

Schmitt, M.

Zur Erfassung des moralischen Urteils: Zwei standardisierte objektive  
Verfahren im Vergleich.

P.I.V.-Bericht Nr. 4

## INHALT

	Seite
1. Prolog	1
2. Die Instrumente von REST und LIND zur Erfassung des moralischen Urteils	2
2.1 Der DIT von REST	3
2.2 Der m-u-t von LIND	7
3. Vergleich von m-u-t und DIT	9
3.1 Unterschiede in der Form der die Entwicklungsstufen des moralischen Urteils repräsentierenden Argumente/ Fragen	9
3.2 Unterschiede in der Vorgabesystematik der stufenspezifischen Argumente	10
3.3 Unterschiede in der abhängigen Variable	10
3.4 Unterschiede in der empirischen Überprüfung der Gütekriterien der Instrumente	11
3.5 Unterschiede in den aus den Rohwerten ableitbaren Maßen für die "moralische Reife"	11
4. Einige allgemeine Fragen der Quantifizierung des moralischen Urteilsniveaus	12
4.1 Rekognition vs. Konstruktion	12
4.2 Struktur vs. Inhalt	12
4.3 Qualitative vs. quantitative Maße der moralischen Reife	13
5. Schlußfolgerungen	15
Literatur	16

## 1. Prolog

Die vorliegende Arbeit entstand im Rahmen eines Forschungsprojektes über die "Entwicklung interpersonaler Verantwortlichkeit und interpersonaler Schuld."

Es ist eine Grundannahme in der theoretischen Konzeption dieses Forschungsprojektes, daß Helfehandeln nicht allein Ausdruck interindividuell variierender dispositioneller Hilfsbereitschaft ist, sondern ein Handlungsprozeß, der eine komplexe Wechselwirkung personaler und situativer Einflüsse umfaßt (MONTADA 1981). Als Folge dieser Grundannahme stellt sich die Aufgabe, nach jenen situativen und personalen Merkmalen zu suchen, die für das Beschreiben, Verstehen und Erklären von prosozialem (und antisozialem) Handeln wichtig sind. Dabei sind Theoriekonstruktion und ihre empirische Überprüfung gleichermaßen wichtig.

Ein Personmerkmal, für das ein theoretisch begründeter Zusammenhang mit Helfehandeln erwartet werden kann, ist das moralische Urteilsniveau im Sinne der kognitiven Entwicklungstheorie von PIAGET (1954) und KOHLBERG (z.B. 1969, 1976). Voraussetzung zur Prüfung theoretisch begründeter Zusammenhangshypothesen ist die Verfügbarkeit eines Instrumentes zur Quantifizierung inter- (und intra-) individueller Unterschiede im moralischen Urteilsniveau.

Die vorliegende Arbeit verfolgt das Ziel, eine begründete Empfehlung für die Wahl eines der verfügbaren Instrumente zur Erhebung des moralischen Urteilsniveaus zu geben.

Neben dem von KOHLBERG und Mitarbeitern entwickelten und vielfach modifizierten Standard-Interview (vgl. ECKENSBERGER, VILLENAVE-CREMER & REINSHAGEN 1980) gibt es Versuche besonders von REST (z.B. REST, COOPER, CODER & MASANZ 1974, REST 1976, 1979a, b, 1980) und LIND (1978, 1980, 1982, LIND & WAKENHUT 1980), das moralische Urteil mit standardisierten Verfahren zu messen.

Im folgenden werden zunächst der DIT (Defining Issues Test) von REST und der m-u-t (Moralisches-Urteil-Test) von LIND kurz beschrieben und verglichen. Im Anschluß daran werden einige allgemeine Fragen der Quantifizierung des moralischen Urteils behandelt.

## 2. Die Instrumente von REST und LIND zur Erfassung des moralischen Urteils

Der DIT wurde von REST und seinen Kollegen Anfang der 70er Jahre in den USA entwickelt. Der m-u-t wurde von LIND im Rahmen des Forschungsprojektes "Hochschulsozialisation" an der Universität Konstanz entwickelt und von KRÄMER-BADONI & WAKENHUT (1978) in einer modifizierten Fassung als m-u-f (Moralisches-Urteil-Fragebogen) vorgestellt. Sowohl der DIT als auch der m-u-t wurden in Anlehnung an die kognitive Entwicklungstheorie des moralischen Urteils von KOHLBERG entwickelt. LIND und REST bewegten im wesentlichen drei Motive zur Konstruktion ihrer standardisierten Fragebögen: (a) Die Interviewmethode von KOHLBERG kann zwei wesentliche klassische Testgütekriterien nur unbefriedigend erfüllen, nämlich das Kriterium der Objektivität und das Kriterium der Reliabilität. (b) Die Interviewmethode von KOHLBERG ist unökonomisch (aufwendige Interviewer-Schulung, aufwendige Interview-Auswertung). (c) Nach Meinung von LIND und auch von REST kommt es im Alltag wesentlich häufiger vor, daß jemand zu Argumenten (reaktiv) Stellung nehmen muß, die einen Tatbestand nach moralischen Gesichtspunkten entweder rechtfertigen oder verurteilen, als daß er eigene Argumente zur moralischen Beurteilung/Bewertung eines Tatbestandes generieren muß.

Die Forderung nach Objektivität, Reliabilität und Ökonomie hat sowohl bei REST als auch bei LIND zur Konstruktion von Fragebögen geführt, die folgende Merkmale aufweisen: Im Unterschied zu KOHLBERGs Interviewmethode müssen in diesen beiden Verfahren die Versuchspersonen keine Argumente zur Beurteilung eines moralischen Dilemmas generieren; vielmehr müssen sie vorhandene Argumente rekonstruieren und auf ihre Wichtigkeit bzw. ihre Akzeptabilität zur Beurteilung eines moralischen Dilemmas prüfen. Dabei entsprechen die zu beurteilenden Argumente jeweils einer der von KOHLBERG vorgeschlagenen Entwicklungsstufen des moralischen Urteils. Mit anderen Worten: Die Person liest ein moralisches Dilemma, das entweder im Sinne des einen oder des anderen Wertes gelöst wird. Im Anschluß an das Dilemma findet sie eine Reihe von Argumenten, die die Tat (d.h. die Lösung des Dilemmas) entweder rechtfertigen oder verurteilen. Die Versuchsperson soll nun angeben, für wie wichtig (bei REST) bzw. für wie akzeptabel (bei LIND) sie die jeweiligen Argumente für die moralische Bewertung des Tatbestandes erachtet.

