The Reliability and Validity of Objective Indices of Moral Development

Mark L. Davison and Stephen Robbins
University of Minnesota

The present paper addresses three issues surrounding Rest's *Defining Issues Test*, an objective test of moral development based on Kohlberg's six-stage theory of moral development. Those issues are (1) the stability of test scores over time; (2) correlation of scores with Kohlberg's interview measure of moral development; and (3) the insensitivity of its scoring procedure, which ignores responses to all items keyed to lower stages. In two age heterogeneous samples, total score test-retest reliabilities were generally in the high .70's or low .80's, regardless of which of several scoring schemes was used. In another age heterogeneous sample, the correlation with scores on Kohlberg's test was .70; but in two age homogeneous samples, the correlations were about .35 and .20. These validity coefficients suggest that (1) the common variance shared by Rest's and Kohlberg's tests in age heterogeneous samples can be attributed to the fact that scores on both tests increase with age and (2) the two tests cannot be considered equivalent measures of the same construct differing only in format. Results also indicated that an empirically weighted scoring scheme is more sensitive to longitudinal change than is Rest's *P* score. This sensitivity to longitudinal trends is an important property for tests such as Rest's which claim to be developmental and are frequently used to assess educational change. The empirically weighted sum had a significantly higher test-retest reliability ($p < .05$) than did a simple sum of item responses, and it had a significantly higher correlation with Kohlberg's measure than did a theoretically weighted sum.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 3 Summer 1978 pp. 389–401

Kohlberg (1969, 1971) has outlined a six-stage theory of moral development. Each successive stage in the theory is said to be characterized by moral reasoning which is more complex, more comprehensive, more integrated, and more differentiated than the reasoning of earlier stages. According to Kohlberg's theory, the child develops by movement through a sequence of steps with the following bases of moral order:

Stage 1: Internal compulsion and power,
Stage 2: Simple interpersonal exchange and need satisfaction,
Stage 3: Maintaining positive, interpersonal relationships,
Stage 4: Maintaining social order,
Stage 5A: A legitimate social contract,
Stage 5B: Intuitive individualism and humanism and,
Stage 6: Individual conscience.

Kohlberg has developed an interview procedure for assessing an individual's level of moral development.

Rest (1974; Rest, Cooper, Coder, Masanz, & Anderson, 1974) has developed an objective measure of moral level, called the *Defining Issues Test* (DIT), which consists of six stories describing moral dilemmas. Rest (1976b) reports that the DIT has been used in over 100 studies involving 5,000 subjects. After reading each story, the subject is first asked to answer a yes-no question indicating how he/she thinks the

central character of the story ought to respond to the dilemma. The respondent is next asked to rate each of several issues on a five-point scale of importance in deciding what ought to be done. Finally, the respondent is asked to rank order the four issues which he/she thinks are the most important. Each of the issues reflects reasoning characteristic of either Stage 2, 3, 4, 5A, 5B, or 6 in Kohlberg's theory.

Because Rest's issues can be divided into six sets, one for each of Stages 2, 3, 4, 5A, 5B, and 6 in Kohlberg's theory, the test readily yields subscale scores or stage scores. The problem is to combine information from all six stage scores into a single index of the individual's overall developmental level. Loevinger (Loevinger & Wessler, 1970; Loevinger, Wessler, & Redmore, 1970) would call such an index a measure of the respondent's core functioning.

As an overall measure of development, Rest (1974; Rest et al., 1974; Rest, 1975) uses the "*P* score." This is computed by giving each person four points for each Principled Issue (item keyed to Stage 5A, 5B, or 6) ranked first; three points for each Principled Issue ranked second; two points for each Principled Issue ranked third; and one point for each Principled Issue ranked fourth. Typically, *P* is reported as a percentage of the maximum possible *P*, and this convention will be followed in the present paper. Although Cooper (1972) and Rest (1975) have presented convincing empirical support for *P*, Loevinger (1976, p. 223) has pointed out that *P* lacks intuitive appeal as a measure of overall development because it incorporates no information from issues keyed to the first three stages.

The logical alternatives to *P* include a sum of importance ratings for all items or, possibly, a weighted sum of item importance ratings. In the following research, the reliability and validity of Rest's *P* score have been compared to that of a simple sum of item responses, a theoretically weighted sum of responses, and an empirically weighted sum of responses. Besides indicating an alternative to replace *P*, the following research updates the available information on the reliability and validity of Rest's test of moral development.

