Georg Lind

Testing Moral Judgment Competence. A critical review of the attempt to calculate C-score from the DIT

1996

Contact: Prof. Georg Lind
University of Konstanz
FB Psychologie
78457 Konstanz
E-Mail: Georg.Lind@uni-konstanz.de

For further information and publications on this topic see
www.uni-konstanz.de/ag-moral/b-publik.htm

Psychology of Morality &
Democracy and Education

Abstract

The Moral Judgment Test (MJT) measures moral judgment competence. Its design is distinct from common tests of moral judgment in several respects: counterarguments to probe subjects' ability to apply moral principles consistently; the subject's own moral ideals as a criterion; dilemmas that require 'Principled' moral reasoning; the balanced orthogonal test design to disentangle major judgment factors; the C index that extracts relational information from individuals' response pattern and attributes consistency to the person rather than to the test; and the test's unbiasedness about developmental regression. These features allow us to prove a) that moral judgment and behavior have a strong competence aspect, b) that moral judgment competence can be fostered through education, and c) that people's moral competencies regress if their education is insufficient. For Rest, Thoma and Edwards (this volume), the MJT misses their benchmarks. However, as this paper shows, their evaluation is inadequate: Their analysis misses most of the distinct features of the MJT. Moreover, their C index is based on an attitude test rather than on a competence test and thus has a completely different meaning than our C score.

The Moral Judgement Test (MJT), which we designed in the early 1970s and revised once in 1977 (Lind, 1978; Lind & Wakenhut, 1985), is one of the most widely used measurement instruments in the field of moral development research. Culturally validated versions of the MJT have been used in research projects in France, Hungary, Israel, Italy, Mexico, the Netherlands, Poland, Saudi-Arabia, Slovenia, Spain, Turkey, and the United States. The findings from this research have been summarized in a series of articles and books, most of them in German, although some are in English (see Lind, 1985a; 1985b; 1993; 1995; 1996a; 1996b).

Dr. James Rest and his colleagues (Rest et al., 1997) recently evaluated the MJT in a very knowledgeable and extensive critique of the MJT. The main thrust of their critique is aimed at "the key notion of Lind's work [. . .] his distinction between stage preference and stage consistency." On the basis of numerous analyses of data produced with their research instrument, the Defining Issues Test (DIT; see Rest, 1979), they conclude that the scoring stage consistency does not provide better measures of moral development than scoring person's stage preferences, and that, therefore, the DIT outperforms the MJT.

In this paper, I will challenge their conclusion. My main point is that the most important feature of the MJT is its in-built moral task. The scoring of judgment consistency is only meaningful in connection with this moral task. Because the MJT has a built-in moral task, but not the DIT, any evaluation of the MJT must be based on data produced with this instrument. Therefore, it is inadequate to use DIT data for benchmarking consistency scoring as Rest et al. (1997) did. I will show in this paper that, if both tests are properly evaluated, the MJT clearly outperforms the DIT in measuring moral judgment competence.

Kohlberg (1964) once defined moral judgment competence as "the capacity to make decisions and judgments that are moral (i.e., based on internal principles) and to act in accordance with such judgments" (p. 425). This definition of moral judgment competence, I felt, represented a major breakthrough in moral psychology and education (Lind, 1993; in press). It defines moral behavior as a function not only of one's moral attitudes and values, but also of one's ability to apply these values to concrete decision-making.

Yet, Kohlberg's definition left one important question open: What is an appropriate moral task by which to test moral judgment competence? In research, three major criteria have been considered, which can be discussed here only in short: a) external behavioral standards, b) the level of moral reasoning that a per-

son prefers or produces in a dilemma situation, and c) the universality or 'behavioral necessity' of a person's moral knowledge. As Blasi (1980), Pittel and Mendelsohn (1966), Kohlberg (1958; 1984) and others have convincingly argued, behavioral standards defined by the researcher, do not provide universally valid indicators of moral judgement competence.

Therefore, many researchers have proposed to use a person's moral attitudes instead of their behaviors as an indicator of moral development (Levy-Suhl, 1912; Piaget, 1965; Kohlberg, 1958; Rest, 1979). In psychology, the term 'attitude' is commonly defined as "a psychological tendency expressed by evaluating a particular entity with some degree of favor or disfavor. [. . .] Psychological tendency refers to a state that is internal to the person, and evaluating refers to all classes of evaluative responding, whether overt or covert, cognitive, affective, or behavioral" (Eagly & Chaiken, 1994, p. 1). So attitudes are indicated both by the type of moral reasoning that a person prefers and that he or she produces. Differences between preference and production should then reflect differences only regarding linguistic ability.

