermögen.

Wir sagen "vermögen", da diese Personen offenbar eine Fähigkeit besitzen, die andere nicht besitzen, nämlich die Fähigkeit, das Verhalten durch Vernunft und Überlegung leiten zu lassen, statt bloß von Intuition und Gewohnheit. Diese Fähigkeit schätzen wir vor allem deshalb, weil bei Meinungskonflikten nur durch sie eine vernünftige, diskursive, gewaltfreie Einigung ermöglicht wird. Wenn sie fehlt, läßt sich ein Konflikt nur gewaltsam, das heißt durch einen Kampf lösen, bei dem eine Meinung dominiert und die andere unterdrückt wird, ohne daß geklärt ist, welche Meinung nun die moralisch richtigere oder vernünftigere ist. Selten wird dieser Kampf auf Meinungen beschränkt, sondern bezieht auch die Vertreter dieser Meinung mit ein, so daß aus der Unfähigkeit zum vernunftgeleiteten Verhalten nicht selten gewaltsame

Auseinandersetzungen mit dem Ziel der Unterdrückung oder Vernichtung anderer Menschen resultieren. Eine humanitäre, demokratische Gesellschaft ist daher auf dem Prinzip der wechselseitigen Achtung und der friedlichen Konfliktlösung gegründet, die von jedem Bürger verlangt, daß er oder sie versucht, gegensätzliche Meinungen über moralische, soziale oder politische Probleme durch vernünftige Überlegung und demokratischen Diskurs einander anzugleichen.¹²

Daß wir berechtigt sind, hier von einer besonderen Fähigkeit zu sprechen, ergibt sich bereits aus der bisherigen Forschung. So hat bereits Keasey (1974) empirisch belegt, daß die "Meinungsübereinstimmung" von Argumenten für das faktische Urteilsverhalten vieler Menschen eine überragende Bedeutung hat, und auch, daß die Entwicklung der Fähigkeit, die eigene Meinung im Lichte moralischer Prinzipien zu revidieren, offenbar an Entwicklungsprozesse gebunden ist. Meinungskonformität wird auch von Döbert und Nunner-Winkler (1978) als wesentlicher Faktor des psychischen "Apparats" bezeichnet. Bereits Piaget (1928) hat in seinem klinischen Interview seine Probanden mit Argumenten konfrontiert, die gegen deren jeweilige Meinung standen, um dadurch die Festigkeit ihrer kognitiven Struktur zu prüfen. Auch Kohlberg (1958) bediente sich anfänglich dieser Technik (*probing*), um den Grad der moralischen Determination des Urteilsverhaltens durch moralische Prinzipien zu prüfen (vgl. Eckensberger & Reinshagen 1980), auch wenn sie in der revidierten Fassung von Kohlbergs MJI (s. Kapitel 3) diese Funktion nicht mehr hat.

4.3 Experimentelle Fragebogen zur Erfassung psychologischer Dispositionen

Im MUT spielen *meinungskonträre* Argumente die Rolle eines besonderen Kriteriums für die Frage, ob eine befragte Person sich tatsächlich an moralischen Prinzipien orientiert oder ob sie diese nur benutzt, um die eigene Meinung

12 Ein wesentliches Element einer auf diesen Einsichten fußenden Demokratie-Erziehung ist daher die systematische Förderung der Fähigkeit, in Dilemma-Situationen Argumente der Gegenseite zu verstehen und sich mit ihnen auseinander zu setzen. Dabei ist es wichtig, wie Döbert (1987, S. 509) schreibt, "darauf zu insistieren, daß gerade Argumente, die der eigenen Position nicht entsprechen, stark gemacht und verteidigt werden". Vgl. auch Kohlberg 1987, Lind 1987, Oser und Althof 1992; auch Kapitel 8.

nachträglich zu stützen. Anlage und Auswertung des MUT sind so konzipiert, daß nur diejenige Person einen *hohen* Wert auf der Skala für moralische Urteilsfähigkeit erhält, die Argumente auch dann konsistent nach ihrer *moralischen* Qualität beurteilt, wenn diese der eigenen Meinung zur Lösung eines Dilemmas widersprechen. So ist die Absicht, aber wie wird das Ausmaß der moralischen Determiniertheit des Urteilsverhaltens im MUT konkret gemessen?

