Journal of Educational Psychology
1997, Vol. 89, No. 3, 498–507

Alchemy and Beyond: Indexing the Defining Issues Test

James Rest, Stephen J. Thoma, Darcia Narvaez, and Muriel J. Bebeau
University of Minnesota, Center for the Study of Ethical Development

For over 20 years, the Defining Issues Test (DIT) has used the P index. In view of criticisms, a search has been underway for a new index. The authors propose a working definition of construct validity, systematically reanalyze existing data sets ("classic" studies) with new indexes, and make comparisons to trends obtained using the P index. The criteria for construct validity are (a) sensitivity to educational interventions, (b) differentiation of age–educational groups, (c) upward movement in longitudinal studies, (d) correlations with moral comprehension, (e) correlations with prosocial behavior, and (f) correlations with civil libertarian attitudes. As meta-analysis demonstrates, a new index, N2, generally outperforms the P index.

In the early 1970s, Larry Kohlberg found amusement by comparing the research project on the Defining Issues Test (DIT) to alchemy. The alchemist's dream of the middle ages had been to transmute the "base metals" into gold. At the time, Kohlberg was beginning work on revising his scoring system and was mindful of the complexities involved in analyzing moral judgments and the arduous work required of a scorer. Kohlberg pointed out similarities between alchemy and the attempt to derive a measure of moral judgment from a multiple-choice test. Obtaining moral development scores by simply asking participants to rate or rank statements seemed too good to be true—it was like trying to turn lead into gold. Nevertheless he was supportive of the exploration of new sources of information on moral judgment (Kohlberg, 1979), although he did make sure that we were alert to the possible problems with such an approach. Years later, once we had developed computer programs to score the DIT, we joked with Kohlberg about doing morality research "untouched by human hands"—the very thought of which he also found preposterously amusing.

For over 20 years, DIT researchers have relied on the P score to index moral judgment and for that long we have tried to find a better index. Not only did we aspire to find a way to use both ratings and rankings to measure moral judgment (the DIT uses only ranking data), we also had a second aspiration: to find a new index that would use existing DIT rating and ranking data—involving no additional work of the participants—and would produce more powerful data trends than we were then getting. In short, we were searching for a new index that would produce better results without costing anything more. Again, the similarities of that endeavor with the alchemists has been painfully obvious.[1] The practical point of this paper is to show that we have found a way to use both rating and ranking data from the DIT, which is generally (but not invariantly) better, called N2.

An *index* is the overall score by which a participant is characterized. In other words, an index is the number used to represent a participant's development. The most used index for the DIT for over 20 years has been the P index, which is based on a participant's ranking of prototypic items written for Kohlbergian Stages 5 and 6. The P index is interpreted as the relative importance participants give to principled moral considerations (Stages 5 and 6) in making a moral decision. The features of the DIT have been discussed extensively in a previous article (Rest, Thoma, & Edwards, 1997) along with the validation strategy for the DIT. The present article is a sequel to the previous article.

The P index was adopted in the 1970s after considering many other alternative indexes. The P index survived because it consistently gave better trends for the theoretically expected findings than did other ways of indexing. Secondarily, the P index also has the virtue of relatively easy computation and of straightforward interpretation.

Two criticisms of the P index have predominated in discussions of the DIT. The first issue might be called the *qualitative–quantitative issue,* and the second, the *problem with throwing away data.*

Qualitative–Quantitative Issue

Kohlberg's developmental theory emphasizes qualitatively different stages in understanding the social–moral

James Rest, Department of Educational Psychology, University of Minnesota; Stephen J. Thoma, Child Development Center, University of Alabama; Darcia Narvaez, Department of Curriculum & Instruction and Department of Educational Psychology, University of Minnesota; Muriel J. Bebeau, Department of Health Ecology, University of Minnesota. All authors are members of the Center for the Study of Ethical Development, College of Education and Human Development, University of Minnesota.

This article is part of a larger on-going project to devise better methods of measuring development in moral judgment, sponsored by the Center for the Study of Ethical Development. We thank the following researchers for use of their data: DeWitt Baldwin, Mark Davison, Deborah Deemer, Laura Duckett, Jean Evens, William Penn, Donnie Self, and Muriel Ryden. Special thanks to Lynn Friedman for her expertise and advice in matters meta-analytic.

Correspondence concerning this article should be addressed to Stephen J. Thoma, P.O. Box 870158, University of Alabama, Tuscaloosa, Alabama 35487-0158. Electronic mail may be sent via Internet to sthoma@ua1vm.edu.

[1] An anonymous reviewer wryly suggested a different medieval notion: "Searching for an index superior to P strikes me as something akin to searching for the methodological Holy Grail."

world. The stages are described in terms of different lines of moral argument, qualitatively different moral logics for organizing various concepts. With this emphasis on qualitative differences, the P score—using a continuous scale ranging from 0 to 95—seems out of place in substituting a quantitative dimension for qualitative distinctions. Kohlberg represents development in terms of stages, that is, assessing participants' development is done in terms of their being in one stage or another (e.g., a Stage 2 participant or a Stage 4 participant). In contrast, the DIT P score assigns a number (e.g., 37, 54) to a participant. How can a quantitative measure represent a Kohlbergian view?