Although moral attitudes represent a necessary aspect of a person's morality, namely his or her own internal moral ideals rather than external standards, they are not sufficient for defining morally mature behavior.[1] Most people prefer high-level, 'Principled' moral reasoning (Rest, 1969; Lind, 1985a), because many people, even young children, often use high-stage arguments to justify their decisions and opinions on moral problems (Colby et al., 1987). Yet only when people show that they can competently apply their moral principles to everyday decision-making, we call their behavior morally mature. So the observation neither of rule-conformity nor of moral attitudes allows us to infer unambiguously a person's moral judgment competence.

Although, at first glance, the MJT looks so similar to the DIT that several authors have confused them, it differs greatly from this and other tests of moral judgment attitudes regarding seven distinct features. Its first and most important feature is the technique of counterarguments borrowed from Jean Piaget's 'clinical method.' I believe that counterarguments should not be neglected. The evaluation of counterarguments is a very common task in everyday life and fundamental for democratic interaction. Piaget used this technique in his research on mathematical and logical thinking (cited by Lourenço & Machado, 1996, p. 146). Common tests cannot be interpreted unambiguously because, by merely counting the number of correct solutions of mathematical tasks, they confuse

mathematical abilities with other factors, e.g., learning by heart and test-taking motivation. With counterarguments, Piaget showed, one can eliminate confounding factors and get a purer measure of a person's mathematical competence. The technique of countersuggestion was also used by Deanna Kuhn (1991) in her study of people's causal thinking. The first application of this technique in the moral domain, it seems, was in Kohlberg's (1958) study. To find out whether a subject really understood a certain level of moral reasoning, Kohlberg asked his interviewees to reason like a person whose decision was contrary to his or her opinion. "We are more interested," Kohlberg (1958) explained, "in assessing the degree to which any of an individual's judgments approximates the criteria of a moral judgment than in assessing how many of them he makes or acts upon" (p. 7). Unfortunately, Kohlberg and his colleagues (Colby et al., 1987) gave up this idea later. Contrary to Piaget's clinical method and to Kohlberg's earlier statement, these authors score the MJI merely by counting the number of arguments that a subject produced on each Stage, whereby they assign more weight to pro arguments than to con arguments. Counterarguments are merely seen as a potential source of measurement error. Colby et al. (1987) believe that "subjects may sometimes use less mature arguments in support of the nonchosen [decision] than that of which they are capable" (p. 161). For a more detailed critique, see Lind (1989; 1995).

Counterarguments are used in the MJT to probe a subject's ability to reason from a moral perspective. Subjects are asked not only to judge the acceptability of arguments supporting their decision on a (hypothetical) behavioral dilemma but also of arguments opposing their decision. A subject gets a high moral competency score © score) only if he or she rates arguments consistently by their moral quality instead by their agreement with his or her decision on the dilemma. A person gets a low C score if nonmoral considerations, like opinion-agreement, guide his or her judgments of others' arguments. That is, by virtue of test design, the C score reflects the degree to which moral rather than other factors determine subjects' judgment behavior.

In contrast, the DIT does not contain counterarguments and thus does not provide the possibility to probe a person's moral judgment competence. Rest (1979) himself interprets the DIT's P index as an attitude measure, namely "as the relative importance given to Principle moral considerations on making a moral decision" (Rest, 1979, p. 101). So while these two tests look alike, they pursue two completely different aims of measurement.

As a second important feature of the MJT, moral judgment competence is scored by the subject's own moral standards, rather than by external standards provided by the test designer. I designed the MJT's C index to be logically independent from any particular moral view. To get a high C score, the subjects need not prefer Principled moral reasoning. Yet they need to apply the same level of moral judgment to supportive arguments as to counterarguments, that is, to use moral principles consistently. The fact that the C index is calculated without reference to Principled moral standards, implies some interesting advantages. It opens the possibility to test empirically Piaget's (1976) hypothesis of affective-cognitive parallelism without committing a tautology, and without viewing the competence and the attitudinal aspect of morality as two separate domains of behavior (Lind, 1985a; 1985b). As Piaget (1976) noted, "affective and cognitive mechanisms are inseparable, although distinct" (p. 71). Moreover, the use of internal moral standards for scoring people's moral judgment competence, assures the cultural fairness of the MJT. If in some culture, subjects would think that the dilemmas in the MJT should be solved on the level of Conventional reasoning, they still could get a maximum C score if they were consistent in their judgments. In contrast, in such a case, the DIT and the MI would yield medium or low scores.

A third feature of the MJT is, as already mentioned, that both MJT dilemmas pull a Principled level of moral reasoning. If we had included dilemmas that pull only Conventional (Stage 3 and 4) or Preconventional (Stage 1 and 2) moral reasoning, we would have biased the MJT against high C scores. Not every dilemma needs to be solved on the highest level of moral principles. "The solution," Kohlberg (1958) noted, "must do justice both to what the self believes and yet meet the situation" (pp. 128-129). In fact, people's preferred level of moral reasoning varies considerably with dilemma-type (Krebs et al., 1991).