Zwar verfolgen beide Instrumente dieselbe Meßintention, aber sie unterscheiden sich in einer Reihe von Merkmalen, nämlich (a) der Anzahl der Dilemmata, (b) der Form der zu beurteilenden Argumente, (c) der abhängigen Variable (Wichtigkeit vs. Azeptabilität) und (d) der verwendeten Meßwerte/Indices der "moralischen Reife". Die beiden Verfahren werden getrennt beschrieben.

## 2.1 Der DIT von REST

Es gibt zwei Versionen des DIT, eine vollständige Version bestehend aus 6 Dilemmata und eine Kurzversion bestehend aus drei Dilemmata. Da es hauptsächlich zur vollständigen Version Forschungsergebnisse gibt, rät REST von der Verwendung der Kurzversion ab. Insgesamt gibt es nach Aussagen von REST (1979b) bisher über 200 Studien zum DIT. REST hält den DIT für das beste Instrument zur objektiven Erfassung des moralischen Urteils. Interessanterweise verwendet REST weder in der Instruktion noch in den Geschichten und auch nicht in den Argumenten den Begriff "moralisch". Vielmehr spricht er von "sozialen Problemen" und bittet die Versuchspersonen um ihre Meinung zu Argumenten, die bei der Beurteilung der Problemlösungen eine Rolle spielen könnten. Eigentlich sind die Argumente noch keine Argumente, sondern vielmehr Fragen nach Gesichtspunkten des Problems und der Problemlösung (d.h. nach weiteren Informationen) auf die sich die moralische Beurteilung der Problemlösung stützen sollte.

Zu jedem Dilemma gibt es 12 Fragen nach Informationen, auf denen die entsprechenden Argumente aufbauen würden. Jede dieser 12 Fragen gehört einer der folgenden Kategorien an: den Stufen der KOHLBERG'schen Theorie (2, 3, 4, 5a, 5b, 6), einer Kategorie A (Anti-Establishment-Orientierung, vgl. REST 1979b, p. 5.2) oder einer Kategorie M (= meaningless, aber verbal komplex; die Kategorie M ist eine Kontrollkategorie, siehe unten). Es gibt also 8 Kategorien für die 12 Fragen pro Dilemma, d.h. die Fragen sind nicht über alle Kategorien gleich verteilt. Das ist ein wesentliches Problem des DIT von REST (siehe unten).

Pro Dilemma werden drei unterschiedliche Maße (Reaktionsklassen) der Versuchspersonen erhoben: (a) Wird dem Ausgang des Dilemmas zugestimmt? (ja, nein, unentschieden). (b) Jede der 12 Fragen muß nach ihrer Wichtigkeit zur Beurteilung des Konfliktes auf einer fünfstufigen

Skala eingestuft werden. (c) Die vier wichtigsten Fragen müssen bezüglich ihrer Wichtigkeit in eine Rangreihe gebracht werden.

Die von REST vorgeschlagenen Maße der "moralischen Reife" werden nur aus den Rangplätzen der vier wichtigsten Fragen/Argumente abgeleitet. Das wichtigste Argument, d.h. das Argument auf Rangplatz 1 erhält 4 Punkte, das Argument auf Rangplatz 2 erhält 3 Punkte, das drittwichtigste Argument erhält 2 Punkte und das viertwichtigste Argument erhält 1 Punkt. Der Stufenrohwert eines Dilemmas ergibt sich also aus dem Rangplatz, den das Argument, das diese Stufe vertritt, bezüglich seiner Wichtigkeit zur Beurteilung des Dilemmas nach moralischen Gesichtspunkten von der Versuchsperson zugewiesen bekommen hat. Für jede Stufe werden die Punkte über die Geschichten zusammengezählt. Das ergibt die Rohpunkte pro Stufe/Kategorie (auch für die Kategorien A und M). Insgesamt gibt es (wenn keine missing data in den Rangplätzen vorliegen) 60 Punkte pro Test. Die Prozentwerte, die für die einzelnen Stufen/Kategorien abgeleitet werden, ergeben sich aus den Rohwerten pro Stufe/Kategorie im Verhältnis zu der Gesamtzahl der möglichen Punkte pro Test.

Von REST wird das "P"-Maß als Maß der moralischen Reife vorgeschlagen. "P" ist die Summe der Prozentwerte, die auf die Stufen 5a, 5b und 6 entfallen (Anteil postkonventioneller Präferenzen an allen Präferenzen (gewichtete Rangsumme)). "P" kann streuen von 0 - 95%, weil für 3 Geschichten nur drei postkonventionelle Argumente verfügbar sind, d.h. nur die Rangplätze 1 - 3 und damit nur 11 anstatt 12 Punkte vergeben werden können.

Ein von DAVISON (vgl. REST 1979, p. 3.6) vorgeschlagener Index "D" soll "P" ersetzen. "D" beruht im Gegensatz zu "P" nicht nur auf den Prozentwerten der postkonventionellen Präferenzen, sondern auf den Prozentwerten aller Stufen der Entwicklung des moralischen Urteils. "D" und "P" korrelieren allerdings sehr hoch (ungefähr .80) miteinander. Außerdem ist "D" nur auf der Grundlage von Gewichtskoeffizienten berechenbar, die DAVISON einer Unfolding-Analyse von Daten der Eichstichprobe entnimmt. Auf die Berechnung von "D" kann (oder muß) deshalb in vielen Fällen verzichtet werden.

