Child Development, March/April 2001, Volume 72, Number 2, Pages 535–548

Hierarchical Structure of Moral Stages Assessed by a Sorting Task

Jan Boom, Daniel Brugman, and Peter G. M. van der Heijden

Following criticism of Kohlberg's theory of moral judgment, an empirical re-examination of hierarchical stage structure was desirable. Utilizing Piaget's concept of reflective abstraction as a basis, the hierarchical stage structure was investigated using a new method. Study participants (553 Dutch university students and 196 Russian high school students) sorted statements in terms of moral sophistication. These statements were typical for the different stages of moral development as defined in Colby and Kohlberg. The rank ordering performed by participants confirmed the hypotheses. First, despite large individual variation, the ordering of the statements that gave the best fit revealed that each consecutive Kohlbergian stage was perceived to be more morally sophisticated. Second, the lower the stage as represented by the items, the higher the agreement among the participants in their ranking; and the higher the stage as represented by the items, the lower the agreement among the participants in the rankings. Moreover, the pivotal point depended on the developmental characteristics of the sample, which demonstrated a developmental effect: The ordering of statements representative of moral stages below one's own current stage was straightforward, whereas the ordering of statements above one's own stage was difficult. It was concluded that the Piagetian idea of reflective abstraction can be used successfully to operationalize and measure the hierarchical nature of moral development.

INTRODUCTION

Hierarchy is a fundamental property of developmental stages (Piaget, 1960). Nevertheless, the notion of hierarchy has been discussed almost solely by the philosophically minded (e.g., Wren, 1990) and only in the context of moral development. Few empirical studies have been devoted to hierarchy. Studies related to ordering in chronological time are mostly irrelevant. It is important to distinguish between ordering in time (an invariant stage sequence) and ordering in an evaluative or logical sense (hierarchical stage patterning). In this article, a new approach to the issue of hierarchy is proposed. The approach is based on the view that there is a link between hierarchy and the developmental mechanism.

Hierarchical ordering, in any stage theory, implies a sequence of stages that forms a developmental pattern based on a particular type of relation, sometimes described in terms of a developmental logic (see Van Haaften, Korthals, & Wren, 1997). For example, in Kohlberg's theory of moral judgment, the stages represent different deep structures that are assumed to be increasingly adequate, in philosophical terms, as justifications for moral choices and actions. Change from one stage to another involves a reorganization of thought and structural transformations (Colby & Kohlberg, 1987). In Rest's classic empirical study "The Hierarchical Nature of Moral Judgment," hierarchy is explained as follows (Rest, 1973, as cited in Puka, 1994, p. 34): "Each succeeding stage is said to be an advance over the preceding stage in being a more differentiated and integrated structure. In other words, the stages are said to be *hierarchically* related: a new stage does not simply replace a previous stage, nor is it added to it, but rather the new stage is a transformation of elements of the old along with new elements into a new emergent structure."

The theoretical model tested in the present study is an interpretation of Kohlberg's theory based on Habermas (1990) and Piaget (1977a, 1977b). The issue of hierarchy can be addressed in an empirically meaningful way if the Kohlberg model of moral development is interpreted as a reconstruction of changes in the moral intuitions of subjects as described by Habermas (1984). This theoretical reconstruction should reveal the structure and the rules by which this domain is organized. Whether these structures and rules are correctly reconstructed in the theory can be tested by assessing whether actors actually recognize them as the structures and rules that guide their thinking on moral problems. Because one important feature of the way the moral domain is organized reveals itself in the hierarchical order of the structures of thinking, looking at the hierarchy of the stages as perceived by the participants makes sense.

This interpretation of the Kohlberg model of moral development makes even more sense when the developmental process is taken into consideration. It focuses on reflective abstraction, in the Piagetian sense, construed as the self-organization of meaningful content (Piaget, 1977b; see Boom, 1991). In reflective abstraction, the basic principle is critical reflection on one's own manner of thinking. This, however, is not necessarily a conscious process. Reflective abstraction leads to a form of inclusion of previous thinking in the next stage, but in a reorganized form (Boom, 1991). For example, once you come to see that, in orienting yourself toward the immediate social group (a typical Stage 3 view), everybody else is probably doing the same, you will realize the inherent limitation of this view. In other words, grasping the structure of your own thinking by generalizing your previous thinking activities is the crucial step in proceeding to the next stage (in this case the social system perspective of Stage 4).

From these considerations on critical self-reflection, hierarchical ordering, reconstruction of the developmental process, and mechanisms for developmental processes, it is suggested that an important aspect of moral reasoning competence is to be able to evaluate the arguments of others, at least those of lower stages than one's own current stage. The type of thinking characteristic of lower stages is assumed to have been subjected to critical reflective abstraction in the course of a normal subject's developmental trajectory. Therefore, reasoning at stages below one's own current stage should be recognized as defective, regardless of the position (pro or con a certain course of action) taken in these arguments, whereas arguments above one's own stage cannot be expected to be judged consistently. Thus, an individual's current stage is assumed to be adequately characterized by the ability to criticize earlier ways of thinking as inadequate by form.