However, the dilemma-type varies within the level of Principled moral reasoning. The MJT contains a Stage 6-type dilemma (mercy killing/doctor) that seems to require Stage 6 reasoning, and a Stage-5 type dilemma (theft/workers) that might be optimally solved using Stage 5 moral reasoning. Empirical studies support this interpretation. On average, subjects rate Stage 6 as most acceptable in the mercy-killing dilemma, and Stage 5 in the theft-dilemma. However, the differences in ranking are small (Lind, 1985a). Overall, subjects rank the Stages of the arguments very much in the Kohlbergian order. In our study of 2098 first semester university students, more than 84% of the variance in the MJT can be

accounted for by the stages to which the items are keyed.[2] In a study of vocational school students, Heidbrink (1995) reports a mean Spearman rank-correlation of r = .93 (r^2 = .86) between the theoretical Stage numbers and the subjects' preference order of the Stages of moral reasoning. These figures compare well to the r^2 = .64 reported by Davison (see Rest, 1979, p. 231) for the DIT.

In contrast, in the DIT and the MJI, the differential moral 'pull' of dilemmas used has hardly received any attention. Some dilemmas used in these tests seem to demand only Conventional moral thinking. Therefore, these tests might inadvertently restrict subjects from getting high scores. For example, the Joe Dilemma in Kohlberg's MJI seems solvable without invoking Principle moral reasoning. So, normally we cannot expect any subject to get a MMS score higher than Stage 4 unless pressed by the interviewer to reason on a higher stage.

As a fourth feature, the MJT combines the above mentioned three potential determinants of a person's judgment behavior to form a three-factorial single-subject experiment: The first factor is dilemma-type. Two dilemmas seemed sufficient for a test that is exclusively used in research and evaluation studies. (The MJT, as most other tests of moral development, has not been designed for diagnosing individual persons.) The second factor, called opinion-agreement, also consists of two categories: arguments that support, and those that oppose, the subject's opinion on a particular dilemma. If persons do not state an opinion, we try to infer it from his or her ratings of the arguments (Lind, 1985a). The third factor is the moral quality, or Stage, of reasoning to which the arguments are keyed. We used all of Kohlberg's original 6-Stages to cover as wide a developmental range as possible with the MJT. So the MJT consists of 24 argument-items, six pro and six con arguments, in each dilemma, each argument representing a different type, or Stage, of moral reasoning as described by Kohlberg (1958; 1984). Within each dilemma, subjects are asked to judge the protagonist's decision (e.g., "Was the doctor right or wrong to help the mortally ill woman to die?") and to judge the pro and con arguments made by others.

Most test-items (arguments) of the MJT were selected from interview-material. All items were rated by experts on Kohlberg's stage model and were pretested with about 20 subjects. This helped to ensure that the arguments represented the six Kohlberg-stages of moral reasoning, and to weed out items that sounded too far-fetched, cynical, funny, or bombastic. To find a pro and con statement on every stage and for every dilemma-type was difficult, yet it was possible. Inte-

restingly, it was harder finding arguments for lower stages than for higher stages (which, by the way, we did not anticipate). In spite of this difficulty, I felt that a strict orthogonal balanced design is extremely important because it assures that the factors do not correlate with each other. This feature of the MJT reduces considerably the problem of inferential ambiguity because it eliminates the problem of collinearity among these three factors (Anderson, 1991). Only an orthogonal design lets us unambiguously identify the contribution of each of these three factors to an individual's response pattern. I see no convincing reason for mimicking some 'natural distribution' of arguments, no matter whether we could ever achieve such an aim. We constructed the MJT to provide a strict and unbiased test for developmental theories. So we did not select any items or coding procedures to maximize the correlation of the C score with age, which would bias the test against the detection of regressions.

The fifth distinct feature of the MJT is its C index, whose construction is described in detail by Rest et al. (this volume). The C index is designed to reflect the degree to which one is able to judge both pro arguments and counterarguments consistently. To a small degree, the C index also reflects judgment consistency across dilemma-types. Therefore, we also developed a more complex C* index that accounts for this fact by correcting the total response variances by dilemma-related variances (Lind, 1978; 1985a). However, the empirical differences between the C and the C* were so small that we decided to use the C index, which is less complex. The meaning of the C index, however, depends fully on the special design of the MJT. If the MJT had not been designed as a competence test, no indexing method could have made up for this shortcoming. "If a test is to yield stage structure," Kohlberg (1984) once noted, "a concept of that structure must be built into the initial act of observation, test construction, and scoring" (p. 402). In contrast, Rest and his colleagues use the C score with the DIT, neglecting the fact that the C score has a totally different meaning with the DIT, which has not been designed as a test of moral competence. The DIT-C score may reflect something like moral rigidity or insensitivity to the differential moral demands of a dilemma, rather than moral judgment competence. The overall low correlation of the DIT-C score with level of education supports this interpretation.