Method

Subjects

In evaluating the reliability and validity of the DIT, the following seven subject groups were used:

Group 1 consisted of 160 subjects: 40 junior high, 40 senior high, 40 college undergraduates, and 40 graduate students. The graduate students consisted of 25 male seminary students and 15 male doctoral students in moral philosophy. The remaining 120 subjects were approximately evenly split between males and females. This sample was used to estimate the test's internal consistency reliability, its correlation with a measure of cognitive ability (*DAT Manual*, 1966), its correlation with a measure of comprehension of moral issues, and its correlation with two measures of attitudes related to moral development. (For a more complete description of this group and the instruments administered to it, see Rest et al., 1974.)

Group 2 contained 1,080 subjects ranging in age from 15 to 82. The subjects were equally split between those with a junior high school, senior high school, college, and graduate education. Of these subjects, 424 were male and 452 were female. The sex of 203 subjects was unknown. The data from this group were used solely to estimate the empirical weights for the empirically weighted sum. As will be shown, these weights are of theoretical interest in their own right.

Group 3 consisted of 123 subjects. All were subjects in moral education projects, and all had taken the DIT twice over an interval of from one week to five months. Evaluations of the individual moral education projects showed no significant change in DIT scores as a result of any project. These subjects ranged in age from 16 to 56. Of the 88 for whom sex information was available, 43 were males and 45 were females. Thirty-eight had a junior high school education,

24 a senior high education, 34 a college education, and 27 some type of graduate training. These data were used to estimate test-retest reliabilities.

Group 4 was used to obtain a second test-retest reliability estimate. This group contained 19 ninth graders, called Group 4a, and 33 Australian college-age students from a study by McGeorge (1975), called Group 4b.

Group 5 was used to estimate the correlation between scores on Kohlberg's interview measure and scores on Rest's paper-pencil measure. This sample contained a total of 213 subjects. Seventy-four were ninth grade males living in Minnesota (McColgan, 1975). The remaining 139 were college students, about half male and half female, enrolled in an introductory psychology course at the University of Texas (Froming & McColgan, 1977). Rest (1976a) reported that only 2 of 23 studies found significant sex differences in moral development. Consequently, the fact that the ninth grade group contained only males is probably unimportant.

Group 6 contained 54 subjects, 21 subjects who were ninth graders when first tested in 1972 and 33 who were eleventh graders when first tested in 1972. Their ages ranged from 14 to 18. All 54 were retested in 1974 and again in 1976. The group, which is more completely described by Rest (1975), included 21 males and 33 females. Their data were used to study longitudinal changes as assessed by each moral development index.

Group 7 contained 21 subjects who were high school juniors and seniors when first tested in 1974. All were retested in 1976. Of these 21 subjects, 9 were males and 12 females. Like the data from Group 6, the data from Group 7 were used to study longitudinal changes as assessed by each moral development index.

Indices

As stated above, four indices were compared: (1) Rest's *P* score; (2) a simple sum of item responses; (3) a theoretically weighted sum of item responses; and (4) an empirically weighted sum. The *P* score has been described above. The simple sum (*SS*) is a sum of the importance rating for each item after reverse scoring items keyed to the lowest stages (2, 3, and 4). Both weighted sums (*TS* and *ES*) are weighted sums of double-centered item ratings. The double-centered rating of person i for item j, $\tilde{x}_{ij}$ is defined as:

$$\tilde{x}_{ij} = x_{ij} - x_{i.} - x_{.j} + x_{..} \qquad [1]$$

where

X_{ij} = the importance rating given by person i to item j.

$x_{i.} = (\frac{1}{n}) \sum_{j} x_{ij}$, the mean importance rating given by person i to the n items.

$x_{.j} = (\frac{1}{N}) \sum_{i} x_{ij}$, the mean response to item j in the standardization sample of 1080, and

$\tilde{x}_{..} = (\frac{1}{Nn}) \sum_{i} \sum_{j} x_{ij}$, the grand mean of responses in the standardization sample.

Double centering item ratings has two effects. First, subtracting the mean response to each item ($x_{.j}$) adjusts for the fact that some items are more popular regardless of their stage content. Second, subtracting the mean importance rating of subjects i, $x_{i.}$, adjusts for a subject's tendency to rate items as more or less important regardless of their content. In other words, double centering adjusts for popularity of items independent of their content and for subjects' tendency to rate items as important regardless of their content.