To address this question, we distinguish two issues: (a) the issue of depicting qualitative differences in social–moral understanding and (b) the issue of scoring participants. Regarding the first issue, stages depict how the various perceptions and considerations are assimilated into a coherent logical whole. Stages depict types of reasoning. However, regarding the second issue, we realize that one person is not confined to one stage. People are aware of various stages of reasoning; people are not consistently "pure types." The types of thinking that are used by a particular person are influenced by many factors, including various situational variables (e.g., the kind of dilemma, the participant's familiarity with the type of dilemma, group ideology, what others are saying about the dilemma; Rest, 1979). Therefore we think of the question of assessment not as what stage is a person in (assuming stage consistency and limitation to one stage) but rather as to what extent and under what conditions does a person manifest particular stages of thinking. According to this understanding of the stage model and assessment, qualitative differences are emphasized in discussing types of thinking (stages), but quantitative differences are emphasized in discussing assessment. In fact, theorists who are often regarded as emphasizing qualitative features (e.g., Kohlberg, Loevinger [1976]) do use quantitative information in assessment (e.g., for Kohlberg, quantitative information determines major and minor stage assignment [Colby & Kohlberg, 1987, pp. 185]; Loevinger used the empirically derived "ogive" rules for assessing ego level [Loevinger, 1976, pp. 236]). Although Kohlberg and Loevinger assess development by assigning participants to a stage, "development" for us means that people over time come to use higher stages more and lower stages less (not that people move out of one stage and into another). Therefore, development is a matter of shifting distributions of stages rather than the move from one stage completely into the next. Movement is gradual—matters of more or less—and assessment needs to consider quantitative dimensions of stage use. Note therefore that DIT researchers acknowledge that the stages are qualitatively different ways of moral thinking, and make qualitative distinctions in designating stage differences to different items; however, we use quantitative differences (the degree to which a type of thinking is manifest) in depicting the developmental scores of participants. (See Rest, 1979, pp. 48–74, for further discussion.)

Throwing Data Away

The second problem that is associated with the P score is more serious; namely, in computing a P score, none of the information about the lower stage items is used, nor is any of the rating data used (in contrast to the ranking data). A reasonable question to ask is whether or not a better index would come from using the full range of data.

In the late 1970s, Mark Davison (Davison, 1977; Davison & Robbins, 1978) first reported results of scaling DIT items. An index based on scaling items offered the prospect of using all the items (low-stage items as well as high-stage items). A scaling approach also offered a way to make adjustments to items according to how well they were working empirically. (For instance, a high-stage item that was working well would have a high item weight; an item that was intended to be a high stage item that was not working so well would have a lower weight that would not affect a participant's score as much as the good item.) Davison's research eventuated in the D (for developmental) index. Since the 1970s, we have calculated a D index along with the P index, but over the years we have found that P generally outperforms the D index.

Jean Evens (1995) undertook to apply Davison's (1979) scaling methods to a much expanded database unavailable to Davison in the 1970s. Drawing from a pool of over 58,000 DITs, Evens constructed several distributions on which to derive scaling values (e.g., a normal distribution of P scores, a "rectangular" distribution in which every range of P scores is equally represented, a distribution of equal numbers of participants having exceptionally high scores on each of the stages, etc.). Because the effect of different distributions on scaling techniques is not known, Evens explored several sets of scaling weights from several distributions, so as to determine the effect on scaling values of different distributions. After deriving scale weights for the items from these different distributions, she then constructed different indexes based on the different sets of scale values. She put each index in competition with each other by examining the statistical trends generated by each of them in a number of studies. Evens computed dozens of new scaling indexes and systematically compared the statistical results from each new index with the P score. We were astounded at the results. Every new index that she devised did much worse than the P index. We were baffled in finding that the application of sophisticated scaling techniques to an enriched database did not move us forward—it had moved us backward!

Since Evens' (1995) study, we have organized an ongoing project to explore a large number of ways of indexing the DIT. First, we have a working definition of construct validity because the DIT has five criteria. Namely, we assume that an index of moral judgment should do the following: (a) differentiate naturally occurring groups that vary in terms of presumed expertise in moral judgment (e.g., junior high school students from Ph.D. students in moral philosophy); (b) correlate with moral comprehension; (c) show longitudinal change as a function of age and of enriching experiences; (d) be sensitive to moral education interventions (i.e., show pre- to posttest differences from moral education programs); (e) link moral judgment to behavior; and (f) link moral judgment to civil libertarian attitudes. In addition, we look at internal reliability. (See Rest, Thoma, & Edwards, 1997, for a more extensive discussion of construct validity and for details of our proposal for a working definition for the DIT.)

Second, we assembled several specific data sets from studies referred to as "classic" studies. By classic studies, we mean that the studies have been cited in the literature as major studies for building the construct validity of the DIT, that they represent diverse ways of studying moral judgment, that they represent many thousands of participants studied from various parts of the country, and that they contain some old and some new data.

Third, our strategy then is to see if we can devise a new index that performs better than the summary statistics in the classic studies that have used the P index. We assume that an index that outperforms the P index on these studies is a better index. In other words, our strategy is to see if we can beat the benchmark statistical trends established by the P index in the classic studies.

When one considers the many ways that a score can be put together, one begins to realize the vast possibilities. For instance, do we use rating data or ranking data or both? Do we weight the ranking data by 4, 3, 2, 1, or 7, 6, 4, 1, or 1, 1, 1, 1, and so on. Do we use empirically derived scaling weights for the rates of items? Which set of scaling weights do we use—from scaling, from factor analysis? How do we handle missing data? If we combine various elements going into an index, how do we combine them? Do we use Kohlberg's classifications for grouping items or some other groupings?

Beating the benchmark statistics of the P index in the classic studies has turned out to be difficult. Almost always, the new indexes have performed more poorly than the P index. In late 1995, we did come up with an index based on scaling analyses that seemed to perform slightly better than the P index. Finding something at last that was not clearly inferior to the P index was heartening, because heretofore all the new ideas had produced results that were decidedly much worse than the P index. We called this index (somewhat prematurely) *NewIndex,* and urged others to check it out in their studies alongside results from the P index. In this article, we are supplanting NewIndex with something else, the N2 index, which is better than anything else we have derived.

We have tried dozens of new indexes, some suggested by the indexes devised by other researchers. One of these indexes by other researchers proved to be especially valuable to the development of N2. Credit is due to Georg Lind of Germany for identifying many years ago that development is manifest in the greater differentiation of the ratings of high-stage items from the ratings of low stages. Back in the 1970s (Lind, 1979, p. 57), Lind had directed attention to the difference between ratings for Kohlbergian high stages from ratings written to exemplify Kohlbergian low stages. He showed that this gap increases from high school students to college students. In the ensuing years, Lind has gone on to devise his C index (Lind, Hartmann, & Wakenhut, 1985; Lind, 1995), which captures this phenomenon (the increasing gap between ratings of low stages from ratings of high stages) using sophisticated individual multivariate analysis of variance (MANOVA) procedures. As discussed in Rest et al. (1997) we did not find that the specific formulation of the C index by Lind was an improvement over the P index on DIT data.