Sixth, the MJT uses a person's moral judgment consistency as an indicator of his or her judgment structure. The structure of a person's moral judgment behavior is not manifested in any single response but becomes visible only in the re-

lationships between his or her responses (Mischel & Shoda, 1995). Inconsistent response patterns, which may look "meaningless" or "random," are thus viewed as an expression of low moral judgment competence. Eliminating persons with apparently 'meaningless' or 'random' responses from further analysis thus would not only be wasteful but would also bias this analysis against low scoring subjects. Furthermore, we do not regard a person's response consistency as an attribute of the test, but as an attribute of an individual's interaction with the situation specified by the test. So with the MJT, no items have been added, or discarded, to maximize 'reliability,' that is, to increase artificially response consistency or stability over time.

This is in contrast with classical test theory. According to this theory, the most important criterion for test construction should be high 'reliability.' To get high reliability, test-items must be constructed (or selected) so that the test maximizes individual differences and becomes insensitive to change. Structural differences between people are not accounted for at all but are viewed as unreliability of the test. Obviously, any test streamlined to meet these criteria is biased against cognitive-structural theories.

Yet another approach is taken by Rest et al. who use reliability criteria to purify their data. "We have found that purging the subjects gives much clearer results and better reliability than leaving all the subjects in" (Rest, 1979, p. 93). Rest reports a loss of DIT questionnaires in the 2 to 15% range (p. 93). Gielen and Markoulis (1994) report even higher losses in cross-cultural research. Sometimes more than 50% of the sample had to be discarded because of the M items and the Consistency check. However, such practice is not justified, or even implied, by classical test theory to which the idea of an "inconsistent person" is unknown. Moreover, through artificially boosting 'reliability' DIT research is biased against cognitive-structural theories. Rest (1979) once wrote that he prefers high "test-retest reliability" over any kind of "theoretical advantage" (p. 104), although his own findings speak against the applicability of reliability theory. So I conclude that Rest's approach is methodological absolutism rather than old-fashioned empiricism.

Similarly, the scoring of the MJI seems to employ techniques to reduce inconsistency in the data rather than using this information to characterize a person's level of moral development (Krebs et al., 1990; Lind, 1989; Rest, 1979, pp. 69-70). Moreover, Kohlberg overlooked structural information contained in individuals' response patterns because he believed that already "each item in the ma-

nual clearly reflects the structure of the stage to which it is keyed" (Kohlberg, 1984, p. 403).

The seventh feature of the MJT is that it is a developmental measure. To allow us to measure not only the progression of moral development but also its regression, the validity of the MJT was established without reference to criteria like chronological age or invariant sequence. As derived from Kohlberg's (1964) definition, the only criterion for developmental progress (or regression) is the subjects' ability to apply their moral ideals consistently to pro and con arguments. The validity of this criterion is established both through the experimental design of the MJT, and through research that shows that the C score cannot be faked upward.

In contrast, Kohlberg stated frequently that the major validity standard for the MJI was a high positive correlation with age and the corroboration of the invariant-sequence hypothesis. "We wish to provide evidence for [. . .] age differences in various formal attributes of moral thinking," he wrote in 1958 (p. 17). "The validity criterion of moral judgment development is [. . .] that of an organization passing through invariant stages," he also insists in 1984 (p. 194). Kohlberg saw the MJI invalidated by any case of regression that could not be attributed to 'measurement error,' and revised his test several times to meet this criterion. Therefore, I doubt, as Anne Colby (1995) maintains, that the MJI "is perfectly capable of picking up either progression or regression." For Rest (1979), age trends were not only sufficient, as for Kohlberg, but necessary, or "crucial," for proving the validity of his DIT (p. 143).

Testing the theory

"Entia non sunt multiplicanda sine necessitate" (Do not multiply your constructs without necessity), advised William of Occam (1290-1350) to curb the scholasticism of the Middle Ages. I felt that Occam's razor should be applied to strip the theory of unnecessary overhead, including the stipulation of invariant developmental sequence. This all resulted in a more concise theory, which I call the Dual-Aspect-Theory of moral behavior and development (Lind, 1985a; 1985b; 1993; in press). It consists of three basic testable hypotheses:

1. Morality has both an affective aspect (i.e., moral attitudes, values etc.) and a cognitive or competence aspect. I prefer the word 'aspect' over 'component' to signify that cognition can be distinguished but not be separated from affect. "Affective and cognitive mechanisms," as Piaget (1976) noted, "are inseparable, although distinct: the former depend on energy, and the latter depend on structure" (p. 71; see also Lind, 1985b). The hypothesized distinguishableness of attitudes and cognitive competence can be tested, as Rest (1979) and Emler and his colleagues (1983) have suggested, by so-called 'faking experiments.' In these experiments, subjects are asked to fill out a moral judgment test twice, first regularly and a second time under the experimental instruction to fake high scores. Emler et al. (1983) believed the instruction to fake like a liberal/leftist person, to be most useful because of the correspondence of Principled moral reasoning and liberal/leftist political ideologies (see also Gross, 1996; Krebs et al., 1991). Indeed, they found that subjects could easily simulate P scores higher than their own. Two further faking experiments (Markoulis, 1985; Barnett et al., 1995) also support Emler et al.'s (1983) conclusion that the DIT is not a cognitive-structural (=competence) test, but rather varies with affective-political commitments of the individual taking the test.
Kohlberg's interview method, the MJI, has been also suspected to be fakable. Quoting Rom Harré and Robert Hogan, Krebs et al. (1991) argue that the MJI "is susceptible to impression management, and thus, the stages individuals display may depend on the audiences to which they direct their communication" (p. 156). However, Ann Colby (1995), the senior author of the MJI scoring manual (Colby et al, 1987), contradicts this allegation: "According to my reading of the literature, the MJI cannot be faked high." Neither side quotes any empirical evidence dealing with this issue.

In contrast, the MJT proved resistant to faking in three independent faking experiments. Using the same experimental instruction and the same kind of subjects as Emler et al. (1983), Lind (1993) found that university students could not be instructed to fake the C score of the MJT upward. Kindergarten teachers, who were instructed to simulate the moral judgment of real persons, namely their peers, could not do this when their peer's C actual score was higher than their own (but high scorers could, of course, imitate low scorers). Moreover, the simulation of low scoring target persons by high scoring subjects was more precise than the simulation of high scoring target persons by low scoring subjects. The correlations between simulated and actual C scores in these two groups were r = .71 and .23, respectively (Wasel, 1996). Bühn (1995), who used yet another instruction, also found that the C score could be faked downward but not upward. Rest and his colleagues hypothesize that these findings may be accounted for by subjects' tiredness or drop in motivation when taking the MJT a second time. Yet none of these studies observed any sign of tiredness or drop of motivation. If such factors had prevented subjects from getting high MJT scores, they should have been even more forceful in preventing high P scores in DIT research as the DIT is much longer, and more tiresome to take, than the MJT. As we have seen, however, many subjects did fake DIT scores upward.

2. So the MJT seems not to vary, as the DIT does, as a function of the test taker's intentions and attitudes, at least as tapped in faking experiments. Therefore, the educational effects documented by the MJT must be regarded as true gains in moral judgment competency. They cannot be discarded as 'Hawthorne-effects,' that is, as effects resulting from the students' tendency to respond favorably to We hypothesized that moral competency, like other competencies, must be taught, that is, moral competencies do not emerge independently of educational support (Lind, 1996a). This contrasts with more traditional cognitive-developmental perspectives that play down explicit instruction as a moving force in development (Piaget, 1965; Kohlberg, 1984). MJT research clearly supports the education hypothesis. In my work, educational attainment has been the factor that explains individual differences of level of moral judgment competence. No other factor (age, gender, socioeconomic status etc.) has a comparable impact when education is controlled for (Lind, 1985a; 1993; in press).

3. We hypothesized that moral competencies can regress. We believe that both an active individual and an active environment (e.g., involving explicit instruction) are necessary conditions for moral learning. That is, the lack of either condition may stall or reverse the developmental process. However, I predict that regression will not occur if people have reached a high enough level of moral competence for 'self-sustaining' development to set in (Lind, 1996a), most likely after completion of college education (e.g., Kohlberg, 1984). The regression-hypothesis is sharply rejected by many Piagetian theorists (Colby et al., 1987; Kohlberg, 1984; Rest, 1979; Rest et al., this volume).

Two major MJT studies support the regression-hypothesis (Lind, 1993). First, in a cross-sectional study of adolescents between 14 and 23 years of age, all of whom had graduated from grade 9 or 10 to enter vocational schools and jobs, we found a marked erosion of C scores following the departure from formal schooling. Second, in a longitudinal study, medical students, compared with other students, experienced a drop of C score. Interestingly, in a longitudinal study of Finish medical students, Helkama (1987) reports the same drop in moral judgment competence, using Kohlberg's interview method. In both cases of drop (i.e., among vocationalists and medical students), I believe, students' moral competence eroded because opportunities for practicing moral judgment competence became scarce. This interpretation is supported by a recent study documenting a positive correlation between C score gains and opportunities for role-taking and guided reflection (Lind, 1996a).

Little can be said about the empirical findings that Rest et al. presented. As I explained above, the C score gets its meaning only through the particular design of the MJT. Used with the DIT, the C index seems to reflect moral rigidity, rather than moral judgment competence. Therefore, we would not expect the DIT-C index to correlate with level of education. Rest et al.'s data (Tables 2 to 6) support this interpretation. The unique correlation of the DIT-C index with level of education is very low.