Both the theoretically and empirically weighted sums have the form:

$$s_i = \sum_{j} w_j \tilde{x}_{ij} \qquad [2]$$

where

s_i = the sum for person i

w_j = a weight for item j, and

x_{ij} = subject i's double-centered response to item j.

For the theoretically weighted sum (*TS*), the weight w_j was simply the number of the stage to which the item is keyed. The theoretically weighted sum is an analog of Kohlberg's Moral Maturity Score, the scoring scheme used with his interviews. Because of its similarity to Kohlberg's scoring scheme, the theoretically weighted sum was included for study in this research.

For the empirically weighted sum (*ES*), the weight was proportional to the item's projection onto the first principal component of the double-centered response matrix in the standardization group of 1,080. The rationale for choosing the empirical weighting scheme lies in previous work. Davison (1977) and Davison, Robbins, and Swanson (in press) have proposed and presented evidence in support of a unidimensional metric unfolding model for objective moral judgments.

As applied to developmental data in the unidimensional case, the metric unfolding model assumes that (1) each person can be characterized by a score, y_i, representing his or her developmental level; (2) each item can be characterized by a score, z_j, representing the developmental level of the content stated in that item; and (3) the importance of item j for subject i is inversely related to the squared difference between y_i and z_j:

$$x_{ij} = 6 - (y_i - z_j)^2 \qquad [3]$$

In effect, Equation 3 states that subjects will assign high importance ratings to a statement if the developmental level reflected in its content corresponds to the subject's own developmental level. The constant 6 in this equation simply reverses scoring so that high importance ratings correspond to high scores on variable x_{ij}.

Schönemann (1970) has shown that the least squares estimate of the person score (y_i) is a weighted sum of double-centered responses to items, where the weights are proportional to the projections of items onto the first principal component of the double-centered response matrix. Except for an additive and multiplicative constant, the item weights are estimates of item scale values (z_j) along the hypothesized moral development dimension. Because item weights are linearly related to item scale values, items keyed to higher stages should have higher weights.

The empirical weight estimation scheme represents empiricism at its blindest. If it had turned out that higher stage items did not have higher weights, there would have been no logical basis for arguing that the empirical sum reflected the hierarchical moral development sequence in Kohlberg's (1969) theory. On the average, however, higher stage issues did have higher weights, as will be reported below.

In the following section, empirical item weights will first be reported. Then the four scoring schemes will be compared in terms of their internal consistency reliabilities, test-retest reliabilities, correlations with other measures, and sensitivities to longitudinal trends.

Results

Empirical Item Weights

Only the empirical weight estimates from Group 2 were used in computing the empirically weighted sum; nevertheless, Table 1 reports the weight estimates derived from Samples 1 and 2. In Group 1 (consisting of 160 subjects) the mean weight for items keyed to Stages 2, 3, 4, 5A, 5B, and 6 were −.34, −.17, −.13, .34, .35, and .31, respectively. Stages 2, 3, and 4 were ordered as expected, but not Stages 5A, 5B, and 6. However, both Stages 5A and 6 included one statement with a highly deviant weight, which partially accounts for the standard deviations from these two stages being higher than those for any others. Removing these two statements yielded the adjusted means given in parentheses. Taking these adjusted means as more accurate measures of central tendency, the stages were ordered as predicted, except for a reversal involving Stages 5A and 5B.

Table 1
Means and Standard Deviations of
Weights for Items Keyed to Each Stage[1]

	Stage					
	2	3	4	5A	5B	6
Group 1						
Mean	-.34	-.17	-.13	.34 (.40)	.35	.31 (.42)
S.D.	.10	.19	.15	.25	.09	.27
Group 2						
Mean	-.33	-.26	-.03	.29	.40	.31
S.D.	.12	.16	.21	.22	.05	.23

[1]Schönemann's algorithm was used to derive a weight estimate or scale value for each item in Rest's test. For the items keyed to a given stage, the several weight estimates were averaged together to obtain the means shown above. In short, these means are average weight estimates computed across items keyed to a given stage.