Reexaming the Bebeau and Thoma study (1994; a moral intervention study in a professional dentistry school using the DIT as a pre- and posttest), we rediscovered the effect that Lind had mentioned in the 1970s. We found that there were two effects of the educational intervention: one effect was the acquisition of new thinking (increases in P score—the familiar effect); the second effect was systematic rejection of simplistic thinking (decreases in Stages 2 and 3). From a practical educational point of view, both kinds of developmental progress are desirable: gaining more sophisticated moral thinking and also becoming clearer about what ideas should be totally rejected for their simplistic and biased solutions. Thus the two components of N2 were suggested by the Bebeau and Thoma data: a measure of prioritizing the high stages and a measure of discrimination and rejection of the lower stages.

In this article, we present data supporting the claim that the N2 index generally outperforms the P index. To do this, we present comparisons of the P index with the N2 index on each of the validity criteria, plus reliability. We also comment on the theoretical implications of the indexing work for the construct, moral judgment.

Method

Because this paper presents data from many previously published studies, details about the participants and methods will not be repeated here. In this section, we comment on calculating the indexes, P and N2.

Briefly, the regular DIT consists of six dilemmas (including the famous dilemma from Kohlberg's procedure of Heinz and the drug).[2] Each dilemma is followed by 12 items, representing various issues that might be considered in making a decision about what to do in the dilemma. The participant's task is to rate each item in terms of how important it is, and then to rank the most important items (the top four ranks). Various devices and checks for participant reliability are built into the DIT (see Rest, Thoma, & Edwards, 1997).

A P score is calculated on the basis of ranking data. If a participant ranks a principled item as "most important," then this increases the P score by 4 points (in second place, by 3 points; in third place, by 2 points; in fourth place, by 1 point). The P score is the total number of points across the six dilemmas. The P score is converted from a base of 60 points to a percentage (with a base of 100). P scores can range from 0 to 95 (not 100 because every dilemma does not have four possible P items). Missing data (i.e., leaving some ranks blank) is dealt with by adjusting the P score on the basis of responses given (for instance, if a participant leaves out the third rank on one story, the P score is recalculated on the basis of 58 points instead of the full 60 points).

An N2 score has two parts: the degree to which P items are prioritized (almost identical to the P score) plus the degree to which

[2] The short form consists of three dilemmas and takes about 30–40 min to complete, whereas the regular six-story form takes about 45–55 min to complete. Comparisons of the six-story form with the three-story form will be provided in footnotes.

the lower stages are rated lower than the ratings of the higher stages. First, the prioritization of P items follows the procedure above for the P score except in the handling of missing data. If a participant leaves out a rank, then in N2 no adjustment is made for that omission—omitting a rank, in effect, is the same as not prioritizing a P item. Leaving out all ranks for one whole dilemma is adjusted by basing a total score on the other five dilemmas. If more than one dilemma is omitted, the whole protocol is invalidated for we assume there is a problem in test motivation in general, not an occasional ambiguity.

The second part of N2 is based on rating data, not ranking data. The main idea is that "discrimination" is measured in terms of the average rating given to items at Stages 2 and 3 (the lower stages) subtracted from the average rating given to items at Stages 5 and 6. Hence the distance of Stages 2 + 3 from Stages 5 + 6 is the measure of discrimination. Average ratings are standardized by dividing this difference (Stages 5 + 6 − Stages 2 + 3) by the participant's standard deviation of Stages 2 + 3 + 5 + 6. Occasional missing rates are supplied by filling in the average rating for the story. If rates for one whole dilemma are left out, then the score is adjusted, on the basis of the other five dilemmas. However, if more than eight rates for two dilemmas are missing, then the whole protocol is invalidated.

The two parts of N2 are combined into one score per participant by adding the P score to the rating data weighted by three. (We weight the discrimination component by three because this component has about ⅓ the standard deviation of the P scores; therefore weighting equalizes the two parts of the N2 index.) N2 scores are adjusted to have the same mean and standard deviation as the P score on the 1995 standardization sample (n = 1,115) so that comparisons between P and N2 can be made easily.

Note that the N2 index uses the same ranking data from the same participants as the P score but also uses the rating data from the protocols. Because the N2 score uses both rating and ranking data, and because it has more stringent rules for handling missing data than the P index, more protocols are invalidated for missing data in the N2 index than for the P index. In every comparison of the P index with N2 that follows, we report on samples using the same participants for both indexes. Because we are eliminating more protocols for the N2 index than for the P index, the samples may sometimes be smaller than in previous reports (e.g., Rest, Thoma, & Edwards, 1997) and there are some slight discrepancies in the means, standard deviations, and so on, of the P index in this article with the statistics reported in previous studies because of the smaller sample sizes.

Results

Sensitivity to Educational Intervention

Because studies of educational interventions provided the first lead to the N2 index, we will consider first this criterion. As discussed in Rest, Thoma, and Edwards (1997), the expectation is that an index of moral judgment will be sensitive to educational programs designed to facilitate moral judgment development (i.e., matched t tests on the pre- and posttest will be statistically significant). Table 1 reports four studies of this sort. It shows the pre- to posttest changes in terms of matched t tests. For instance, for the Bebeau and Thoma (1994) study, the matched t test for the P index is 4.39 and the matched t test for the N2 index is 6.48. Given this result, we conclude that the N2 index is more sensitive to educational intervention on this study than the P index—and in this regard, N2 is outperforming P.

Table 1
Pre–Post Change on Educational Intervention Studies

Study	P index matched *t* test	N2 index matched *t* test
Bebeau and Thoma, 1994 (n = 114)	4.39	6.48
Duckett and Ryden, 1994 (n = 209)	7.64	9.71
Penn, 1990 (n = 48)	9.01	10.41
Self and Baldwin, 1994 (n = 131)	3.55	3.86
Combined sample (n = 502)	11.20	13.94

Note. All matched t tests are statistically significant at $p < .01$ level.