Empirical Evidence for Hierarchy in Moral Development

The empirical evidence for the existence of a hierarchical ordering of stages (as in Kohlberg's approach to morality) is indirect. According to Walker, de Vries, and Bichard (1984), the general pattern obtained across several experimental studies was for participants to prefer the reasoning of stages higher than their own current stage, and to reject lower stage reasoning (as measured by their ratings of stage-prototypic statements). In addition, participants' understanding (as measured by their ability to paraphrase stage-typical statements accurately) was limited to no more than one stage higher than their own current stage. Walker et al. (1984) criticized these studies for not using the full spectrum of stages and not ruling out the role of the complexity of the language of stage-prototypic statements as an alternative explanation. In their own study, Walker et al. used five stages, addressed the possible effect of nonstage features, and concluded that the hierarchical pattern was still found even after controlling for language difficulty. The fact, however, that understanding (i.e., giving a stage-adequate paraphrase) of statements representing one stage higher than the stage assigned to the participants is possible remains puzzling (see the Discussion section for more on this and on additional concerns).

Another type of empirical evidence comes from questionnaires yielding score profiles that allow a statistical pattern analysis of evaluations. These score profiles are obtained by asking a participant to evaluate (e.g., in terms of preference, rejection, or importance) a series of items representing different stages. This approach can be seen as a generalization of the experimental approach. The Defining Issues Test (DIT) designed by Rest (1979) is a well-known instrument that provides score profiles. In the normal use of this questionnaire the ratings of the items are not directly employed. Davison, Robbins, and Swanson (1978), however, used a unidimensional metric unfolding technique to analyze the ratings of statements and found support for the ordering of the stages. Boom and Molenaar (1989) used LISREL to analyze such score profiles and introduced a hybrid model that took into account the fact that in samples where not all participants have reached the highest possible stage, the higher stages are more difficult to rate. They found that participants could rate DIT statements (in terms of preference) below their own current stage, but could not do so consistently for statements representing higher stages. Effects were not strong, however, and this result has not been replicated by others. Although the use of evaluations of moral statements is also found in the instruments devised by Gibbs and coworkers (e.g., the Sociomoral Reflection Objective Measure–Short Form, SROM-sf; Basinger & Gibbs, 1987), there are no known empirical studies relevant to this study's research question based on score profiles obtained with this instrument.

In a recent publication, Rest, Thoma, Narvaez, and Bebeau (1997) introduced a new index for the DIT that they claim is an improvement. Whereas the old P-index actually measured only the degree to which higher stage (postconventional) thinking is endorsed, the new index (N2) also takes into account the degree to which lower stage reasoning is rejected. This finding is taken to support the assumption on critical self-reflection in this study.

Although the discussion about the best way to obtain moral judgment data continues (see Langford, 1997; Rest, Thoma, & Edwards, 1997), the issue of hierarchy is not addressed directly by these recent discussions among test designers. This omission is addressed here.

A New Test of Hierarchic Structure

The Moral Judgment Sorting Task (MJST) is a sorting task designed to test whether subjects can recognize the inherent hierarchical ordering of stages as defined in the official Standard Issue Scoring Manual (Colby & Kohlberg, 1987). The task involves statements representing all the Kohlberg stages (Stages 1 through 5) as well as intermediate stages (resulting in nine categories). Intermediate stages were included in this study because they are sufficiently demarcated in the manual as separate ways of responding, although admittedly, the theoretical status of these in-between stages still needs further clarification. Participants were asked to sort these nine statements, indicative of moral structure, from simplistic to wise. The intention was to investigate the relation between the ordering of the stages proposed by Kohlberg and the average ordering of the stages in the MJST given by the participants.

First, it was expected that the ordering proposed by Kohlberg would be reproduced, but second, because of a developmental effect, the ordering would be more clearly reproduced for the statements representing the lower stages than for those representing the higher stages. It is possible to approach both expectations in several different ways.

The first possible approach to the first expectation is to compute for each of the items (statements) the average rank they elicited in the particular sample of participants. If the averages of the ranks of each of the nine items forms a smoothly increasing series, then the hierarchy is corroborated; otherwise a hierarchical ordering becomes questionable. Unfortunately, there is a problem with this approach: The average ranks provide no indication of the fit of the model. There might be too much individual variance, such that, notwithstanding an increasing series, results are still questionable.