Um diesen kognitiven Aspekt des individuellen Urteilsverhaltens unter variierenden Bedingungen (Dilemmatyp, Meinungskonformität) messen zu können, haben wir uns des Konzepts des multivariaten Experiments (statt der klassischen Testtheorie) bedient. Um hervorzuheben, daß es sich dabei nicht um Reiz-Reaktions-Experiment im Labor handelt, sondern um eine Befragungssituation zur Messung individueller Dispositionen, habe ich diese Methode als "Experimentellen Fragebogen" bezeichnet (Lind 1982). Angeregt wurde dieses Konzept durch einige methodische Ansätze in der Psychologie, die teilweise in Vergessenheit geraten sind oder in ihrer Bedeutung nicht voll erkannt wurden, so zum Beispiel von Brunswiks (1955) "diakritischer Methode". Brunswik hat diese Methode vorgeschlagen, um ein altes Problem der psychologischen Diagnostik zu lösen: Wie kann auf experimentellen Wege unter mehreren möglichen Dispositionen jene gefunden werden, die ein Verhalten tatsächlich bestimmten? Brunswik hat seinen Vorschlag dem Vorgehen beim Experiment entlehnt, in dem mehrere unabhängige Variablen in einem geeigneten Forschungsdesign so kombiniert werden ("orthogonales Design"), daß ihr Einfluß auf eine bestimmte abhängige Variable unabhängig voneinander geschätzt werden kann. Brunswiks diakritische Methode unterscheidet sich aber von der üblichen Verwendung experimenteller Designs, insofern er vorschlug, nicht die untersuchten Variablen zu variieren (das ist bei Dispositionen auch schlecht möglich), sondern die externen Reize. Als Beispiel führt er folgendes Gedankenexperiment an: Ein Vogel fliegt von einem Punkt los in eine bestimmte Richtung. Will er den Mast oder den Baum anfliegen, die beide in einiger Entfernung vor ihm liegen? Wie ließe sich herausfinden, was die Absicht oder das Motiv des Vogels ist, bevor er sein Ziel erreicht? Solange Mast und Baum in einer Linie zur Position des Vogels liegen, lassen sich im Verhalten des Vogels keinerlei Anhaltspunkte für die Richtigkeit der einen oder der anderen Vermutung finden. Eine reine Beobachtung führt hier nicht zum Ziel. Der Experimentator kann nur dann eindeutige Hinweise im Verhalten des Vogels erhalten, wenn er das eine Objekt, hier den Mast, so

versetzt, daß Vogel, Baum und Mast nicht mehr in *einer Linie* stehen. War es die Absicht des Vogels, den Mast anzufliegen, müßte er jetzt seine Flugrichtung ändern, womit auch für den Experimentator klar wäre, wohin der Vogel wirklich fliegen will.

Vielfach wird angenommen, daß mit experimentellen Designs nur Hypothesen über die Wirkung von Reizen geprüft werden können, weswegen der Einsatz von Experimenten in der Psychologie fast nur auf psycho-physische Fragestellungen begrenzt ist. Wie Brunswiks Anwendung der diakritischen Methode in bezug auf eine dispositionelle Fragestellung aber zeigt, können mit experimentellen Methoden auch Hypothesen über die Wirkung bestimmter Dispositionen im Verhalten untersucht werden, und zwar nicht nur global für mehrere Personen, sondern individuell für eine bestimmte Person. In einem Reiz-Experiment muß immer das zu untersuchende Objekt vom Experimentator variiert werden. Wenn Kausalhypothesen der Form "wenn *A* dann *B*" geprüft werden, muß "*A*" systematisch variiert werden, um den Kausalzusammenhang testen zu können. Anders liegen die Dinge, wenn eine Dispositionshypothese der Form "das Verhalten einer Person *P* ist durch die - positive oder negative - Einstellung zu *E* determiniert" geprüft wird. Da Dispositionen Verhaltenskonstanten sind, muß, um ihre Vorhandensein in einer bestimmten Situation feststellen zu können, die Umwelt von "*P*" systematisch in bezug auf "*E*" variiert werden. (Die in der psychologischen Einstellungsforschung übliche Annahme, daß Dispositionen selbst über längere Zeit konstant sind, ist hierfür nicht notwendig.)