Looking at the four studies, we see that N2 produces a bigger matched t statistic than the P index in all four studies. (N2 beats the intervention benchmarks of the P index.) The combined sample simply includes the participants from all four studies (total n = 502). Again, as in the separate studies, the combined sample produces a higher matched t for the N2 index than for the P index (13.94 vs. 11.20), supporting the contention that N2 is a better index than P.[3]

Differentiation of Educational Groups

We expect Ph.D.s in moral philosophy to score higher than junior high school students on a measure of moral judgment. Usually, one of the first kinds of studies performed on any measure that purports to be developmental is its differentiation of age–educational groups. For the DIT, a sample of about 1,000 participants from junior high school, senior high school, college, and graduate school was compiled (the 1995 composite sample—see Rest, Thoma, & Edwards, 1997, for details). An additional sample of similar composition (about 1,000 participants also from four education levels) was compiled by Davison in 1979 (the 1979 composite sample).

Table 2 shows age–educational trends of the two samples in terms of one-way ANOVA. The dependent variable is either P score or N2 score, and the independent variable is level of education. N2 differentiates the educational groups better than P.[4]

Longitudinal Trends

One longitudinal study spans 10 years (Rest, 1986) and contrasts three testings for 49 participants. (Participants

[3] The short 3-story form still produces a statistically significant result for both P and N2 indexes (n = 502, $p < .001$), but the shorter form produces much lower t-test values for both indexes, with short-form N2 (matched t = 10.36) still outperforming short-form P (matched t = 7.99).

[4] The short 3-story form produces a statistically significant result for both P and N2 indexes. For the 1995 composite sample (n = 955) for the P index, F = 186.1; for N2, F = 210.2. For the 1979 composite sample (n = 1,012) for the P index, F = 158.8; for N2, F = 211.0.

Table 2
Analysis of Variance of Age–Educational Group Differentiation of P and N2 Indexes

		F	
Source	*df*	P index	N2 index
1995 composite sample ($n = 955$)			
Education	3	249.2	293.4
Residual	951	(175.5)	(163.5)
1979 composite sample ($n = 1{,}012$)			
Education	3	203.3	294.0
Residual	1008	(176.1)	(157.2)

Note. Education is the main effect, coded at four levels. Residual mean squares are given in parentheses. All *F* tests are statistically significant, $p < .01$.

were either male or female, college educated or non-college educated, from a large city or from a small town in the upper Midwest.) Table 3 compares the P and N2 indexes on gains over time in terms of repeated measures analysis of variance (ANOVA). Time 1 represents scores in high school, Time 2 is 2 years later, Time 3 is 10 years later. Note that both are highly significant, and the N2 index shows a stronger trend in the *F* statistic.

The McNeel (1994) data include 263 participants in a longitudinal study of college students, from freshman to senior status. On the P index, students gained from a mean of 35.8 ($SD = 11.59$) to 45.81 ($SD = 13.58$). On the N2 index, students gained from 39.23 ($SD = 11.74$) to 49.18 ($SD = 12.18$). The matched *t* test for the P index is 11.13 but for N2 is 12.86. Although both the P index and the N2 index show highly significant gains over the college years ($p < .001$), the N2 index shows stronger longitudinal trends.[5]

In the 10-year longitudinal study, Deemer (1986) developed "life experience codes" from extensive interviews with the participants about their experiences during the previous 10 years. The point of Deemer's research was to go beyond the mere passage of years as an indicator of presumed development and to characterize "richness" of life experience, using terms such as *continued stimulation, richness of social environment,* and similar clinical judgments about the lives of the participants during the previous 10 years. Table 4 shows the correlations for these codes with the two DIT indexes at the later longitudinal testing. As can be seen, both indexes are significantly correlated with the life experience codes, performing at about the same level.[6]

Table 3
Analysis of Variance From 10-Year Longitudinal Sample

		F	
Source	*df*	P index	N2 index
Testing (3)	2	23.07***	29.26***
Within cells	96	(101.75)	(93.56)

Note. $N = 49$ cases tested at 3 times, repeated measures, from Rest (1986). Mean square errors are given in parentheses. Mauchly sphericity test is nonsignificant in both analyses and therefore no correction for degrees of freedom is necessary.
***$p < .001$.

Table 4
Correlations of Defining Issues Test Indexes With Indicators of "Richness" of Life Experience

	Indicators of life experience			
Index	Continued intellectual stimulation[a] ($n = 91$)	Composite richness code[b] ($n = 96$)	Richness of social environment[c] ($n = 66$)	Educational completion[d] ($n = 93$)
P	.58***	.59***	.66***	.52***
N2	.56***	.58***	.64***	.54***

Note. Numbers are correlations of the P or N2 index with various indicators of life experience. Correlations of the P index are not significantly different from the correlations of the N2 index with the life experience variables.
[a]Continued intellectual stimulation is Deemer's (1986) code for ongoing cognitive stimulation and support of learning over the 10-year period. [b]Composite richness code is a composite variable constructed by Evens (1995) from five of Deemer's codes. It is the sum of those codes. [c]Richness of social environment is Deemer's code for the stimulation from the social environment (spouse, friends, institutional affiliations). [d]Educational completion reflects how much schooling has been completed (high school only, some college, college graduate, graduate–professional school).
***$p < .001$.

Correlations With Moral Comprehension

A test of moral judgment should be correlated with a measure of moral comprehension. As a measure of moral comprehension, we use a procedure that first presents a participant with a paragraph expressing a moral argument. Then the paragraph is followed by four shorter sentences; the participant's task is to choose which of the four sentences best expresses the gist of the paragraph. The Moral Comprehension Test has 11 paragraph-and-sentence units, and the score is simply the number of correct matches (0–11).