REST hat in seinen Fragebogen einige Kontrollen eingebaut, mit denen die Validität der Werte einer Person überprüft werden können soll. (a) Hat die Person einen hohen "M-Wert" (Rohwert > 8 oder Prozentwert > 14%), sollten ihre Testwerte nicht interpretiert werden. (b) Die Wichtigkeitseinstufungen pro Argument müssen mit der Rangreihe der Wichtigkeit der Argumente übereinstimmen: Auf dem Rangplatz 1 muß das Argument mit der höchsten Wichtigkeitseinschätzung stehen, auf Platz 2 das Argument mit der zweithöchsten Wichtigkeitseinstufung. Ist dem nicht so, soll ein Inkonsistenzkoeffizient (REST 1979b, p. 3.5) gebildet werden. Ab einem kritischen Inkonsistenzkoeffizienten ( $\geq 8$ ) sollen die Testwerte nicht mehr interpretiert werden.

REST und seine Kollegen haben einen sogenannten Scoring-Service eingerichtet, in dem sie ausgefüllte Testbögen per Computer automatisch auswerten und dem Autor die Roh- bzw. Prozentmaße jeder Versuchsperson mitteilen. Wiewohl dieser Service lobenswert erscheint, kann die Auswertung auch von Hand relativ rasch vorgenommen werden, besonders wenn die Auswerter geschult sind.

Die verschiedenen Indices/Maße der moralischen Entwicklung sind folgendermaßen zu interpretieren: Im Unterschied zu KOHLBERG will REST keine Zuordnung von Personen zu qualitativen Stufen vornehmen, sondern eine Zuordnung von Personen auf einem quantitativen Entwicklungskontinuum (obwohl auch er qualitativ unterschiedliche Denkformen und Formen der Argumentation annimmt).

Der am häufigsten verwendete "P"-Index ist ein Maß für die relative Wichtigkeit, die eine Person Argumenten auf einer postkonventionellen Ebene für die Lösung moralischer Dilemmata zuschreibt. Äquivalent zum "P"-Index gibt es Maße für die Stufen 2, 3, 4 und A. Wie bereits erwähnt ist der "D"-Index ein Maß, das sich als eine gewichtete Summe der Prozentwerte aller Stufen bilden läßt; es sollte ursprünglich den "P"-Index ersetzen, hat in einigen Untersuchungen seine Überlegenheit bewiesen (Kriterien sind unklar ...), nicht aber in allen Untersuchungen. Da die Berechnung des "D"-Maßes nicht von Hand vorgenommen wird, sondern die Verwendung des Algorithmus von DAVISON verlangt, verliert dieser Index an Attraktivität.

Der moralische Reifestand kann also abgelesen werden: (a) aus dem "P"-Index allein, (b) aus dem "D"-Index allein, (c) aus dem Profil der Prozentwerte zu den Stufen 2, 3, 4, 5a, 5b, 6, A, P, D.

Hier einige Angaben zur Reliabilität und Validität des DIT, die von REST et al. (1974) und REST (1979b) gemacht werden:

- Retest-Reliabilität des "P"-Maßes: .70 bis .80.
- Retest-Reliabilität für die Scores der Stufen 2 - 4 wesentlich niedriger (um .50).
- Die vollständige Version, bestehend aus 6 Geschichten, ist reliabler als die Kurzversion, bestehend aus 3 Geschichten.
- Es gibt einige Untersuchungen zur Kriteriumsgruppen-Validität, beispielsweise haben Studenten der Moralphilosophie wesentlich höhere "P"-Werte als 15-jährige High School-Schüler.
- Longitudinale Veränderungsvalidität (= theoretische und/oder Konstruktvalidität): Es gibt eine altersabhängige Zunahme von "P"- und "D"-Werten in Längsschnittuntersuchungen.
- Konvergente/diskriminative Validität (= Konstruktvalidität): Der DIT korreliert am höchsten mit anderen Instrumenten zur Erfassung des moralischen Urteils, aber auch nicht viel weniger hoch mit Maßen zur Erfassung der Intelligenz.
- Validierung durch Resistenz gegen experimentelle Entwicklungsintervention (theoretische und/oder Konstruktvalidität): DIT-Maße haben sich als resistent gegen Versuche der experimentellen Entwicklungsbeschleunigung (Trainings) erwiesen.
- Validierung durch Nachweis der Insensitivität der DIT-Maße gegen Veränderungen der Instruktion (theoretische und/oder Konstruktvalidität): Die Anweisung, "gute" und "schlechte" Ergebnisse zu erzielen, hat zu folgendem Resultat geführt: Personen, die die Anweisung bekamen, möglichst gute Ergebnisse zu erzielen, schafften dies nicht. Umgekehrt gelang es Personen, "möglichst schlechte" Ergebnisse zu erzielen.

Einige ausgewählte Ergebnisse aus Untersuchungen mit dem DIT (vgl. auch REST 1979a, 1980):

Mit "P"- bzw. "D"-Maßen korrelieren am höchsten: IQ, Bildung und sozio-ökonomischer Status. Ferner: Alter, aber nur so lange, wie Alter und Bildung konfundiert sind, d.h. die moralische Reife nimmt nach Abschluß der Schulzeit und Berufsausbildung nicht mehr zu.

Die moralische Reife ("P"-Werte) ist korreliert mit der geographischen Region (Personen aus dem konservativen Süden der USA haben wesentlich niedrigere "P"-Werte als Personen aus dem Norden). "P" korreliert mit Parteizugehörigkeit: Die niedrigsten Werte erzielten Republikaner, dann kommen die sogenannten Unabhängigen, die höchste "moralische Reife" weisen Demokraten auf.

## 2.2 Der m-u-t von LIND

Die Zielsetzung des Verfahrens ist es, sowohl kognitiv-strukturelle als auch die motivational-inhaltlichen Aspekte als Teile des moralischen Bewußtseins zu beschreiben und ihr Zusammenwirken bei der Bildung des moralischen Urteils zu analysieren. Unter struktureller Komponente versteht LIND die Konsistenz in der Beurteilung der Angemessenheit einer Begründungslogik (Konsistenz als interindividuell variierendes Personenmerkmal!). Unter der evaluativ-inhaltlich-motivationalen Komponente versteht LIND eine Haltung gegenüber einem Wert (Zustimmung vs. Ablehnung einer Handlung, die einen Wert verletzt bzw. erfüllt).