The second approach, therefore, is to look at the pattern of rankings given by each individual participant and compare this with the theoretically derived ordering of the nine categories given by Kohlberg. The obvious way to investigate this relation is by calculating for each participant a straightforward Pearson Product Moment Correlation coefficient. Participants can have low, even negative, correlations, but many are expected to have positive and high correlations. Of particular importance is the average computed over all participants because this is an index for how well the actual rankings fit the theoretically derived ordering. Moreover, this overall index can be compared with alternative orderings of the statements. Suppose that, on the basis of inspection of the average ranks per item (first approach), the decision is made to reorder the items; then the individual correlations would change too. Only when this leads to a clear rise in the average correlation over all participants does this reordered sequence describe the data better.

The third approach generalizes this latter idea by determining objectively the best ordering and spacing of the items that best describe the sample. Because the nine categories need not necessarily be equally distanced in a computation of the correlation coefficient, an optimal scaling technique can be used to quantify optimally the rankings of Categories 1 through 9. That is, the distances between the categories are varied in such a way that the *average* of the correlations between Kohlberg's ordering of the categories and each participant's ordering of the categories is maximal. This variation of distances may imply a revised ordering. If sorting is consistent but deviates from the expected ordering principle, maximization of the average correlation will be possible only by quantifying the statements according to their preferred alternative ordering. If the sample is not consistent in their sorting, maximization of the average correlation will not be possible at all.

Keeping the above approaches in mind, the first hypothesis of this study was formulated more precisely: It was expected that the optimal quantifications for the ranking of the Kohlberg stages would result in a reasonably high average correlation and would display a clear, smooth increase. Only insofar as this was the case, would the stages as defined and operationalized in the Colby and Kohlberg manual be corroborated, particularly in their hierarchical sequence but also, by implication, in many other respects.

The second expectation of this study concerned the developmental effect. It was based on the idea that the ordering of statements that represent lower stages than the participants' own current stage would reflect the hierarchical ordering according to the theory, whereas ordering of statements higher than their own current stage would reflect that these participants were not acquainted with these stages. Although assessing someone's actual stage in a reliable and precise way is difficult, it was nevertheless assumed that, from the point of view of the participants, statements representative of higher stages than their own current stage would appear undifferentiated and elicit no clear orderings. Moreover, even if these statements were given a rank higher than the participants' own current stage, it is unlikely that the subtle hierarchy among those statements would be appreciated; the less so, the more these items are removed from someone's own current stage. To test the second expectation, several approaches were again possible.

The first approach is to look at the average and the distribution of the ranks per item. If, for a particular sample, two statements are perceived as alike, the differences in ranking in the sorting task will be randomly distributed and the average will cancel out this random variation, resulting in equal average ranks. If a sharp distinction is made between two items, however, the average ranks will differ significantly.

Second, the quantifications obtained for each stage after optimization may be considered. These too reveal, for a particular sample, which pairs of stages are, on average, perceived as different and which are perceived as alike. Therefore, the quantifications for lower stages are expected to be further apart (more different) and for higher stages, to be closer together (almost indistinguishable).

The third approach is based on an external criterion. A moral judgment score is obtained by using an instrument that assesses moral development (the SROM-sf was used in this study). Participants scoring lower on moral development are expected to be more inconsistent and thus to give lower individual correlation with the theoretical ordering proposed by Kohlberg than participants with higher scores on moral development. Thus, a substantial correlation is expected between the score on the SROM-sf and the individual score on the MJST.

Fourth, stage groups based on the SROM-sf score can also be created and compared. For samples lower on moral development and for younger samples, more deviations from the ideal pattern are expected than for more morally mature or older samples.

Thus, based on the above four approaches, the second hypothesis was formulated more precisely: The quantifications after optimization were expected to be further apart for lower stages and closer together for higher stages, depending on the average SROM-sf score of a sample. The higher the range of the SROM-sf score of the sample, the more the increase in quantifications would extend to the statements representing the higher stages and the higher the average correlation would be.

METHOD

Participants

Results for university students were obtained by using two cohorts of Dutch undergraduates: 275 in cohort 1998 (age: M = 23.4 years, SD = 7.35, *range* = 18–58; 56 male and 219 female) and 278 in cohort 1999 (age: M = 20.7 years, SD = 3.42, *range* = 18–51; 52 male and 226 female). With this relatively high educational level, it was hoped that a close correspondence with the ordering in the scoring manual would be found. The high school sample consisted of Russian students from two different cities (Moscow and a city near Moscow). The youngest group (grade 8) consisted of 103 adolescents (age: M = 14.0 years, SD = .32, *range* = 12.6–15.0; 49 male and 54 female). The older group (grade 10) consisted of 93 adolescents (age: M = 16.0 years, SD = .48, *range* = 15.0–16.8; 34 male and 59 female).