One data set comes from the Rest, Cooper, Coder, Masanz, and Anderson (1974) study. This is a sample of 140 participants (after elimination of incomplete data), composed of four educational groups: junior high, senior high, college, and graduate students. The correlation of moral comprehension with P is .67 and with N2 is .69 (both significant, $p < .001$). A replication comes from the Rest (1986) longitudinal sample, consisting of 96 participants, (ages 25 to 28)—a somewhat homogeneous sample therefore restricting the range of variables. The correlation of

[5] The short 3-story form on the Rest, 1986, longitudinal study produced an $F = 14.37$ for the P index and an $F = 16.91$ for the N2 index. On the McNeel (1994) longitudinal study, on the short form, the P index produced a *t* test of 9.90 ($n = 263$, $p < .001$) and the N2 index produced a *t* test of 10.37.

[6] The short 3-story form on the P index gives correlations with continued intellectual stimulation ($r = .57$), composite richness code (.58), richness of social environment (.67), and educational completion (.47). The comparable correlations for N2 are .55, .57, .66, and .50 (all correlations are significant, $p < .001$).

Table 5
Correlations of P and N2 With "Prosocial" Behavior

Index	Prosocial indicators Duncan Work Scale[a] ($n = 95$)	Community involvement[b] ($n = 75$)	Civic responsibility[c] ($n = 57$)
P	.38***	.32**	.39**
N2	.42***	.35**	.37**

[a]Duncan Work Scale is a sociological measure based on reported occupation of the social prestige of the occupation. [b]Community involvement is a Deemer (1986) experience code based on interviews that indicates extent to which the individual identifies with the community. [c]Civic responsibility is a Deemer experience code based on interviews that indicates amount of service activity in the community.
$^{**}p < .01$. $^{***}p < .001$.

moral comprehension with P is .34 and with N2 is .34 (statistically significant, $p = .001$).[7] Therefore, with regard to moral comprehension, both the P and N2 index are significantly correlated, but both are correlated at about the same level.

Links to Behavior

Table 5 presents correlations of P and N2 with various measures of prosocial behavior. Although both indexes are significantly linked to these measures, the two indexes are correlated at about the same levels.[8]

Links to Civil Libertarian Attitudes

Table 6 presents correlations of P and N2 with several measures of political attitudes regarding free speech, giving civic authorities excessive power, toleration of religion, and so on (see original references for more description of the attitude measures). Interestingly, in this comparison of P with N2, P is slightly more highly correlated with law and order attitudes than N2. What do we make of this reversal?[9]

The Law and Order scale was originally developed specifically with the Stage 4 to principled morality shift in mind. At this point of development there is a shift in attitudes toward civic libertarian issues (see Rest, Narvaez, Bebeau, & Thoma, 1997). That is, at this specific point in moral judgment development, there is a shift in political attitudes from prioritizing social order, unquestioned deference to authorities, and rejection of deviance to the prioritizing of individual welfare, questioning of authority, and tolerance of deviance. Whereas the narrowness of the P index is usually a limitation for a general measure of moral judgment development, in the case of relating to law and order measure, its narrowness is a slight advantage. Whereas the broader N2 index is usually superior to P, in this case, its broadness is a slight disadvantage—although N2 still picks up the shift in the law and order measure, and P has an advantage of only a few correlational points (e.g., −.58 in contrast to −.53). In support of the "broadness–narrowness" interpretation of these findings, consider an even more specific index of moral judgment, P − Stage 4 (formed by simply subtracting Stage 4 from principled morality). The P − Stage 4 index in effect isolates the shift between Stage 4 and principled morality by increasing the score for gains in P at the expense of Stage 4. In the Rest et al. (1974) data, the P − Stage 4 score is even more strongly correlated with law and order than either N2 or P: $r = -.65$ (contrasted to −.53 or −.58). However, the P − Stage 4 index is not a good general index of development, for instance, as manifest in its correlation with moral comprehension ($r = .58$ compared with .69 with N2). Moreover, the other statistical comparisons of P − Stage 4 in other criteria and samples also show this index to be a poorer general index in comparison with N2 or P. Nevertheless, the pattern of results supports the broadness–narrowness explanation as to why N2 generally outperforms the P index but not in the case of correlations with law and order.

Table 6
Correlations of P and N2 With Civic Libertarian Attitudes

Index	Rest et al. (1974) study ($N = 140$) Political toleration	 Law and order attitudes	Rest (1986) study ($N = 96$) Political awareness	 Law and order attitudes
P	.59***	−.58***	.51***	−.61***
N2	.59***	−.53***	.50***	−.56***

$^{***}p < .001$.

Table 7
Internal Reliability (Cronbach's Alphas) of P and N2 Indexes

Sample	Cronbach's α P index	 N2 index
1995 composite sample ($n = 932$)	.78	.83
1979 composite sample ($n = 994$)	.76	.80

Note. Only participants who gave complete data on every story are included in this analysis.

Internal Reliability

Table 7 presents Cronbach's alpha on the two indexes from two heterogeneous samples, the 1995 composite sample and the 1979 composite sample, each with about

[7] On the short form, the comparable correlations with moral comprehension in the Rest et al. (1974) study are P at .68 and N2 at .67—about the same as for the long form. In the Rest (1986) study, the correlations are P at .31 and N2 at .33—again, about the same as for the long form.

[8] On the short form, the correlation of P with the Duncan Scale is .34, with community involvement is .28, and with civic responsibility is .42; the comparable correlations with N2 are .40, .29, and .42, respectively. All correlations are statistically significant at $p < .01$.

[9] On the short form, the correlations of P with the four corresponding measures of political attitudes in Table 6 are .57, −.56, .49, and −.52; for N2, the corresponding correlations are .58, −.52, .47, and −.50.