The high correlation between the P index and progressive political ideologies (Table 9) confirms my assumption that the DIT is largely an attitude test. Interestingly, the effect-sizes of moral teaching are low for both DIT-C and P indices (Table 8), much lower than for Kohlberg's MJI (Lind, 1996b). The only exception is the Penn-study, which used a 'direct teaching' method, aiming at attitude change rather than competence development. Unsurprisingly, the P index coded as it is with regard to attitudes, but not the C index, was highly sensitive to

methods of attitude change (d=1.25, r = .49).

To provide a better basis of comparison, I have re-analyzed our four-wave longitudinal study of N=2098 German university students, conducted in the early 80's (Lind, 1985a), calculating a P score for the MJT analogously to Rest's P score. As MJT-P index I used the correlation between an individual's actual rating of others' arguments with an ideal rating in which Principled moral reasoning is preferred to all other Stages of reasoning. This correlation is 1.00 (and the P score is 100) if a subject prefers Principled moral arguments over arguments that represent lower stages. If the subject prefers some or all lower stage arguments over Principled arguments, his or her P score is lower. Unlike the C score, the P score can be -100 if the preference order is completely reversed. The comparison of both the MJT-P and C indices reveals some interesting differences. First, the average MJT-P scores are much higher (their Øs ranged from approx. 76 to 81) than the average C scores (Øs ranging from approx. 40 to 48). This supports my assumption that the P score reflects largely moral attitudes rather than moral competence. Second, during the study, the average MJT-C index changed much more pronouncedly than the average P index, namely by 5 versus 2 points, though the P score can vary much more than the C score, namely from -100 to +100. This supports my assumption that the C index is more sensitive to educational effects. Third, unexpectedly, the C index showed a much higher test-retest correlation over two-year-intervals than the MJT-P index:

2-year test-retest correlation r	1st to 5th semester (N=830)	5th to 9th semester (N=454)	9th to 13th semester (N=328)
MJT-C scores:	0.47*	0.54*	0.56*
MJT-P scores:	0.25*	0.20*	0.36*

* $p<0.001$

The comparably high test-retest correlations of the C index surprised us because, as I explained above, the MJT was not designed to achieve any kind of differential stability. It may reflect the fact that over the tested time range, all students experience a similar quality and quantity of education. The low correlation of the P index may look surprising to Jim Rest. It challenges the hypothesis that

the P index beats all benchmarks.

In summary, the crux of the MJT is its design as a moral competence test. It has shown to be valid and unbiased. Therefore, the MJT fills a void not only in research and theory development but also in the practical application of cognitive-developmental psychology. Kohlberg's interview method, still among the best measures, substantially deviates from his original ideas, confounding moral competencies with moral attitudes. Other measures like the DIT tap merely moral attitudes. At best, these tests can be used as indirect measures of moral competencies because, as we have found, moral attitudes correlate very high with moral judgment competence when the motivation to fake is absent (Lind, 1985a; 1995). However, for some research purposes like the evaluation of moral education, these tests are insufficient because they do not allow us to discard other explanations (e.g., faking, Hawthorne-effect) for the effects measured. Moreover, the DIT-P index is insufficiently sensitive to moral-cognitive interventions with children and adolescents because it is confined to Principled moral reasoning (Lind, 1996b; Rest, 1979, p. 244). So new scoring methods, like Stephen Thoma's (1994) 'Utilizer (U) score,' enhance the usefulness of tests like the DIT but not their validity. In contrast, the MJT-C index cannot be faked upward and is sensitive to education-induced changes in children and in adults (Lind, 1993). So the MJT enables us to weed out popular, yet inefficient, education programs like value-indoctrination, and to show the benefits of high-quality moral teaching like the Moral Dilemma Discussion method (Blatt & Kohlberg, 1975; DeVries & Zan, 1994), the Just Community (Althof, 1992; Power et al., 1989) program, and the Democratic School (Mosher, 1993).

References

Althof, W. (Ed.) . (1992). Moral Education Forum, 17 (special issues). With contributions by W. Althof, P. Dobbelstein, D. Garz, G. Lind, F. Oser, S. Reinhardt, and H. Schirp.

Anderson, N.H. (1991). Moral-social development. In N.H. Anderson (Ed.), Information integration theory, Vol. III: developmental theories (pp. 137-187). Hillsdale, NJ: L. Erlbaum.

Barnett, R., Evens, J., & Rest, J. (1995). Faking moral judgment on the Defining Issues Test. British Journal of Social Psychology, 34, 267-278.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. Psychological Bulletin, 88, 1-45.

Blatt, M., & Kohlberg, L. (1975). The effect of classroom moral discussion upon children's level of moral judgment. Journal of Moral Education, 4, 129-161.