Der m-u-t weist folgende strukturell-methodologische Merkmale auf: Er ist ein experimenteller Fragebogen mit drei (Meßwiederholungs)-Faktoren/Facetten: (a) Die Stufe der moralischen Argumentation (nach KOHLBERG), (b) die Meinungskonformität des Beurteilers mit dem jeweiligen Argument (das den Handlungsausgang des Dilemmas entweder verteidigt oder angreift) und (c) die Art/der Typ des moralischen Dilemmas.

Der m-u-t besteht aus zwei Dilemmata, die folgendermaßen aufgebaut sind: Zunächst wird das Dilemma (Arbeiter, Arzt) in Form einer Geschichte beschrieben. Die Versuchsperson muß zuerst auf einer siebenstufigen Skala angeben, wie sehr sie das Verhalten des Akteurs des Dilemmas für richtig oder falsch ansieht. Danach soll sie 6 Argumente für den Handlungsausgang und 6 Argumente gegen den Handlungsausgang lesen und bezüglich ihrer Akzeptabilität zur Rechtfertigung bzw. zur Verurteilung des Handlungsausganges einschätzen. Diese 6 Argumente pro vs. contra Handlungsausgang entsprechen den 6 Stufen der Entwicklung des moralischen Urteils nach KOHLBERG. Abschließend soll die Versuchsperson angeben, ob es sich bei dem beschriebenen Problem in erster Linie um ein rechtliches, politisches, moralisches, gesellschaftliches oder humanitäres Problem handelt.

Aus dem m-u-t lassen sich folgende Meßwerte zur Indikation der moralischen Reife ableiten: LIND unterscheidet zwischen Strukturwerten und Inhaltswerten. (a) Strukturwerte: Anteil der intraindividuellen Urteilsvarianz, die durch den Faktor "Stufe" aufgeklärt wird (quadrierte Produkt-Moment-Korrelation zwischen diesem Faktor und dem individuellen Urteilsverhalten = "DETER STUFE"). (b) Inhaltswerte: Modale Stufe (= höchste stufenspezifische mittlere Akzeptabilität (nach KOHLBERG wäre dies ein Strukturwert!)). (c) Außerdem unterscheidet LIND (1978, p. 196) eine ganze Reihe von Hilfswerten (z.B. Verteilung von missing data).

LIND führt u.a. folgende Untersuchungsergebnisse als Belege der Validität des m-u-t an: (a) Es gibt eine theoriekonforme Entwicklungsveränderung (Längsschnittdaten?) der Präferenzdaten. (b) Der Strukturwert "DETER STUFE" korreliert mit den Präferenzwerten der Stufen und zwar negativ mit den Präferenzwerten der Stufen 1, 2 und 3 und positiv mit den Präferenzwerten der Stufen 5 und 6. Die "moralische Reife" (dargestellt in Form der stufenspezifischen Präferenzwerte) nimmt also mit zunehmender Konsistenz (intraindividuell, dargestellt im Strukturwert) zu.

LIND meint, der m-u-t könne nicht sinnvoll nach klassischen Testgütekriterien beurteilt werden (z.B. LIND 1982, p. 84). Das meine ich nicht! Zunächst ist der Test fraglos objektiv. Auch die Reliabilität des Tests läßt sich als Retestreliabilität sinnvollerweise prüfen, vorausgesetzt, das Zeitintervall ist lange genug gewählt, um Erinnerungseffekte auszuschließen und kurz genug gewählt, um Entwicklungsveränderungen auszuschließen. Dagegen ist die Bestimmung einer Paralleltestreliabilität (Korrelation der Kennwerte der beiden Dilemmata) nicht sinnvoll, da die Konsistenz des moralischen Urteilsniveaus über verschiedene Dilemmata nicht ä priori als gegeben angesehen werden kann. Vielmehr ist der Einfluß des Inhaltes eines Dilemmas auf die Struktur der moralischen Beurteilung seiner beiden Lösungen empirisch zu klären.<sup>1</sup> Hier handelt es sich um ein allgemeines Problem, das sich in fast allen Bereichen der Psychologie als Frage nach der Konsistenz des Verhaltens stellt (vgl. auch SCHMITT 1980). Schließlich läßt sich auch die Validität

<sup>1</sup>LIND (1978, p. 176) spricht in diesem Zusammenhang "Segmentierung".

des m-u-t sinnvoll empirisch bestimmen: Die inhaltliche und Augenscheinvalidität der Dilemmata und Argumente ließe sich über Expertenurteile bestimmen. LIND hat dies offensichtlich getan, bezeichnet die "kategoriale Repräsentativität" der Argumente aber als Kriterium für theoretische Validität (1982, p. 84). Auch die Konstruktvalidität des m-u-t ließe sich bestimmen, etwa über den Vergleich der Präferenzurteile oder Strukturwerte bei unterschiedlichen Kriteriumsgruppen oder über die Analyse der Zusammenhänge dieser Maße mit anderen einschlägigen Personvariablen (im Sinne der konvergenten und diskriminativen Validität LIND (1982, LIND & WAKENHUT 1980) LIND (1982, LIND & WAKENHUT 1980) hat natürlich recht, dass die empirisch bestimmte Güte eines Tests (etwa anhand des Kriteriums der internen Konsistenz) ebensowenig die Gültigkeit einer Theorie garantiert wie umgekehrt ein ungültiger Test die Falsifikation einer Theorie bedeuten kann. Aber auch das ist ein Problem, mit dem sich jede Disziplin, die den Anspruch verfolgt, Theorien empirisch zu prüfen, konfrontiert sieht (vgl. auch MONTADA & SCHMITT 1982). Sollten sich beispielsweise in Längsschnittuntersuchungen mit dem m-u-t nicht die theoretisch erwarteten Entwicklungsveränderungen im moralischen Urteil finden lassen, kann nicht entschieden werden, ob die Theorie der Entwicklung des moralischen Urteils ungültig ist (bzw. für die untersuchte Stichprobe nicht gilt) oder der m-u-t.