1,000 participants. For each index, a story score was computed (for each of the six dilemmas), and the Cronbach alpha represents the internal consistency of the six dilemma scores. Although internal reliability is not one of the five major validity criteria, it is the case that N2 is clearly superior to P in internal reliability.[10]

Correlations Between P and N2

So far we have emphasized differences between the P index and the N2 index. Table 8 indicates that the two indices are highly correlated in the .90s. Separating the two components of N2 (the ranking component—essentially the P score) and the rating component (the difference in rating the high-stage items from rating the low-stage items), we see that the P index is virtually identical with the first component of N2 (there is only a minor difference in handling missing data). But the correlation between the P index and the second component of N2 is also considerable (r = .83 and .80, in the two samples). Recall that the P index and Part 2 of N2 are based on different data. The slight difference between the two allows the N2 index to have slightly different properties than the P index, and enables N2 to outperform the P index.

Meta-Analytic Summary

Table 9 summarizes the comparisons of the two indexes of the DIT, P and N2, in a meta-analysis. The meta-analysis does the following: (a) groups the findings from different studies and different measures by the validity criteria, (b) expresses the combined trends for each index in terms of "effect size estimates," (c) expresses the difference in effect sizes by comparing P with N2, and (d) finally expresses the overall effect size and effect size difference (combining across studies, measures, and criteria) and gives a probability estimate for N2 being better than P. The bottom line is that the N2 index significantly accounts for a greater portion of the variance in the classic studies as a whole than the P index (p < .01).

More specifically, here is what was done in the meta-analysis: (a) Following the approach of Rosenthal (1994) using the "r family" of estimators, the individual sample results were transformed into a common effect size estimate. (b) Effect sizes for the separate samples were summed and averaged for each validity criterion (thus grouping trends across the separate samples and measures) in terms of Z transformations of measures of association. (c) Given our interest in contrasting the two indexes, P and N2, the difference between the two averages were statistically compared using an error term corrected for correlated observations (e.g., Dunn & Clark, 1969). As can be seen in Table 9, N2 statistically outperforms P on the intervention criteria, on differentiation of educational groups, and on longitudinal gains; N2 is equal to P on three other contrasts, and is in one instance, slightly weaker (i.e., correlations with civil libertarian attitudes). (d) Continuing to pool effect sizes across all studies, measures, and criteria, the overall effect size of P (for the classic studies) is .69 and the overall effect size for N2 is .94. Note that the overall estimate of the difference in effect sizes is somewhat unstable given the use of multiple effect sizes from the same sample. (e) The difference between the overall effect size of P and N2 is significant at p < .001 level, whether the different sample sizes are taken into account or not (.69 vs. .94, or .68 vs. .87). Therefore, in sum, the meta-analysis indicates that N2 beats P on the benchmark statistics of the classic studies as an indicator of general development in moral judgment.

Table 8
Correlations of P With N2 and N2 Components

Sample	N2	N2 Part 1 (Ranks: Stages 5 + 6)	N2 Part 2 (Rates: High–low)
1995 composite sample (n = 932)	.9509	.9958	.8272
1979 composite sample (n = 994)	.9392	.9953	.7967

Note. Only participants who gave complete data on every story are included in this analysis.

Table 9
Meta-Analysis Comparisons of P With N2

Validity criteria	N	Effect size		Z test of Difference[a]
		P index	N2 index	
Sensitivity to educational intervention	503	.23	.31	4.07***
Differentiation of cross-sectional education groups	2,033	.85	1.29	49.08***
Longitudinal gains	359	.76	.99	7.42***
Correlations with life experiences	331	.67	.66	−.01
Correlations with moral comprehension	230	.63	.65	.02
Correlations with prosocial behavior	215	.38	.41	.03
Correlations with civic libertarian attitudes	457	.66	.62	−2.03*
Total (weighted by sample sizes)	4,125	.69	.94	36.16***
Total (unweighted by sample sizes)		.68	.87	27.15***

[a] Z test of difference is the Z transform of the effect size estimate for N2 minus the Z transform of the effect size estimate for P2 (see Rosenthal, 1994). A positive value indicates that N2 had a greater effect size than P; a negative value indicates that P had a greater effect size than N2.
*p < .05. ***p < .01.

[10] On the short form, for the 1995 composite sample, the Cronbach alpha for P is .65 and for N2 is .74; for the 1979 composite sample, for P the Cronbach alpha is .63 and for N2 is .70. Shortening the test from six stories to three stories lowers the internal reliability by 9 to 13 points.

Discussion

The short conclusion of this paper is that a new index, N2, is generally better than the P index. N2 generally beats the benchmark statistics of the P index in many of the classic studies—not always but generally. The meta-analysis indicates the kinds of studies where N2 outperforms the P index and the studies in which the difference is slight. It now remains to be seen if other researchers who use the DIT also find that N2 outperforms the P index in their studies. If so, this means that researchers using N2 will find better trends in their studies without collecting any additional information from participants. Although we have not attained the alchemists' dream of creating gold out of lead, we have achieved two goals that years ago seemed too good to be true: (a) useful information about moral judgment development that comes from a multiple-choice recognition task; and (b) a new index that can outperform the P index at no extra cost.

A longer conclusion involves taking account of the larger context of instrument development—placing the job of searching for an index in broader context—and considering the implications of what we have learned about the construct, moral judgment, from this research. From a broad perspective, there are at least four sets of decisions that must be made in devising a measure of moral judgment development: (a) What features of thinking are to be used in characterizing development? (e.g., Kohlberg's stages?) (b) What information-collecting procedure will be used? (e.g., Kohlberg's interviews or DIT's recognition task?) (c) How does one index a developmental score for each participant? (This is the main question of the current paper.) (d) How does one validate a measure? (How operationally does one define construct validity?) These issues are discussed in some depth in various references. The first issue (regarding a Kohlbergian approach) is discussed in Rest, Narvaez, et al. (1997). The second (regarding method of data collection) is discussed in Rest, Thoma, & Edwards (1997). The third (regarding indexing) is discussed in the present paper as well as in Rest, Thoma, & Edwards. The fourth (regarding validity) is discussed in Rest, Thoma & Edwards and in Thoma et al. (1997). Recently, Sanders, Lubinski, and Benboe (1995) have raised the question of the discriminative validity of the DIT apart from verbal ability—Thoma et al. (1997) directly respond to this challenge.