In the larger Group 2 (consisting of 1,080 subjects) the weights for Stages 2, 3, 4, 5A, 5B, and 6 were −.33, −.26, −.03, .29, .40, and .31, respectively; but again, Stage 6 had one item whose deviance was reflected in the large standard deviation of weights for that stage. As in Group 1, it was larger than for any other stage; as in Group 1, the deviant Stage 6 item was Item 29. Removing the deviant item yielded the Stage 6 mean weight, shown in parentheses. Taking the adjusted mean as a more accurate measure of central tendency, the stages were again ordered as predicted.

The mean weights in Table 1 provide clear support for the hierarchical ordering of Kohlberg's Stages 2, 3, and 4 and somewhat more ambiguous support for the relative ordering of Stages 5 and 6, which were unambiguously higher than 2 through 4. The ordering of the weights (or scale values) suggests that the reasoning of adjacent stages is more similar than that of nonadjacent stages. Consequently, there is an ordering of the issues which corresponds to the theoretically predicted stage sequence. Because the empirical weights mirrored Kohlberg's hierarchically ordered moral development sequence, there are grounds for believing that the sum based on those weights reflects a subject's level along Kohlberg's moral development sequence. And the weights provided clear evidence for the sequentiality of Kohlberg's stages, an hypothesis for which Kurtines and Greif (1974) have asserted that evidence has been lacking.

Verifying the stage ordering in Kohlberg's theory requires more than just evidence that empirically derived weights fall in the hypothesized ordering. The hypothesized stage ordering requires a logical analysis of reasoning at different stages to determine whether the reasoning of higher stages is more comprehensive, incorporating the reasoning of lower stages. The hypothesis requires longitudinal and cross-sectional data on age trends in moral reasoning, such as that of Holstein (1976), Kuhn (1976), and Rest (1975). To the longitudinal and cross-sectional data on stage ordering, the present study adds evidence that the reasoning of adjacent stages is more similar than the reasoning of nonadjacent stages. Consequently, the data suggest that there is an ordering to the stages

roughly corresponding to the predicted ordering.

Reliability

Table 2 shows reliability data for the four scoring schemes and for Rest's six-stage scores. Since a number of researchers (Dortzbach, 1975; Erickson, Colby, Libbey, & Lohman, 1976; Guttenberg, 1975; Morrison, Tawes, & Rest, 1973; Sanders, 1976; Troth, 1974) have used a three-story version of the DIT to save subject time, Table 2 contains reliability data on the shorter version as well. To permit computation of standard errors of measurement (e.g., Rest, 1975), Table 3 reports standard deviations for test and retest scores. The discussion, however, will focus on the indices of overall development derived from the full six-story version of the DIT.

All four indices of overall development exhibit fair to good reliabilities in these groups. The two scores most commonly employed by users of the DIT—*P* and the empirically weighted sum—generally had reliabilities in the upper .70's and .80's. Two trends in this table are

Table 2
Internal Consistency and Test-Retest Reliabilities[1]

	Internal Consistency (alpha)		Test-Retest		
Group	1	3	4	4a	4b
Six Stories					
P	.77	.82	.76	.81	.71
SS	.70	.75	.72	.52	.73
TS	.90	.67	.77	.90	.67
ES	.79	.87	.76	.92	.67
Stage 2	.50	.44	.62	.78	.27
Stage 3	.51	.55	.66	.66	.67
Stage 4	.52	.61	.76	.66	.80
Stage 5A	.60	.65	.66	.57	.68
Stage 5B	.28	.60	.51	.49	.56
Stage 6	.43	.72	.54	.57	.49
Group	1	3	4	4a	4b
Three Stories					
P	.76	.77	.65	.58	.67
ES	.71	.83	.71	.81	.63
Stage 2	.30	.32	.69	.70	.42
Stage 3	.32	.48	.52	.54	.50
Stage 4	.27	.56	.66	.47	.74
Stage 5A	.53	.69	.63	.60	.64
Stage 5B	--	.58	.41	.45	.39
Stage 6	.00	.50	.47	.26	.52

[1]The test items in the DIT are factorially complex, and consequently not parallel to each other. In such cases, alpha is a lower bound to, not an estimate of, reliability defined as the ratio of true variance to total variance.