LIND meint, der m-u-t sei in dem Maße "nützlich", wie er es erlaube, intraindividuelle Antwortvarianz sinnvoll zu rekonstruieren, d.h. der Test ist umso "nützlicher", je weniger Varianz auf den Faktor "Rest" (das ist die intraindividuelle Urteilsvarianz, die nicht durch die Faktoren "Stufe", "Argumentationsrichtung" und "Dilemma" erklärt werden kann) entfällt.

### 3. Vergleich von m-u-t und DIT

#### 3.1 Unterschiede in der Form der die Entwicklungsstufen des moralischen Urteils repräsentierenden Argumente/Fragen

(1) Beim DIT werden Argumente in Form abstrakter Fragen nach Gesichtspunkten, die für die moralische Bewertung der Lösung des Dilemmas wichtig sein könnten, angeboten.

(2) Beim m-u-t werden Argumente bereits konkret formuliert, der Bezug zum Handlungskontext des Dilemmas ist unmittelbarer, es muß deshalb keine

kognitive Zwischenleistung (Transformation eines allgemeinen Prinzips auf ein konkretes Dilemma) von der Versuchsperson vorgenommen werden.

(3) Man kann erwarten, daß beim m-u-t subjektive Interpretationsunterschiede weniger zur Urteilsvarianz beitragen als beim DIT. Das macht den m-u-t meines Erachtens überlegen.

### 3.2 Unterschiede in der Vorgabesystematik der stufenspezifischen Argumente

(1) Beim DIT ist die Anzahl der Argumente nicht über alle Kategorien (2, 3, 4, 5a, 5b, 6, A, M) gleich verteilt, und zwar weder pro Dilemma noch über alle Dilemmata hinweg. Vielmehr sind Argumente der konventionellen Ebene (Stufen 3, 4) gegenüber prä- und postkonventionellen Argumenten deutlich überrepräsentiert. Daraus ergibt sich für konventionelle Argumente eine höhere Wahrscheinlichkeit, auf einen der vier Rangplätze zu kommen als für prä- und postkonventionelle Argumente. Selbstverständlich wirkt sich dieses Problem nur aus, wenn Argumente "unsystematisch" in eine Rangreihe gebracht werden, denn systematische und Zufallsvarianz sind nicht mehr unabhängig, d.h. die Fehlerwahrscheinlichkeit ist über die Kategorien ungleich verteilt. Je mehr der Varianz Fehlervarianz ist, desto stärker ist ein Verteilungsgipfel der Stichprobenmeßwerte in den Stufen 3 und 4 zu erwarten (tatsächlich sind die Stufen 3 und 4 in den meisten untersuchten Stichproben die am häufigsten vorkommenden modalen Stufen).

(2) Im Unterschied dazu weist der m-u-t die Merkmale eines vollständig gekreuzten experimentellen Designs auf: Es gibt pro Dilemma 12 Argumente, und zwar für jede Stufe ein Argument, das den Ausgang des Dilemmas rechtfertigt, und ein Gegenargument. Die Faktoren "Stufe" und "Argumentationsrichtung" sind orthogonal, und die Varianzanteile dieser Faktoren lassen sich ebenso wie der Varianzanteil des Faktors "Dilemma" für jede Person bestimmen.

### 3.3 Unterschiede in der abhängigen Variable

(1) Beim DIT wird nach der Wichtigkeit einer Frage zur Suche nach Argumenten für die moralische Wertung des Dilemmas gefragt.

(2) Beim m-u-t wird nach der Akzeptabilität bereits formulierter Argumente gefragt.

(3) Welche der beiden Formen einen valideren Indikator für die Struktur des moralischen Urteils darstellt, ist ohne die Verfügbarkeit eines Kriteriums, dessen Validität gesichert ist, nicht entscheidbar.

### 3.4 Unterschiede in der empirischen Überprüfung der Gütekriterien der Instrumente

(1) REST (1979a, b, 1980) meint, die Zuverlässigkeit und Gültigkeit des DIT sei in einer Vielzahl von Untersuchungen belegt worden.

(2) Im Vergleich dazu sind die Untersuchungen zum m-u-t (und m-u-f) noch spärlich, auch wenn ein Zuwachs zu verzeichnen ist (vgl. LIND 1982).

(3) Andererseits ist mir keine Untersuchung zu der von LIND & SCHMIDT (o.J.) ins Deutsche übersetzten Version des DIT bekannt, was ihre Verwendung weniger attraktiv macht.

### 3.5 Unterschiede in den aus den Rohwerten ableitbaren Maßen für die "moralische Reife"

(1) Prinzipiell erlauben beide Tests die Bildung sowohl stufenspezifischer Kennwerte (mittlere Wichtigkeit und Akzeptabilität, die eine Versuchsperson stufenspezifischen Argumenten zugesteht) als auch die Bildung der modalen Stufe einer Versuchsperson (Stufe mit der höchsten mittleren Wichtigkeit oder Akzeptabilität bezüglich der sie repräsentierenden Argumente): REST schlägt das "P"- bzw. "D"-Maß als Maß der Wahl für mit dem DIT gewonnene Daten vor. Andererseits schlägt LIND (1978, p. 197) eine Fülle von Maßen vor, die aus m-u-t-Rohdaten gewonnen werden können. Zum Teil ist der theoretische Stellenwert, den LIND diesen Maßen zuspricht, noch offen. Aber es gibt die sehr sinnvollen und informativen Maße der Anteile intraindividuelle Urteilsvarianz, die auf die Faktoren "Stufe", "Argumentationsrichtung" und "Dilemma" entfallen. Diese Maße lassen eine Aussage darüber zu, wie sehr die Urteilsvarianz nach den Maßgaben der kognitiven Entwicklungstheorie konsistent (auf Faktor "Stufe" muß ein hoher Varianzanteil entfallen) ist versus bedingt wird durch inhaltliche Aspekte des zu beurteilenden Sachverhaltes (Art des Dilemmas, Richtung der vorgegebenen Argumente).