In contrast to Lind (1995), who claimed that the DIT can be reduced to sociopolitical attitudes, Sanders, Lubinski, and Benbow (1995) stated: "The DIT is simply another way of measuring verbal ability. . . . If we are to continue using the DIT in psychological research . . . it is imperative that a well-established marker of verbal ability be used" (p. 502). Sanders et al. suggested that they are the first to raise this issue: "Yet [no studies] have specifically examined the uniqueness of moral reasoning" (p. 499). They based their interpretation on the finding that the DIT does not correlate with their selection of personality variables.

Actually, DIT research has attended to discriminant validity for over 20 years. In Rest's 1979 book (from which Sanders et al., 1995, cite), a section entitled, "The Distinctiveness of Moral Judgment" (p. 198–203) cites six studies that partial out verbal ability and general cognitive ability. For instance, after partialing out the Differential Abilities Test (DAT), the semi-partial correlation of the DIT with moral comprehension is .51 ($n = 73, p < .01$); partialing out the DAT, the semi-partial correlation of the DIT with political attitude (libertarianism) is .36 ($n = 73$, $p < .01$). Moreover, challenges to the discriminant validity of the DIT is not limited to verbal ability (or general cognitive ability). Other researchers have proposed reducing the DIT to sociopolitical attitudes, to years of education, and to gender orientation (see Rest et al., 1997). For over 20 years, DIT researchers have been monitoring the question of discriminant validity, and the evidence does not favor reducing the DIT to any simpler variable than moral judgment (see Thoma et al., 1997).

The evidence that Sanders et al. (1995) produced is that correlations of the DIT with personality trait measures are nonsignificant (especially after partialing out verbal ability). Rest (1979) stated: "Of approximately 150 correlations between the DIT personality measures, most are nonsignificant. . . . The DIT is more related to cognitive processes than to personality traits" (pp. 197–198). The pattern of findings in the Sanders et al. study is consistent with the findings reported in Rest. The only hypothesis relevant to the findings of Sanders et al. is that the DIT is correlated with everything—a hypothesis that we have not advanced. The actual validity claims for the DIT (for instance, as outlined in this article) are not even addressed by Sanders, and, in fact, the evidence for 20 years has been in support of the discriminant validity of the DIT on all the criteria.

The main focus of this paper is on indexing. For a long time, we have assumed that the key to building an index better than P was to devise some sort of scaling algorithm. Scaling adjusts the item weights empirically, according to the actual choice behavior of participants. Scaling procedures seemed to offer three advantages: (a) In scaling, all the items are used—items designed for low stages as well as items designed for the high stages. Therefore, an index based on scaling in effect uses more information than the P items (Stages 5 and 6). (b) The scale values for every item are derived independently of the theoretical stage designations. Scaling seemed to offer a kind of empirical hindsight, in which the better items of a test determine more a participant's general score than the poorer items. (c) Scaling weights are calibrated in small units rather than in terms of only six groupings and therefore seem to offer greater precision.

Unexpectedly, our experience has been that building an index that uses empirically derived item weights turns out not to be the best way to build an index for the DIT. We will illustrate this point with one example. Looking at the matched t-test statistic in the combined intervention sample, in Table 1 it was reported that the "benchmark" t-test value for the P index was 11.21 and for N2 the statistic was better (13.94). Using the D index (an index based on scaling, by Davison [1979]), the t-test value is only 8.53. If an N2 index is modified to take account of item weights (rather than counting each item equally according to its a priori stage designation) the statistic is 12.93 (less powerful than the trend using the plain, nonweighted N2 index). The improvement of the item-weighted N2 index over the P index (12.92

vs. 11.21) is more likely due to its hybrid structure than due to using item weights. Counter to our expectation for many years that empirically derived item weights would improve an index—if we could only figure out how to arrive at the weights—we have not found a better index that uses weighted items. This is generally true no matter how the item weights are derived (multidimensional scaling, factor analysis, correlations with education, etc.), no matter which validity criteria are used (e.g., age–education trend studies, intervention studies), no matter which statistic describing the trend is used (correlations, *t* tests, ANOVAs), and no matter which samples are used (e.g., 1995 composite sample, 1979 composite sample, intervention samples). In sum, weighting the individual items does not add to the strength of trends.

What sense do we make of this? We are not sure, but one possibility is that these investigations are telling us that it is important to regard the items as indicators of stages as classes of items, and not as separate items devoid of their superordinate stage groupings. To put this point another way, Kohlberg's theory in effect represents the items of a stage as belonging to larger systems of meaning (the stages), each stage being composed of many items. On the other hand, scaling theory regards each item on its own, as independent of a system of meaning defined by a stage. Scaling regards each item as if it held a place on its own on the underlying continuum of development, and that the item's membership in a stage system did not matter. What might be called "Kohlberg's truth" is to pay attention to item "clumps"—every item is not on its own. According to this speculation, an index that takes account of *item clumping* (i.e., treats all items of a stage the same rather than adjusting each item by weights) will work better than an index that assumes complete item independence.

Another issue arises from the experience with research leading to N2: Why does a hybrid index work better than a simple index? Two explanations occur to us. One is the relatively straightforward explanation that N2 uses more information than P: It uses both ranking and rating data; hence in effect it makes a longer test (more bits go into the aggregate). A longer test is generally more accurate than a shorter test; the redundancy gives error factors more chance to cancel themselves out. This explanation is not completely satisfactory in that we have tried various ways to combine the lower stages with the higher stages, combining the rating data with the ranking data (e.g., combining the P index with the D index, using low-stage items in combinations with high-stage items), and find that most of these combinations do not outperform the P index even though each combination is using more data than the P index. There seems to be something special about the particular combination of elements that go into the N2 index.