Table 3
Standard Deviations in Test and Retest Groups for Three and Six Story DIT Data

	Group 3		Group 4	
Six Stories	s_{pre}	s_{post}	s_{pre}	s_{post}
P	18.53	19.90	13.40	12.70
ES	10.76	10.90	5.86	5.53
Stage 2	2.91	2.78	3.19	3.49
Stage 3	6.12	6.31	5.29	5.35
Stage 4	7.13	6.97	7.21	7.73
Stage 5A	7.48	7.40	6.18	5.75
Stage 5B	3.65	3.67	3.27	3.23
Stage 6	3.90	4.33	2.60	2.79

	Group 3		Group 4	
Three Stories	s_{pre}	s_{post}	s_{pre}	s_{post}
P	20.43	22.83	15.86	14.63
ES	6.20	6.16	3.91	3.40
Stage 2	3.56	3.51	4.55	4.19
Stage 3	7.83	8.58	6.77	6.12
Stage 4	8.72	8.06	7.92	9.25
Stage 5A	9.51	10.43	8.03	7.78
Stage 5B	3.18	3.24	3.00	2.69
Stage 6	3.53	3.82	3.11	3.36

worth noting. First, the empirically weighted sum is superior to the simple sum in all but one case. It was significantly more reliable in Groups 3 and 4a ($p < .05$). In the one case in which the reliability of the simple sum was higher than that of the empirically weighted sum, it was not significantly higher ($p > .05$). Second, the reliability of *P* and the empirically weighted sum were generally about .80 for the six-story DIT. For the shorter, three-story version, the reliabilities were about .70.

Table 4 shows the test and retest means for the test-retest groups. The means tended to increase with retesting, although the increases were all small and only occassionally reached significance. Whereas the means for Rest's objective measure showed a slight tendency to increase on retesting, scores on Kohlberg's interview measure of moral development tended to drop with retesting (Rest, 1974).

Actually, the evidence for retesting effects in Table 4 is not appreciable. Only the theoretically weighted sum changed significantly in Groups 4, 4A, or 4B. No one has recommended use of this index; it is included here for comparative purposes and because it closely parallels Kohlberg's Moral Maturity Score. All measures changed in Group 3, but these are pre-post data from several moral education studies. While none of these studies reported significant change from pre- to post-test, when the data from the several studies were combined, a significant *t*-statistic resulted. Because the data come from moral education studies, the Group 3 test-retest reliabilities in Table 2 provide a measure of the stability for the relative standing of individuals;

Table 4
Mean Scores for Test and Retest Groups

Group 3	P	SS	TS	ES
Test	40.78*	62.14**	49.76*	22.00*
Retest	44.93	66.36	49.12	23.05
Group 4				
Test	39.58	66.29	47.44	22.94
Retest	39.58	67.69	46.19	23.65
Group 4a				
Test	36.48	59.74	46.62	20.26
Retest	35.78	62.16	46.60	22.05
Group 4b				
Test	41.37	70.06	47.91*	24.48
Retest	41.77	70.88	45.95	24.58

*$p < .05$

**$p < .01$

but the Group 3 mean difference in Table 4 is not simply a measure of retesting effects. Only Group 4, 4a, and 4b mean differences reflect the effect of retesting alone. Neither *P*, the empirically weighted sum, nor the simple sum changed significantly in these groups.

Validity

Table 5 shows the correlations of scores on the DIT with scores on several measures of constructs thought to be related to moral development. Group 1 was used to estimate the correlations between the DIT and measures of Comprehension of Moral Issues, Law and Order Orientation, and Political Tolerance. Rest et al. (1974) have described these measures and have explained their relationship to moral development. The correlations between scores on the DIT and scores on the Differential Aptitude Test (*DAT Manual,* 1966) were computed on the 40 ninth graders in Group 1, the only subjects for whom DAT data were available. Group 5 data were used to estimate the correlations between the DIT and Kohlberg's interview scores.

The theoretically weighted sum did not correlate significantly with the Comprehension of Moral Issues measure. Using a .05 level of significance, both *P* and the empirically weighted sum correlated significantly better than did the theoretically weighted sum with measures of Comprehension of Moral Issues, Law and Order Orientation, and Political Tolerance. But most importantly, the theoretically weighted sum was not significantly correlated with Kohlberg's interview measure of moral development in the total Group 5 or in either subgroup. In Group 5 as a whole, *P*, the simple sum, and the empirically weighted sum all had significantly ($p < .05$) higher correlations with Kohlberg interview scores than did the theoretically weighted sum. Table 5 points to a serious weakness in the theoretically weighted sum; it does not correlate as highly as would be expected with measures which should be related to level of moral development.