Diese Vorteile eines experimentellen Fragebogens (vgl. auch LIND 1982, SCHMITT & MONTADA 1982) machen den m-u-t attraktiv und meines Erachtens dem DIT überlegen.

#### 4. Einige allgemeine Fragen der Quantifizierung des moralischen Urteilsniveaus

##### 4.1 Rekognition vs. Konstruktion

Ein wesentlicher Unterschied zwischen dem Erfassungsinstrument von KOHLBERG (Interview) und den objektiven Fragebögen zum moralischen Urteil ist folgender: Während bei KOHLBERG die interviewte Person die Argumente für bzw. gegen den Handlungsausgang eines Dilemmas selbst konstruieren und formulieren muß, findet sie sowohl in dem Verfahren von REST als auch in dem Verfahren von LIND Argumente vor, die sie rekognizieren und dann beurteilen muß. Dieser Unterschied ist den Autoren des DIT und m-u-t und den Kritikern dieser Verfahren (z.B. REINSHAGEN 1978, SCHMIED 1981) gleichermaßen bewußt. Unterschiedlich sind allerdings die Folgerungen, die daraus gezogen werden. REINSHAGEN z.B. glaubt, die "objektiven" Verfahren zur Quantifizierung des moralischen Urteils würden das Entwicklungsniveau überschätzen. Man kann REINSHAGEN aber ein Argument entgegenhalten, das auch von LIND & WAKENHUT (1980) oder auch von REST et al. (1974) geteilt wird, nämlich folgendes: Die Konstruktion eines Argumentes für oder gegen den Ausgang eines moralischen Dilemmas im KOHLBERG-Interview erfordert neben der kognitiven Leistung der Konstruktion einer Argumentationslogik auch eine Reihe weiterer Kompetenzen, z.B. Verbalisierungsfähigkeit, Angstfreiheit etc. Man kann sich also fragen, ob nicht etwa die Methode von KOHLBERG das Entwicklungsniveau des moralischen Urteils einer Person unterschätzt. Zumindest lassen die recht hohen positiven Korrelationen zwischen "moralischer Reife" einerseits und Intelligenz andererseits dieses Argument gegen den Einwand von REINSHAGEN plausibel erscheinen.

##### 4.2 Struktur vs. Inhalt

Die kognitive Entwicklungstheorie von KOHLBERG versteht unter Struktur die Argumentationslogik, mit der ein Handlungsausgang (z.B. eines moralischen Dilemmas) gerechtfertigt bzw. angefochten wird. Unter die Kategorie "Inhalt" faßt KOHLBERG (z.B. 1976) sowohl sogenannte "Formen" einer

moralischen Orientierung (normative Ordnung, Nützlichkeitskonsequenzen, Gerechtigkeit oder Fairness, Idealselbst) als auch die sogenannten "Werte" oder "moralischen Themen", die er unterscheidet (Sätze und Regeln, Gewissen, personale Rollen der Zuneigung, Autorität, ziviles Recht, Verträge und Austauschgerechtigkeit, Bestrafung und Gerechtigkeit, den Wert des Lebens, Besitzrechte und Werte, Wahrheit, Sexualität und Liebe). Die von KOHLBERG angenommene Unabhängigkeit von Struktur und Inhalt ist eine Idealkonzeption der kognitiven Entwicklungstheorie, die in empirischen Untersuchungen immer wieder widerlegt werden konnte. Während KOHLBERGs Interview und RESTs DIT keine Möglichkeit bieten, den Einfluß des Inhaltes eines moralischen Dilemmas auf die Struktur der moralischen Beurteilung seines Handlungsausgangs systematisch zu analysieren, hat LIND mit seinem m-u-t bewußt Wert darauf gelegt, sowohl den Einfluß struktureller als auch den Einfluß inhaltlicher Gesichtspunkte von Argumenten für und gegen den Ausgang eines Dilemmas unkonfundiert untersuchen zu können. Damit hat LIND auch die Möglichkeit geschaffen, den interindividuell unterschiedlichen Einfluß von inhaltlichen vs. strukturellen Merkmalen von Argumenten auf das moralische Urteil abzubilden. Die Ergebnisse von Untersuchungen mit dem m-u-t weisen darauf hin, daß es tatsächlich eine Reihe von Personen gibt, bei denen inhaltliche Gesichtspunkte von Dilemma und Argument stärker ihr moralisches Urteil bestimmen als strukturelle Gesichtspunkte. Das sind in erster Linie Personen, die nach KOHLBERG noch nicht eine postkonventionelle moralische Reife erreicht haben (vgl. LIND 1978).

#### 4.3 Qualitative vs. quantitative Maße der moralischen Reife

Streng genommen müßte sich nach der kognitiven Entwicklungstheorie von KOHLBERG jede Person einer Stufe der Entwicklung des moralischen Urteils zuweisen lassen. Allerdings hat auch KOHLBERG bereits 1958 bei der Entwicklung eines Scoring-Systems zur Auswertung der Argumente seiner Versuchspersonen insofern ein quantitatives Maß entwickelt, als er die Anzahl von stufenspezifischen Argumenten auszählte und eine Person zu einer Entwicklungsstufe gemäß der Häufigkeitsverteilung ihrer stufenspezifischen Argumente vornahm. Bereits KOHLBERG ging von einer sogenannten Präferenzhierarchie aus, die folgendermaßen aussieht: Die Häufigkeitsverteilung stufenspezifischer Argumente steigt monoton an bis

zur modalen Stufe und fällt danach entweder monoton oder unsystematisch wieder ab. Entwickelte man für jede Stufe und jede Person ein quantitatives Präferenzmaß, so müßte sich nach Meinung von KOHLBERG und auch anderen (z.B. LIND) in der Matrix der Interkorrelationen der Stufenindices ein quasi-Simplex wiederfinden lassen. Tatsächlich kann LIND (1978) mit den mit dem m-u-t gewonnenen Daten Belege für die Richtigkeit der Annahme einer solchen Präferenzhierarchie vorlegen: Die Interkorrelationen der stufenspezifischen Akzeptabilitätsmaße weisen quasi-Simplex-Struktur auf, d.h. die Akzeptabilitätswerte jeweils benachbarter Stufen korrelieren am höchsten (In der Interkorrelationsmatrix fallen die Korrelationskoeffizienten der Höhe nach, je weiter sie von der Diagonale entfernt sind.).