Dispositionen können nicht dinglich separat vom betreffenden Verhalten erfaßt oder unabhängig davon definiert werden. Dennoch ist die Frage nach der Adäquatheit von Dispositionshypothesen nicht trivial. Wenn die Vermutung, daß das Urteilsverhalten einer Person von der Einstellung "*E*" determiniert ist, nicht zutrifft, dann wird dieses Verhalten durch einen bestimmten Wert auf einer Skala, welche die Einstellung zu "*E*" mißt, nicht richtig beschrieben. Oder, einfacher gesagt, der Psychologe mißversteht in diesem Fall seinen Probanden. Eine vermutete Disposition ist zunächst ein "hypothetisches Konstrukt", aber sie bleibt nur solange hypothetisch, bis sie empirisch als existent nachgewiesen ist.

4.4 Das Design des *Moralisches Urteil-Test* (MUT)

Beim *Moralisches Urteil-Test* wurde das Konzept des experimentellen Fragebogens eingesetzt. Damit kann man feststellen, an welchen der folgenden drei Faktoren (bzw. Faktorenkombinationen) sich einzelne Personen orientieren, wenn sie in einer Diskussion über moralische Dilemmas Argumente bewerten:

1. an der *moralischen Qualität* der vorgegebenen Argumente, die verschiedene moralische Orientierungen oder Stufen im Sinne Kohlbergs repräsentieren (wir nennen diesen Faktor kurz "Stufe"),
2. an der *Übereinstimmung oder Nicht-Übereinstimmung* der vorgegebenen Argumente mit der Meinung des Probanden (= Faktor Pro-Contra oder Meinungskonformität),
3. an dem *situativen Kontext* der Argumentation (= Faktor Dilemma).

Diese drei Faktoren bilden im Design des MUT die sogenannten *unabhängigen Variablen*. Es sind die Variablen, die zum Zwecke der kontrollierten Beobachtung hergestellt oder konstruiert wurden, die also unabhängig vom Verhalten der befragten Person sind. Bei der Konstruktion des MUT wurde zudem sichergestellt, daß diese Variablen auch untereinander unabhängig oder orthogonal sind, das heißt, daß ihre Ausprägungen nicht miteinander korrelieren: Die Argumente zu den Dilemmas im MUT wurden so ausgewählt, daß *jede* Ausprägungskombination vorkommt. Sie bilden, technisch gesprochen, ein 3-faktorielles (6 x 2 x 2) Design. Der Faktor Stufe hat sechs Ausprägungen (die sechs Kohlberg-Stufen), der Faktor Meinungskonformität zwei Ausprägungen (Pro- und Contra-Argumente) und der Faktor Dilemma im Standard-MUT auch zwei Ausprägungen (das Diebstahl-Dilemma und das Sterbehilfe-Dilemma). Die *abhängige Variable* im MUT ist die subjektive Beurteilung der Akzeptabilität der vorgegebenen Argumente auf einer Bewertungs-Skala, die gewöhnlich von -4 bis +4 reicht.¹³

Die Darbietung des MUT sieht so aus, daß die befragte Person Verhaltensdilemmas dargeboten bekommt, zu der sie Stellung beziehen soll. Im Sterbehilfe-Dilemma wird beispielsweise gefragt: "Halten Sie es eher für richtig oder eher für falsch, daß der Arzt der sterbenden Frau auf deren

13 Manchmal werden auch Bewertungsskalen von -2 bis +2 verwendet, zumeist dann, wenn anzunehmen ist, daß die befragte Personengruppe mit der differenzierteren Skala überfordert wäre.

Wunsch eine todbringende Spritze verabreichte?" Die Antwort kann die Person auf einer Bewertungsskala von -3 (ganz falsch) bis +3 (ganz richtig) geben. (Manche wird auch eine Skala verwendet, die zwei Punkte länger oder kürzer ist.) Mit ihrer Antwort gibt die Person ihre *Meinung* zum Dilemma an. Sie spricht sich zum Beispiel offen für oder gegen die Sterbehilfe aus. Diese Meinung legt auch fest, welche der nachfolgenden Argumente für und gegen Sterbehilfe aus der Sicht der befragten Person Pro- und Contra-Argumente darstellen. Ist eine Person *gegen* Sterbehilfe, dann stützen die Gegen-Argumente im MUT ihre Meinung, sind also *meinungskonforme* Argumente. Nach diesem Teil werden gemäß dem Design des MUT der befragten Person jeweils sechs Argument für und gegen die Entscheidung des Protagonisten in der Dilemma-Geschichte präsentiert. Diese Argumente muß sie, wie oben beschrieben, auf einer Antwortskala bewerten.