While three of the moral development indices were highly and significantly correlated with Kohlberg's measure in Group 5 as a whole, all had modest to low correlations with Kohlberg's

Table 5
Correlations of Four Moral Development Indices with Measures of Cognitive Ability (DATVN), Comprehension of Moral Issues (COMP), Law and Order Orientation (LO), Political Tolerance (PT), and Kohlberg's Interview Scores

Measure	P	SS	TS	ES
DATVN	.43**	.23	.12	.47**
COMP	.65**	.54**	.12	.63**
LO	-.59**	-.50**	-.14*	-.49**
PT	.62**	.50**	.16*	.55**
Kohlberg Interview				
Group 5	.68**	.63**	.07	.70**
Group 5a	.17	-.12	.22*	.20*
Group 5b	.35**	.32**	.14	.37**

measure in the junior high school sample, Group 5a, and the college subgroup, Group 5b. Prior research (Rest, 1975; Holstein, 1976; Kuhn, 1976) has indicated that older subjects score higher than do younger subjects on both Rest's and Kohlberg's tests. The fact that the two measures correlated only modestly in age homogeneous groups, however, suggests that the majority of their common variance in the total group can be accounted for by the measures' common age trends. The modest correlations in age homogeneous groups further indicate that Rest's and Kohlberg's test cannot be considered measures of the same construct which differ only in format.

As Rest (1974) has stated, his test employs a recognition task, whereas Kohlberg's employs a production task. Rest's test presents subjects with different types of moral issues, each characteristic of reasoning at one of Kohlberg's stages, and it asks subjects to indicate the importance of each issue. Variance in DIT scores represents variation in the kinds of issues considered important by the several subjects. Kohlberg's interview, on the other hand, encourages the subject to discuss moral dilemmas. Scores on his measure reflect the stage of reasoning spontaneously produced by the subject in his/her discussion.

Longitudinal Studies

Table 6 shows the Longitudinal trends observed in Samples 6 and 7. Over the four years in which Group 6 subjects were studied, there was a significant decrease in their Stage 2 and 3 scores and a significant increase in their Stage 5A and 5B thinking. All four indices—*P*, the simple sum, the theoretically weighted sum, and the empirically weighted sum—increased significantly over the period studied. As indicated by the size of the *F* statistics, the strongest longitudinal trend occurred for the empirically weighted sum followed by *P*, the simple sum, and the theoretically weighted sum.

A different pattern was observed in Sample 7. Of the stage scores, only Stage 3 scores changed significantly; and Stage 3 is not reflected by Rest's P. Not surprisingly, *P* displayed no significant upward shift. Applying the conventional .05 level of significance, only the empirically weighted sum changed significantly, although the significance level for the theoretically weighted sum almost reached the .05 level (p = .06). As indicated by the t statistics, the empirically weighted sum again showed the strongest longitudinal trend.

Loevinger (1976) expressed concern that Rest uses *P* as an index of overall development when *P* reflects only the upper stages. The longi-

Table 6
Means of Moral Development Measures by Year in Two Longitudinal Studies

	Group 6			
	1972	1974	1976	F
P	32.87	39.78	44.15	20.06**
SS	59.85	62.72	67.31	7.70**
TS	45.90	48.87	48.05	4.18*
ES	20.26	23.26	24.06	24.86**
Stage 2	6.30	5.00	4.62	6.60**
Stage 3	11.30	8.04	7.56	12.50**
Stage 4	18.15	18.61	17.16	1.30
Stage 5A	12.44	16.48	17.82	18.20**
Stage 5B	4.26	4.83	5.46	3.50*
Stage 6	2.85	2.72	3.20	.60

	Group 7		
	1974	1976	t
P	33.88	36.92	- .97
SS	58.62	58.05	.22
TS	50.11	52.86	-2.00
ES	20.46	22.96	-2.64*
Stage 2	5.00	3.59	1.60
Stage 3	12.38	9.17	2.90**
Stage 4	18.10	19.98	-1.12
Stage 5A	14.29	14.42	- .08
Stage 5B	3.76	3.88	- .15
Stage 6	2.29	3.86	-2.25

*$p < .05$

**$p < .01$

tudinal data in this study serve to amplify her concern. In Group 6 the longitudinal trend in *P* was weaker than that for the empirically weighted sum because *P* does not reflect the observed changes in Stages 3 and 4. The results are more dramatic in Group 7, in which the only significant changes were in lower stage scores not reflected by *P*. As an index of overall development, *P* appears insensitive to change, particularly when that change occurs in lower stages. Such insensitivity should be of particular concern to researchers using the DIT to evaluate outcomes of clinical and educational programs.