The second explanation advances the idea that there is something "synergistic" about the interaction between the two specific elements of N2. That is, the two kinds of information interact with each other: One part boosts the score when the other information source underestimates the score, and decreases the score when the other information source overestimates it. Each information source contributes something that the other lacks and serves as a correction to the other. We illustrate this synergistic effect by reference once again to the combined intervention data. The matched *t*-test statistic for the test–retest effect using N2 is 13.94. The comparable statistic for the first part of the index (the prioritization of Stages 5 and 6) is 11.53. The comparable statistic for the second part (discrimination and rejection of the lower stages) is 12.75. Note that both of these parts independently produce lower statistics than their combination in N2, 13.94. Their combination is more than either part alone or their average together. How does this work? Examining the Bebeau and Thoma (1994) study provides a clue. In this study, about half of the participants were gaining on P and losing on Stage 4 with no appreciable loss of Stages 2 + 3. (These participants produced the gains in P score shown in the intervention study.) The other half of the participants were gaining on Stage 4 with appreciable loss in Stages 2 + 3, and little loss on P. (These participants did not show up as gainers on P but do show up as gaining on differentiation and rejection of the lower stages.) Note that Stage 4 can go up or down—what matters is the amount of differentiation of the high stages from the low stages. (When participants are losing in Stage 4 they are gaining in P, thus increasing the difference between high and low; when participants are gaining in Stage 4 they are losing in Stages 2 + 3, thus also increasing the difference between high and low.) Hence an index that attends to both of these effects works better in producing a general index of development than either alone.

Another way to view this study is that after an immense amount of work, the P index shows that it has captured almost all of the trends of the DIT. The overall improvement of N2 is not spectacular, and there are specific situations in which the P index is still preferable. For instance, the P index is preferable when specific hypotheses about postconventional (principled) morality are at stake. Similarly, other specific indexes may be preferable (such as Stage 4 in relating a "law and order" orientation to other variables, or the difference in ratings between the high stages and low stages [N2 Part 2] when there is specific interest in this effect in assessing an education program). Nevertheless the synergistic effect of N2 (combining N2 Part 1 with N2 Part 2) recommends N2 as the best general index of the DIT.

The most far-reaching inference from the present study is that we are now beyond the methods of alchemy. In the psychological study of morality, we are in a different place than we were several decades ago. We are not captives of hit-or-miss speculation—the field blowing one way or the other, depending on the climate of the times or the sheer tenacity and verbal facility of the advocates. Many ideas that have initial, intuitive appeal can be put through the crucible of empirical tests (e.g., that a recognition task can yield usable data, or that multidimensional scaling will build a better index, or that individual MANOVA techniques will yield a better methodology, or that social cognition is better studied in separate domains, or that "care reasoning" is a separate pathway of development for females than the pathway males take). We will always need new ideas and new approaches, but their usefulness can be checked out through a series of empirical tests. In the process of checking out the new ideas, sometimes unexpected findings will lead to even newer, more fruitful ideas.

References

Bebeau, M., & Thoma, J. (1994). The impact of a dental ethics curriculum on moral reasoning. *Journal of Dental Education, 58,* 684–692.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment: Vol. 1.* New York: Cambridge University Press.

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

Davison, M. L. (1979). The internal structure and the psychometric properties of the Defining Issues Test. In J. Rest (Ed.), *Development in judging moral issues* (pp. 223–245). Minneapolis: University of Minnesota.

Davison, M. L., & Robbins, S. (1978). The reliability and validity of objective indices of moral development. *Applied Psychological Measurement, 2,* 391–403.

Deemer, D. (1986). *Moral judgment and life experience.* Unpublished doctoral dissertation, University of Minnesota.

Duckett, L., & Ryden, M. (1994). In J. Rest & D. Narvaez (Eds.), *Moral development in the professions: Psychology and applied ethics* (pp. 51–70). Hillsdale, NJ: Erlbaum.

Dunn, O. J., & Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64,* 366–377.

Evens, J. (1995). *Indexing moral judgment using multidimensional scaling.* Unpublished doctoral dissertation, University of Minnesota.

Kohlberg, L. (1979). Foreword. In J. Rest (Ed.), *Development in judging moral issues* (pp. vii–xvi). Minneapolis: University of Minnesota Press.

Lind, G. (1979). Moral development—a new issue in higher education research. In G. Framheim (Ed.), *Report of a special seminar in university students—Their training and conception of life* (pp. 52–61). Unpublished manuscript, University of Konstanz, Germany.

Lind, G. (1995, April). *The meaning and measurement of moral competence revisited.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Lind, G., Hartmann, H. A., & Wakenhut, R. (Eds.). (1985). *Moral development and the social environment.* Chicago: Precedents Publishing.

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

McNeel, S. (1994). College teaching and student moral development. In J. Rest & D. Narvaez (Eds.), *Moral development in the professions: Psychology and applied ethics* (pp. 27–50). Hillsdale, NJ: Erlbaum.

Penn, W. Y., Jr. (1990). Teaching ethics—A direct approach. *Journal of Moral Education, 19,* 124–138.

Rest, J. (1979). *Development in judging moral issues.* Minneapolis: University of Minnesota Press.

Rest, J. (1986). *Moral development: Advances in research and theory.* New York: Praeger.

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

Rest, J., Narvaez, D., Bebeau, M., & Thoma, S. J. (1997). *Development, domains, and culture in morality: A neo-Kohlbergian approach.* Manuscript submitted for publication.

Rest, J., Thoma, S. J., & Edwards, L. (1997). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology, 89,* 5–28.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Sanders, C., Lubinski, D., & Benbow, C. (1995). Does the Defining Issues Test measure psychological phenomena distinct from verbal ability? An examination of Lykken's query. *Journal of Personality and Social Psychology, 69,* 498–504.

Self, D., & Baldwin, D. (1994). Moral reasoning in medicine. In J. Rest & D. Narvaez (Eds.), *Moral development in the professions: Psychology and applied ethics* (pp. 147–162). Hillsdale, NJ: Erlbaum.

Thoma, S. J., Narvaez, D., & Rest, J. (1997). *Does the Defining Issues Test measure psychological phenomena distinct from verbal ability? Some relevant data.* Manuscript submitted for publication.

Received November 4, 1996
Revision received February 4, 1997
Accepted February 4, 1997 ■