Es gibt aber auch noch andere Argumente gegen REINSHAGENS (1978) strikte Forderung nach qualitativen Maßen zur Abbildung des Entwicklungsstandes der moralischen Reife: (a) Eine Theorie verliert meines Erachtens an Güte, wenn ä priori Falsifikationsmöglichkeiten ausgeschlossen werden. Wenn also von der kognitiven Entwicklungstheorie angenommen wird, daß es eine eindeutige Zuordnung von Personen zu qualitativ unterschiedlichen Stufen der Entwicklung des moralischen Urteils gibt, ist es nicht mehr als legitim, diese Aussage der empirischen Prüfung zu unterziehen. Natürlich muß man bei widersprüchlichen Ergebnissen mit der Interpretation dieser Ergebnisse sehr vorsichtig sein: Wie oben bereits erwähnt wurde, kann dies gegen die Validität der Theorie sprechen, aber auch gegen die Güte des Meßinstrumentes, d.h. es kann entweder das theoretische Relativ falsch sein, oder aber das numerische Relativ entspricht nicht dem empirischen Relativ, oder aber das empirische Relativ entspricht nicht dem theoretischen Relativ. Die ä priori-Forderung nach qualitativen Stufenzuordnungen von Personen schränkt also unnötig die Falsifikationsmöglichkeiten der Theorie ein, d.h. die Möglichkeit, daß Personen auf mehreren Stufen gleichzeitig argumentieren, wird ä priori verneint. (b) Bei Verwendung quantitativer Maße (wie sie auch von Kohlberg benutzt werden) bedeutet aber eine qualitative Stufenzuordnung unter Umständen erheblichen Informationsverlust, denn die Zuordnung muß anhand einer Verrechnung mehrerer quantitativ abgebildeter Urteile vorgenommen werden und dies mag bei unterschiedlichen Personen ganz unterschiedlich eindeutige Ergebnisse erbringen (beispielsweise mag die Person A immer,

d.h. konsistent auf Stufe 4 argumentieren, während Person B inhalts- und orientierungsabhängig auf den Stufen 2., 3, 4, 5 argumentieren mag). Information über diese Konsistenzunterschiede ginge aber verloren, würde man sich in einem solchen Fall auf die qualitative Stufenzuordnung beschränken.

#### 15. Schlußfolgerungen

Aus den in dieser Arbeit (besonders unter den Punkten 2., 4.1, 4.2 und 4.3) genannten Gründen wird der Verwendung eines objektiven, standardisierten Verfahrens zur Quantifizierung des Entwicklungsstandes des moralischen Urteils im Rahmen des Projektes "Entwicklung interpersonaler Verantwortlichkeit und interpersonaler Schuld" (MONTADA 1981) gegenüber der Verwendung des KOHLBERGInterviews Vorrang gegeben. Von den beiden hier besprochenen Verfahren wird der m-u-t von LIND aus den unter 3.1, 3.2, 3.5 und 4.2 genannten Gründen dem DIT von REST vorgezogen.

LITERATUR

- ECKENSBERGER, L.H., VILLENAVE-CREMER, S. & REINSHAGEN, H. 1980. Kritische Darstellung von Methoden zur Erfassung des Moralischen Urteils. In: ECKENSBERGER, L.H. & SILBEREISEN, R.K. (Ed.) Entwicklung sozialer Kognitionen: Modelle, Theorien, Methoden, Anwendungen. Stuttgart: Klett-Cotta. p. 335 - 377.
- KOHLBERG, L. 1969. Stage and sequence: The cognitive-developmental approach to socialization. In: GOSLIN, D.A. (Ed.) Handbook of socialization theory and research. Chicago: Rand McNally. p. 347 - 480.
- KOHLBERG, L. 1976. Moral stages and moralization: The cognitive-developmental approach. In: LICKONA, T. (Ed.) Moral development and behavior. New York: Holt, Rinehart & Winston. p. 31 - 53.
- KRÄMER-BADONI, T. & WAKENHUT, R. 1978. Möglichkeiten der Skalierung des moralischen Urteilsverhaltens. In: ECKENSBERGER, L.H. (Ed.) Entwicklung des moralischen Urteils. Saarbrücken: Universitätsdruck. p. 379 - 391.
- LIND, G. 1978. Wie mißt man moralisches Urteil? Probleme und alternative Möglichkeiten der Messung eines komplexen Konstruktes. In: PORTELE, G. (Ed.) Sozialisation und Moral. Weinheim: Beltz. p. 171 - 201.
- LIND, G. 1980. Moralisches-Urteil-Test: Hinweise zum Einsatz und zur Interpretation eines Forschungsinstrumentes. Konstanz: Universität Konstanz (unveröffentlichtes Manuskript).
- LIND, G. 1982. Der 'Moralisches-Urteil-Test' (m-u-t) in der Kritik - eine Replik auf Schmieds Analyse. Diagnostica 28, 80 - 89.
- LIND, G. & WAKENHUT, R. 1980. Erfassung von moralischem Urteil mit standardisierten Fragebogen. Diagnostica 26, 312 - 334.
- MONTADA, L. 1981. Entwicklung interpersonaler Verantwortlichkeit und interpersonaler Schuld. Projektantrag an die Stiftung Volkswagenwerk. Trier: P. I. V. - Bericht Nr. 1 (= Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 8).
- MONTADA, L. & SCHMITT, M. 1982. Applied developmental psychology: A life-span perspective. In: BALTES, P.B. & BRIM, O.G.Jr. (Ed.) Life-span development and behavior (Vol. 4). New York: Academic Press (im Druck).