The simple sum was even more inferior than *P* to the empirically weighted sum in its sensitivity to longitudinal trends. In both Groups 6 and 7, the longitudinal trend as reflected in the simple sum was weaker than the trend as reflected in any of the other three indices. On two counts, reliability and sensitivity to longitudinal trends, the simple sum has proven inferior to the empirically weighted sum and *P*.

Discussion

One conclusion suggested by this research is that the empirically weighted sum provides the most desirable measure of overall development. It takes into account information from all stages; and as a result, it is more sensitive to longitudinal change than is *P*, particularly when the change occurs primarily in lower stages. The empirically weighted sum yielded generally higher reliabilities and greater sensitivity to longitudinal change than did the simple sum. As compared to the theoretically weighted sum, the empirically weighted sum yielded a significantly higher correlation with Kohlberg's measure and stronger longitudinal trends.

Although the empirically weighted sum is tedious to score by hand, Davison, Robbins, and Swanson (1977) developed a computer program which uses DIT responses to compute stage scores, *P*, and the empirically weighted sum. The program, which is available from the authors, makes tedious hand scoring unnecessary.

If the reliability data from these samples were to be summarized, then it would have to be said that the overall indices of reliability (*P* and the empirically weighted sum) based on six stories generally have internal consistency and test-retest reliabilities in the high .70's and low .80's. For the three-story DIT, reliabilities of the overall indices generally seemed to fall by about .10 points. Stage score reliabilities, either test-retest or internal consistency, seldom surpassed .70 and were mostly in the .50's and .60's.[1]

As for concurrent validity, the overall indices seemed to correlate in the .40's with a measure of general aptitude, in the .60's with a measure of comprehending moral issues, in the high .40's or .50's with a measure of law-and-order orientation, and in the .50's or .60's with a measure of political tolerance. Correlations with Kohlberg's measure were about .70 in an age heterogeneous group but only about .20 in a group of ninth graders and about .35 in a group of college students.

The reliability data in Table 2 indicate that using the shorter three-story DIT should have little effect in studies where the group means are the focus. In studies of group means, only precise estimates of means are necessary, that is, estimates which have low standard errors. Regardless of a measure's reliability, mean estimates can be made as precise as desired by employing a sufficiently large sample size. In correlational studies, the drop in reliability associated with the shorter version cannot be overcome by an increase in sample size. According to classical reliability theory, the net effect should be a drop in the correlation between the DIT and any outside variable. The drop should be proportional to the square root of the three-story DIT reliability divided by the square root of the six-story reliability. For example, assume that the six-story version correlates about .40 with some variable and has a reliability of .80 in the population of interest, whereas the shorter version has a reliability of .70. The shorter version would be expected to have a correlation of $(\sqrt{.70}/\sqrt{.80}) \times (.40) = .37$. Using the shorter version would always reduce the observed correlation somewhat and may even reduce it to nonsignificance, but the reduction should usually be small.

The data presented in this paper reinforce Rest's (1974) assertion that his test and Kohlberg's interview are not equivalent measures of the same construct which differ only in format. Rest's test is an objective instrument employing a recognition task, whereas Kohlberg's measure employs subjective scoring and a spontaneous production task. The two measures were highly correlated only in an age heterogeneous sample in which the high correlation can be attributed to the common age trend exhibited by scores on both measures.

[1]Except for the results in Groups 4a, 4b, 5a, 5b, and 7, the data in this paper bear on the reliability and validity of the DIT in groups which are heterogeneous in age and education. Further research is needed to determine the test's reliability in groups which are homogeneous in age and education.

References

Cooper, D. *The analyses of an objective measure of moral judgment.* Unpublished doctoral dissertation, University of Minnesota, 1972.

DAT Manual (4th ed.). New York: The Psychological Corporation, 1966.

Davison, M. L. On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika,* 1977, *42,* 523–548.

Davison, M. L., Robbins, S., & Swanson, D. B. *DIT: A FORTRAN IV program for scoring the Defining Issues Test.* Unpublished manuscript, University of Minnesota, 1977.

Davison, M. L., Robbins, S., & Swanson, D. Stage structure in objective moral judgments. *Developmental Psychology,* in press.