Bühn, A. (1995). Die Rolle von moralischer Urteils- und Wahrnehmungsfähigkeit im Prozeß der Fach- und Berufswahl von Studierenden. Unpublished Master thesis [Diplomarbeit], University of Konstanz.

Colby, A. (1995). Personal communication.

Colby, A., Kohlberg, L., Abrahami, A., Gibbs, J., Higgins, A., Kauffman, K., Lieberman, M., Nisan, M., Reimer, J., Schrader, S., Snarey, J., & Tappan, M. (1987). The measurement of moral judgment: Vol. 1. Theoretical foundations and research validation. New York: Cambridge University Press.

DeVries, R. & Zan, B. (1994). Moral classrooms, moral children: Creating a constructivist atmosphere in early education. New York: Teachers College Press.

Eagly, A.H., & Chaiken, S. (1994). The psychology of attitudes. New York: Harcourt Brace.

Emler, N., Renwick, S. & Malone, B. (1983). The relationship between moral reasoning and political orientation. Journal of Personality and Social Psychology, 45, 1073-1080.

Gielen, U.P. & Markoulis, D. (1994). Preferences for principled moral reasoning: A developmental and cross-cultural perspective. In L.L. Adler & U.P. Gielen (Eds.), Cross-Cultural Topics in Psychology (pp. 73-87). Westport, CN: Praeger.

Gross, M.L. (1996). Moral reasoning and ideological affiliation: a cross-national study. Political Psychology, 17, 317-338.

Gulliksen, H. (1987). Theory of mental tests. Hillsdale, NJ: L. Erlbaum.

Hays, W. (1963). Statistics for psychologists. New York: Holt.

Heidbrink, H. (1995). Zur Existenz von Moralstufen. Eine strukturelle Validitätsanalyse. In E.H. Witte (Ed.), Soziale Kognition und empirische Ethikforschung (pp. 77-96). Lengerich/Scottsdale AZ: Pabst.

Helkama, Klaus (1987). The fate of moral reasoning in medical school: a two-year longitudinal study. Unpublished manuscript. University of Helsinki, Finland.

Keasey, C.B. (1973). Experimentally induced changes in moral opinions and reasoning. Journal of Personality and Social Psychology, 26, 30-38.

Kohlberg, L. (1958). The development of modes of moral thinking and choice in the years 10 to 16. Unpublished dissertation, University of Chicago, Illinois.

Kohlberg, L. (1964). Development of moral character and moral ideology. In M.L. Hoffman & L.W. Hoffman (Eds.), Review of child development research: Vol. I (pp. 381-431). New York: Russel Sage Foundation.

Kohlberg, L. (1984). Essays on moral development: Vol. II. The psychology of moral development. San Francisco, CA: Harper & Row.

Krebs, D.L., Vermeulen, S.C.A., Carpendale, J.I. & Denton, K. (1991). Structural and situational influences on judgment: The interaction between stage and dilemma. In W.M. Kurtines & J.L. Gewirtz (Eds.), Handbook of moral behavior and development: Vol. 2. Research (pp. 139-169), Hillsdale, NJ: Erlbaum.

Kuhn, D. (1991). The skills of argument. Cambridge, MA: Cambridge University Press.

Levy-Suhl, M. (1912). Die Prüfung der sittlichen Reife jugendlicher Angeklagter und die Reformvorschläge zum § 56 des deutschen Strafgesetzbuches. Zeitschrift für Psychotherapie, 232-254.

Lind, G. (1978). Wie mißt man moralisches Urteil? In Portele, G. (Ed.), Sozialisation und Moral (pp. 171-201), Weinheim: Beltz.

Lind, G. (1982). Experimental Questionnaires: A new approach to personality research. In A. Kossakowski & K. Obuchowski (Eds.), Progress in psychology of personality (pp. 132-144), Amsterdam, NL: North-Holland.

Lind, G. (1985a). Inhalt und Struktur des moralischen Urteilens. [Content and structure of moral judgement.] Unpublished doctoral dissertation, University of Konstanz.

Lind, G. (1985b). The theory of moral-cognitive judgment: A socio-psychological assessment. In G. Lind, H.A. Hartmann & R. Wakenhut (Eds.), Moral development and the social environment (pp. 21-53). Chicago, IL: Precedent Publishing.

Lind, G. (1985c). Growth and regression in moral-cognitive development. In C. Harding (Ed.), Moral dilemmas (pp. 99-114). Chicago, IL: Precedent Publishing.

Lind, G. (1986). Cultural differences in moral judgment? A study of West and East European University Students. Behavioral Science Research, 20, 208-225.