Die kleinste Auswertungseinheit des MUT ist das *gesamte* Urteilmuster einer Person und nicht einzelne Bewertungen, die eine Person gibt. Ausgewertet werden sowohl der affektive Aspekt wie der kognitive Aspekt des moralischen Urteilsverhaltens einer Person. Bei dem affektiven Aspekt werden gewöhnlich Maße für die *Einstellung* einer Person zu jeder der sechs moralischen Argumentations-Stufen nach Kohlberg gebildet. Dies geschieht in klassischer Weise durch Summierung der Einzelurteile der befragten Person oder, was gleichwertig ist, durch die Bildung von Mittelwerten.

Für den kognitiven Aspekt der moralischen Urteilsfähigkeit können mehrere, eng miteinander verwandte Maße gebildet, die ich an anderer Stelle (Lind 1978; 1985a) ausführlich diskutiert habe. Das am meisten verwendete Maß ist der Grad der *moralischen Determination* des Urteilsverhaltens (kurz auch *Determination Stufe* genannt). Dieses Maß ist im MUT operationalisiert als der Varianzanteil des Faktors Stufe an der gesamten Urteilsvarianz der befragten Person. Es wird berechnet durch eine Zerlegung der Quadratsummen ähnlich wie in der Varianzanalyse (Hays 1963). Der relative Anteil der Quadratsumme des Faktors Stufe (multipliziert mit 100) wird als Meßwert für die Konsistenz des individuellen Urteils in bezug auf die moralische Qualität der beurteilten Argumente verwendet. Der so ermittelte Meßwerte hat ein Minimum von 0 (dieser Wert bedeutet, daß die befragte Person sich beim Beurteilen von Argumenten überhaupt nicht an der moralischen Prinzipien orientierte) und ein Maximum von 100 (dieser Wert bedeutet, daß die oder der Befragte sich ausschließlich an moralischen Gesichtspunkten orientierten). Mit der Konstruktion dieses Meßwertes ist festgelegt, daß einer Person dann eine hohe

moralische Urteilsfähigkeit zugeschrieben wird, wenn sie Argumente nicht nur unabhängig von ihrer Übereinstimmung mit der eigenen Meinung beurteilt, sondern auch unabhängig von dem Dilemmatyp. Obwohl auch bei Kohlbergs Meßmethode dann einer Person eine höhere moralische Urteilsfähigkeit zugeschrieben wird, wenn ihr Urteil vom Dilemmatyp unabhängig ist, kann aus theoretischer Sicht in der Differenzierung des Urteils nach dem Dilemmatyp ein Anzeichen für höhere moralische Urteilsfähigkeit gesehen werden. Tatsächlich haben wir mit Hilfe der Quadratsummenzerlegung weitere Indikatoren gebildet, die diesen Einwand berücksichtigen (Lind 1978). In der Forschungspraxis wurde auf den Einsatz dieser komplexeren Meßwerte bislang jedoch weitgehend verzichtet, da sie noch ohne theoretischen "Unterbau" sind und in empirischen Vergleichen weitgehend dieselben Ergebnisse erbracht haben wie das einfachere Maß "Determination Stufe".

Aufbau und Auswertung des MUT werden an anderen Stellen eingehend beschrieben (Lind 1984, 1985a; Lind & Wakenhut 1985). Auch ist dort die Standardversion des MUT publiziert. Die technische Anleitung zur Auswertung kann beim Autor angefordert werden.

4.5 Resümee

Der *Moralisches Urteil-Test* (MUT) erfüllt die eingangs genannten Anforderungen an ein Meßinstrument, das eine adäquate Prüfung der zentralen Hypothesen dieser Untersuchung erlaubt. Der MUT ist, wie wir in den vorangegangenen Abschnitten gezeigt haben, theorie-konform (valide), objektiv (replizierbar) und leistungsabfordernd.

Der MUT erfüllt die Forderung nach theoretischer Validität, da es damit möglich ist, kognitiv-strukturelle Eigenschaften des individuellen Urteilsverhaltens zu erfassen. Mit dem MUT kann festgestellt werden, ob und in welchem Maß eine Person fähig ist, sich bei der Bewertung von Argumenten auf einen moralischen Standpunkt (*Moral point of view*) zu stellen, das heißt ihr Urteilsverhalten an der moralischen Qualität dieser Argumente auszurichten statt an der Frage, ob diese Argument ihre eigene Meinung stützen oder ihr widersprechen.