- PIAGET, J. 1954. Das moralische Urteil beim Kinde. Zürich: Rascher.
- REINSHAGEN, H. 1978. Forderungen für die Messung des moralischen Urteils mit standardisierten Verfahren. Diagnostica 24, 137-145.
- REST, J.R. 1976. New approaches in the assessment of moral judgment. In: LICKONA, T. (Ed.) Moral development and behavior. New York: Holt, Rinehart & Winston. p. 198 - 218.
- REST, J.R. 1979a. Development in judging moral issues. Minneapolis: University of Minnesota Press.
- REST, J.R. 1979b. Revised manual for the defining issues test: An objective test of moral judgment development. Minneapolis: Minnesota Moral Research Projects.
- REST, J.R. 1980. Development in moral judgment research. Developmental Psychology 16, 251 - 256.
- REST, J.R., COOPER, D., CODER, R., MASANZ, J. & ANDERSON, D. 1974. Judging the important issues in moral dilemmas - an objective measure of development. Developmental Psychology 10, 491 - 501.
- SCHMIDT, U. & LIND, G.o.J: Vorläufige Übertragung von REST's "Defining Issues Test" (DIT) ins Deutsche. Konstanz: Zentrum I Bildungsforschung, Universität Konstanz (unveröffentlichtes Manuskript).
- SCHMIED, D. 1981. Standardisierte Fragebogen zur Erfassung des Entwicklungsstandes der moralischen Urteilskompetenz. Diagnostica 27, 51 - 65.
- SCHMITT, M. 1980. Person, Situation oder Interaktion? Eine zeitlose Streitfrage diskutiert aus der Sicht der Gerechtigkeitsforschung. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 6.
- SCHMITT, M. & MONTADA, L. 1982. Determinanten erlebter Gerechtigkeit. Zeitschrift für Sozialpsychologie 13, 32-44.

## Bisher erschienene Arbeiten dieser Reihe

- MONTADA, L. 1978. Schuld als Schicksal? Zur Psychologie des Erlebens moralischer Verantwortung. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 1.
- DOENGES, D. 1978. Die Fähigkeitskonzeption der Persönlichkeit und ihre Bedeutung für die Moralerziehung. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 2.
- MONTADA, L. 1978. Moralerziehung und die Konsistenzproblematik in der Differentiellen Psychologie. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 3.
- MONTADA, L. 1980. Spannungen zwischen formellen und informellen Ordnungen. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 4.
- DALBERT, C. 1980. Verantwortlichkeit und Handeln. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 5.
- SCHMITT, M. 1980. Person, Situation oder Interaktion? Eine zeitlose Streitfrage diskutiert aus der Sicht der Gerechtigkeitsforschung. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 6.
- SCHMITT, M. & MONTADA, L. 1981. Entscheidungsgegenstand, Sozialkontext und Verfahrensregel als Determinanten des Gerechtigkeitsurteils. Trier: Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 7.
- MONTADA, L. 1981. Entwicklung interpersonaler Verantwortlichkeit und interpersonaler Schuld. Projektantrag an die Stiftung Volkswagenwerk. Trier: P.I.V.-Bericht Nr. 1 (= Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 8).
- SCHMITT, M. 1982. Empathie: Konzepte, Entwicklung, Quantifizierung. Trier: P.I.V.-Bericht Nr. 2 (= Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 9).
- DALBERT, C. 1982. Der Glaube an eine gerechte Welt: Zur Güte einer deutschen Version der Skala von RUBIN und PEPLAU. Trier: P.I.V.-Bericht Nr. 3 (= Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 10).

Andernorts publizierte Arbeiten aus dieser Arbeitsgruppe

- MONTADA, L. 1977. Moralisches Verhalten. In: HERRMANN, T., HOFSTÄTTER, P. R., HUBER, H. & WEINERT, F.E. (Ed.) Handbuch psychologischer Grundbegriffe. München: Kösel. p. 289-296.
- MONTADA, L. 1980. Gerechtigkeit in Wandel der Entwicklung. In: MIKULA, G. (Ed.) Gerechtigkeit und soziale Interaktion. Bern: Huber. p. 301-329.
- MONTADA, L. 1980. Moralische Kompetenz: Aufbau und Aktualisierung. In: ECKENSBERGER, L.H. & SILBEREISEN, R.K. (Ed.) Entwicklung sozialer Kognitionen: Modelle, Theorien, Methoden, Anwendungen. Stuttgart: Klett-Cotta. p. 237-256.
- MONTADA, L. 1981. Gedanken zur Psychologie moralischer Verantwortung. In: ZSIFKOVITS, V. & WEILER, R. (Ed.) Erfahrungsbezogene Ethik. Berlin: Duncker & Humblot. p. 67-88.
- MONTADA, L. 1981. Voreingenommenheit im Urteil über Schuld und Verantwortlichkeit. Trierer Psychologische Berichte, Band 8, Heft 10.
- SCHMITT, M & MONTADA, L. 1982. Determinanten erlebter Gerechtigkeit. Zeitschrift für Sozialpsychologie 13, 32-44.
- DAHL, U., MONTADA, L. & SCHMITT, M. 1982. Hilfsbereitschaft als Personmerkmal. Trierer Psychologische Berichte. (im Druck)
- DALBERT, C. & MONTADA, L. 1982. Vorurteile und Gerechtigkeit in der Beurteilung von Straftaten. Eine Untersuchung zur Verantwortlichkeitsattribution. Trierer Psychologische Berichte. (im Druck)
- MONTADA, L. 1982. Entwicklung moralischer Urteilsstrukturen und Aufbau von Werthaltungen. In: OERTER, R., MONTADA, L. u.a. Entwicklungspsychologie. München: Urban & Schwarzenberg. (im Druck).