Dortzbach, J. R. Moral judgment and perceived locus of control: A cross-sectional developmental study of adults, ages 25–74 (Doctoral dissertation, University of Oregon, 1975). *Dissertation Abstracts International,* 1975, *36,* 4662B.

Erickson, V. L., Colby, S., Libbey, P., & Lohman, G. The young adolescent: A curriculum to promote psychological growth. In G. D. Miller (Ed.), *Developmental education.* St. Paul, MN: Minnesota Department of Education, 1976.

Froming, W. J., & McColgan, E. *A comparison of two measures of moral judgment.* Paper presented at the meeting of the American Psychological Association, San Francisco, August, 1977.

Guttenberg, R. *Videotaped moral dilemmas: Altering the presentation of the stimuli in the Defining Issues Test.* Unpublished manuscript, Brown University, 1975.

Holstein, C. B. Irreversible stepwise sequence in the development of moral judgment: A longitudinal study of males and females. *Child Development,* 1976, *47,* 51–61.

Kohlberg, L. Stage and sequence: The cognitive-developmental approach to socialization. In G. A. Goslin (Ed.), *Handbook of socialization theory and research.* San Francisco, CA: Rand McNally, 1969, 347–380.

Kohlberg, L. From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive Development and Epistemology.* New York: Academic Press, 1971.

Kuhn, D. Short-term longitudinal evidence for the sequentiality of Kohlberg's early stages of moral development. *Developmental Psychology,* 1976, *12,* 162–166.

Kurtines, W., & Greif, E. B. The development of moral thought: Review and evaluation of Kohlberg's approach. *Psychological Bulletin,* 1974, *81,* 453–470.

Loevinger, J. *Ego development.* San Francisco, CA: Jossey-Bass, 1976.

Loevinger, J., & Wessler, R. *Measuring ego development: Construction and use of a sentence completion test* (Vol. 1). San Francisco, CA: Jossey-Bass, 1970.

Loevinger, J., Wessler, R., & Redmore, C. *Measuring ego development: Construction and use of a sentence completion test* (Vol. 2). San Francisco, CA: Jossey-Bass, 1970.

McColgan, E. B. Social cognition in delinquents, predelinquents, and non-delinquents. (Doctoral dissertation, University of Minnesota, 1975). *Dissertation Abstracts International,* 1975, *37,* 199A.

McGeorge, C. The susceptibility to faking of the Defining Issues Test of moral development. *Developmental Psychology,* 1975, *11,* 108.

Morrison, T., Tawes, O., & Rest, J. *An evaluation of a jurisprudential model for teaching social studies to junior high school students.* Study in progress, University of Manitoba, 1973.

Rest, J. *Manual for the Defining Issues Test.* Unpublished manuscript, University of Minnesota, 1974.

Rest, J. R. Longitudinal study of the Defining Issues Test. *Developmental Psychology,* 1975, *11,* 738–748.

Rest, J. *Moral judgment related to sample characteristics.* Unpublished manuscript, University of Minnesota, 1976. (a)

Rest, J. R. New approaches in the assessment of moral judgment. In T. Lackonna (Ed.), *Moral development and behavior: Theory, research, and social issues.* New York: Holt, Rinehart, & Winston, 1976. (b)

Rest, J. R., Cooper, D., Coder, R., Masanz, J., & Anderson, D. Judging the important issues in moral dilemmas—an objective test of development. *Developmental Psychology,* 1974, *10,* 491–501.

Sanders, N. *Pretesting in the pilot study schools of the skills for the ethical action project.* Unpublished manuscript, Research for Better Schools, Inc., 1976.

Schönemann, P. H. On metric multidimensional unfolding. *Psychometrika,* 1970, *35,* 349–366.

Troth, A. G. *An assessment of impacts on students of the "Values" courses.* Unpublished manuscript, St. Olaf College, 1974.

Acknowledgements

This research was supported by a grant from the U.S. Public Health Service (Grant No. 1-R01-

MH27861-01) to the University of Minnesota. The authors thank William J. Froming, Edgar B. McColgan, Collin McGeorge, James R. Rest, and Elaine Wilson for the use of their data.

Author's Address

Mark L. Davison, Department of Social, Psychological, and Philosophical Foundations of Education, 330 Burton Hall, 178 Pillsbury Drive Southeast, Minneapolis, MN 55455