Die Messung der moralischen Urteilsfähigkeit mit dem MUT ist *objektiv*; sie kann von anderen Auswertern und bei anderen Befragten in allen Einheiten überprüft werden. Die Messung ist nicht auf intuitive, für Verzerrungen

anfällige Einschätzungen durch die befragte Person selbst angewiesen. Im MUT braucht die befragte Person nicht anzugeben, wie sie in einer ähnlichen Situation wie dem vorgegebenen Dilemma urteilen *würde*. Vielmehr wird durch den MUT ihr Urteilsverhalten *in situ* beobachtet. Da der MUT auch keine subjektive Einschätzung durch den Auswerter und dessen Theoriekenntnisse erfordert, besteht von dieser Seite keine Gefahr für die Neutralität des MUT gegenüber den in unserer Untersuchung zur Debatte stehenden Hypothesen.

Schließlich ist der MUT "*leistungsabfordernd*"; seine Bearbeitung verlangt von der befragten Person neben moralischen Einstellungen eine Leistung ab. Sie erhält nur dann einen hohen Testwert auf der Skala "moralische Urteilsfähigkeit", wenn sie fähig ist, Argumente konsistent nach ihrer moralischen Qualität zu beurteilen, und zwar auch dann, wenn diese Argumente ihrer eigenen, bekundeten Meinung zur richtigen Lösung eines Dilemma widersprechen. Der experimentelle Beweis für die These, daß die Testwerte des MUT fähigkeitsbedingt sind, ist Teil dieser Untersuchung. Eine Reihe der in den folgenden Kapiteln berichteten Studien befaßt sich direkt mit dieser Frage. So wird einerseits gefragt, ob sich die moralische Urteilsfähigkeit, wie sie im MUT gemessen wird, wie eine Einstellung durch eine einfache Instruktion verbessern läßt, und andererseits, ob fähigkeitsfördernde Programme einen nachweisbaren Effekt auf diesen Meßwert haben. Die leistungsabfordernde Natur des MUT wird in all diesen Studie bestätigt. Anders als die P-Werte des DIT von Rest (1979) lassen sich die MUT-Testwerte nicht durch eine einfache Instruktion verbessern (Kapitel 6). Wie in den Interventionsstudien, die mit dem MJI evaluiert wurden, zeigen auch die MUT-Werte Lerngewinne an. Die MUT-Werte scheinen im Vergleich zu den MJI-Werten sogar "sensibler" auf Effekte zu reagieren; so ergab im HASMU-Projekt (Oser & Schläfli 1986) die Messung mit dem MJI kein Effekt, die Messung mit dem MUT jedoch einen deutlichen Effekt (Kapitel 9).

Beim MUT wird die Stufe, auf der das Dilemma diskutiert wird, nicht als Kriterium für hohe Urteilsfähigkeit vorgegeben. Die Probanden sind frei in der Wahl der moralischen Argumentationsebene, da die Wahl der adäquaten Stufe der Dilemmalösung kaum eine objektive und unstrittige Festlegung von

"außen" möglich ist.¹⁴ Um einen hohen Testwert im MUT zu erhalten, ist nur Voraussetzung, daß sie sich an der selbst gewählten Moralebene *konsistent* orientieren. Der Meßwert "Determination durch moralische Stufe" ist aufgrund dieser Konstruktion *logisch unabhängig* von den MUT-Meßwerten für die moralischen Einstellungen und daher auch geeignet für die Untersuchung von Fragestellungen über den empirischen Zusammenhang beider Aspekte des moralischen Urteilens (Kapitel 2).