Lind, G. (1989). Measuring moral judgment: A review of 'The Measurement of Moral Judgment' by Anne Colby and Lawrence Kohlberg. Human De-

velopment, 32, 388-397.
Lind, G. (1993). Moral und Bildung. Zur Kritik von Kohlbergs Theorie der moralisch-kognitiven Entwicklung. Heidelberg: Asanger.
Lind, G. (1995). The meaning and measurement of moral competence revisited - A dual aspect model. Invited paper presented at the meeting of the SIG 'Moral Development and Education,' 1995 AERA conference, San Francisco.
Lind, G. (1996a). Educational environments which promote self-sustaining moral development. Paper presented at Division E 'Counseling and Human Development,' 1996 AERA conference, New York.
Lind, G. (1996b). The optimal age of moral education. Paper presented at the SIG 'Moral Development and Education,' 1996 AERA conference, New York.
Lind, G. (in press). The psychology of moral competencies. Hillsdale, NJ: Lawrence Erlbaum.
Lind, G. & Wakenhut, R. (1985). Testing for moral judgment competence. In G. Lind, H.A. Hartmann, & R. Wakenhut (Eds.), Moral development and the social environment (pp. 79-105). Chicago: Precedent Publishing.
Lourenço, O. & Machado, A. (1996). In defense of Piaget's theory: a reply to 10 common criticisms. Psychological Review, 103, 143-164.
Markoulis, D. (1989). Political involvement and socio-moral reasoning: Testing Emler's interpretation. British Journal of Social Psychology, 28, 203-212.
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review, 102, 246-268.
Mosher, R. (1993). Preparing for citizenship - The democratic school. New York: Praeger.
Penn, W. (1990). Teaching ethics -- A direct approach. Journal of Moral Education, 19(2), 124-138.
Piaget, J. (1965). The moral judgment of the child. New York: The Free Press.
Piaget, J. (1976). The affective unconscious and the cognitive unconscious. In B. Inhelder & H.H. Chipman (Eds.), Piaget and his school (pp. 63-71). New York: Springer.
Pittel, S.M. & Mendelsohn, G.A. (1966). Measurement of moral values: a review and critique. Psychological Bulletin, 66, 22-35.
Power, F.C., Higgins, C. & Kohlberg, L. (1987). Lawrence Kohlberg's Approach to Moral Education. New York: Columbia University Press.
Rest, J. (1969). Level of moral development as a determinant of preference and comprehension of moral judgments made by others. Journal of Personality, 37(1), 220-228.
Rest, J. (1979). Development in judging moral issues. Minneapolis, MI: University of Minnesota Press.
Rest, J., Thoma, S., & Edwards, L. (1996). Designing and validating a measure

of moral judgment: Stage preference and stage consistency approaches. Journal of Educational Psychology (this issue).
Thoma, S. (1994). Moral judgments and moral action. In J. R. Rest, & D. Narváez (Eds.), Moral development in the professions: psychology and applied ethics (pp. 199-211). Hillsdale, NJ: Erlbaum.
Wasel, W. (in prep.). Is morality a competence? An experimental test of a central assertion of cognitive-developmental theory. British Journal of Social Psychology (submitted for publication).

Author's Note

Major portions of the research reported in this paper have been funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the international longitudinal research project "Hochschulsozialisation" (1973-1986). I thank my former colleagues Tino Bargel, Barbara Dippelhofer-Stiem, Gerhild Framhein, Hansgert Peisert, Johann-Ulrich Sandberger, and Hans Walter, for use of their data. The MJT has been developed in close collaboration with Tino Bargel, Horst Heidbrink, and Roland Wakenhut. For comments and suggestions on earlier drafts and related papers, I thank Anne Colby, Michael Gross, Fritz Oser, Jim Rest, Betsy Speicher, Stephen Thoma, Terry Thorkildsen, Wolfgang Wasel. Speaking English only as a second language, I am especially grateful to Jim Fearns, Gisela Kusche and Michael Pressley, for their suggestions concerning style and readability of my paper.

Inquiries should be sent to Georg Lind, Sozialwissenschaftliche Fakultät, Universität Konstanz, D-78434 Konstanz, Germany, or e-mail: Georg.Lind@Uni-Konstanz.de.

Notes

1. Typically, measures of moral attitudes correlate low to moderately (approx. r = .30) with measures of moral behavior (Pittel & Mendelsohn, 1966; Blasi, 1980); only in some carefully designed experiments, e.g., in the McNamee's study of helping behavior (cited in Kohlberg, 1984, pp. 520) and in Jacob's study of Ss behavior in the Prisoner's Dilemma game (cited in Rest, 1979, pp. 181-185), were impressively clear relationships between test scores and the predicted behaviors found.

2. Calculated as T^2 (Hays, 1963, p. 382) from the linear polynomial of the following means and standard deviations of the acceptability scores of Stages 1 through 6:

0	8.03;	7.02;	9.65;	15.2;	20.6;	21.0
s	5.03	4.76	5.27	4.67	4.26	4.33

	Sum of Squares	df	Mean Square	F	p-level
Effect	327642.8	1	327642.8	10001.7	0.00
Error	60734.6	1854	32.8		