Da der MUT auf einem anderen methodologischen Konzept basiert als der MJI, wird bei diesem Meßverfahren keine Einordnung der Probanden in die sechs (oder fünf) Kohlberg-Stufen vorgenommen. Solche Ein-Stufungen mittels MUT-Werte wurden vorgeschlagen (Lind 1985a; Lind & Wakenhut 1985), um eine Vergleichbarkeit mit MJI-Studien herzustellen. Aber es treffen auf sie ähnliche Einwände zu wie auf den MJI. Die für eine Ein-Stufung erforderliche Vermengung von Meßwerten für den affektiven und kognitiven Aspekt des moralischen Urteilens würde viele Fragestellungen ausschließen. Zudem gibt ein direkter Vergleich von MUT- und MJI-Werten, wie Kohlberg (1985) ihn vorschlug, keinen Aufschluß über die Validität des MUT. Dazu müßte die Validität des MJI selbst zweifelsfrei gesichert sein, was aber, wie wir in Kapitel 7 sahen, nicht der Fall ist. Schmied (1981) fand eine mäßig hohe Korrelation zwischen MUT-Werten einerseits und MJI-Werten andererseits. Es ist jedoch unklar, welche Schlüsse aus diesem Befund zu ziehen sind. In dieser Studie wurden nur Oberschüler und Studierende untersucht, so daß die Varianz beider Meßwerte und damit auch die Korrelation zwischen ihnen *a priori* eingeschränkt war. Auch wurde dort die alte Version des MJI benutzt, die selbst nur mäßig hoch ($r = .30$) mit der neuen Version korreliert (Kohlberg & Candee 1984; s. auch Oser & Althof 1992, S. 75).

Der MUT liegt inzwischen in mehreren Sprachen vor und ist in zahlreichen Untersuchungen in unterschiedlichen Ländern eingesetzt worden (Kapitel 7). Es wurden damit Akademiker ebenso wie Hauptschulabsolventen untersucht, ältere Menschen ebenso wie sehr junge. Die jüngsten mit dem MUT untersuchten Personen sind 11jährige Haupt- und Realschüler. Bei Personengruppen mit geringen Bildungserfahrungen wurden teilweise das Layout (Schriftgröße), die Bewertungsskala (von -2 bis +2) und die Instruktion angepaßt. Wir haben

14 Diese Wahl spiegelt eher kultur- und aufgabenspezifische Einstellungen wieder als ein bestimmtes Niveau der moralischen Urteilsfähigkeit. Siehe Kapitel 7; auch Edwards 1986; Gielen 1986; Locke 1983; Sullivan 1977.

keine Beeinträchtigung der Validität der Meßwerte durch diese Änderungen feststellen können. Gleichwohl muß damit gerechnet werden, daß solche Änderungen, insbesondere aber die Darbietungsform (mit versus ohne Zeitlimits) das generelle Niveau der Testwerte beeinflussen können. Vergleiche über verschiedene Untersuchungen hinweg müssen dies immer berücksichtigen. Mit Ausnahme von der EMNID-Untersuchung (Kapitel 5) wurde der MUT bislang *ohne* Zeitlimit dargeboten.

Zur zusätzlichen, empirischen Sicherung der Validität von veränderten und von in fremde Sprachen übersetzten Versionen haben sich drei Analysen bewährt (Lind 1984): 1. Die Analyse der "Präferenzhierarchie". Gemäß der Theorie sollten die Befragten Stufe-6-Argumente am stärksten akzeptieren, Stufe-5-Argumente etwas weniger usw. und Stufe-1-Argumente am stärksten ablehnen.¹⁵ 2. Die Analyse der Interkorrelationen zwischen den Einstellungen zu den sechs Moralstufen, die nach Kohlberg (1958) eine "Quasi-Simplex-Struktur" bilden sollen. 3. Die Analyse der Korrelation zwischen den Einstellungen der Probanden zu den sechs Stufen auf der einen Seite und dem MUT-Meßwert für moralische Urteilsfähigkeit (s. Kapitel 2). Starke Abweichungen von den prognostizierten Werten waren bislang ausnahmslos durch "technische" Fehler bei der Datenauswertung oder der Übersetzung bedingt; sie verschwanden, sobald diese Fehler behoben waren.

15 Bei klassisch moralischen Dilemmas wie dem Sterbehilfe-Dilemma bestätigt sich diese Annahme fast ausnahmslos - unabhängig von Merkmalen wie Geschlecht, Alter, Schulbildung, Kulturkreis etc. Bei moralisch weniger anspruchsvollen Dilemmas, wie beim Diebstahl-Dilemma, präferieren die Befragten zumeist eine Argumentation auf Stufe 5 stärker als eine auf Stufe 6! Aber bezüglich der anderen Stufen bleibt auch hier die Rangreihe weitgehend gewahrt (Lind 1984; 1985a).