Materials

Moral judgments sorting task. After being presented with a dilemma (three times), participants were asked to rank nine statements justifying what to do. Because the ultimate source for scoring moral statements is the Standard Issue Scoring Manual (Colby & Kohlberg, 1987), all statements (representing the stages 1, 1/2, 2, 2/3, 3, 3/4, 4, 4/5, and 5) were derived from examples given in this manual. The statements were translated and adapted slightly to obtain roughly equal lengths in characters (Set 1: M = 73.9, SD = 3.1; Set 2: M = 64.8.1, SD = 3.6; Set 3: M = 54.2, SD = 7.8). A sentence complexity score was computed for all statements with a maximum score of 100 for highest complexity (Set 1: M = 21.7, SD = 9.5; Set 2: M = 12.6, SD = 8.2; Set 3: M = 16.8, SD = 7.6). None of the correlations of complexity with the stage order reached significance.

Instructions were to sort from "simplistic, LEAST SENSIBLE" to "high-minded, MOST SENSIBLE, wise" irrespective of whether one agreed with the course of action chosen. Participants were not asked which statement was the most moral because such an instruction would not be well understood in the Netherlands. "Moreel" is not a frequently used word among students and has connotations of old-fashioned precepts.

The Colby and Kohlberg manual presents more than 500 so-called Criterion Judgments (CJs). A CJ is a scoreable, coherent piece of reasoning classified by *action choice*, *issue*, *norm*, and *element*, together referred to as content of thinking, and by *stage*, referred to as structure of thinking in the manual. For each CJ, the manual gives detailed criteria including match examples. The statements used in the present study were extracted from examples in Form A. Three sets of statements were created, each belonging to a particular dilemma story (see Table 1). Each set of nine state-

Table 1 Characterization of Items in the Moral Judgment Sorting Task

	Stealing		Promise		Saving Life	
Stage	CJ	Element Content	CJ	Element Content	CJ	Element Content
1	2	Blaming (approval)	2	Seeking reward (avoiding punishment)	1	Good (bad) group consequences
1/2	5	Seeking reward (avoiding punishment)	3	Seeking reward (avoiding punishment)	2	Seeking reward (avoiding punishment)
2	7	Obeying (consulting)	6	Seeking reward (avoiding punishment)	4	Balancing perspectives or role taking
2/3	11	Good (bad) group consequences	8	Good (bad) reputation	8[b]	Blaming (approving)
3	14	Obeying (consulting)	16	Good (bad) reputation	9	Serving human dignity and autonomy
3/4	19	Serving social ideal or harmony	20[a]	Serving social ideal or harmony	19	Serving human dignity and autonomy
4	25	Good (bad) group consequences	29	Good (bad) reputation	24	Serving social ideal or harmony
4/5	29	Maintaining social contract	35	Good (bad) group consequences	31	Maintaining equity
5	33	Maintaining equity	40	Serving human dignity and autonomy	36	Serving human dignity and autonomy

Note: Statements were drawn from Colby and Kohlberg (1987). For Stealing, the issue was *law,* and the norm was *law*; for Promise, the issue was *contract*, and the norm was *contract*; for Saving Life, the issue was *life*, and the norm was *life.* CJ = Criterion Judgment.

[a] CJ 20 had *Affiliation* as the norm instead of *contract.*

[b] CJ 8 had *Conscience* as the norm instead of *life.*

ments was taken from CJs that were identical for action choice and as much as possible, identical for issue, norm and element; they differed only in terms of stage. For good reasons, not all possible combinations are given in the manual; therefore, complete identity of *element* within each set was not feasible. Because statements were derived from the scoring manual and the manual, in turn, was based on actual interviews, it is assumed that the statements are ecologically valid. The first and the last set share the same story on Heinz but have opposite action choices. The first set offers nine reasons for not stealing, whereas the last set offers nine reasons for stealing the drug and saving a life.

Taking the average over two or more dilemmas is a common procedure also used by Kohlberg (Colby & Kohlberg, 1987), Rest (1979), and Gibbs (Gibbs, Widaman, & Colby, 1982). An analysis both per set of statements separately and for the combination of three sets is provided.

The SROM-sf. The SROM-sf (Basinger & Gibbs, 1987) is a written instrument that contains two dilemmas, one of which is the Heinz dilemma. After a dilemma is presented, a participant is offered a series of questions related to this dilemma, each with four alternative answers. Participants are asked for each answer (of four) to indicate whether it is close to the answer they themselves would give. Each of these responses gives a *close* score. Participants are then asked which of the four answers most closely reflects their own (*closest* score). The SROM-sf score is the average of the close and twice the closest scores, multiplied by 100. The statements represent Stage 1 through Stage 4 and no in-between stages. The SROM-sf was used mainly to characterize the samples in terms of general moral development and to split the combined sample into stage groups.

Procedure

For the university student samples, both tasks were computer administered (Brand & Houx, 1992). In the computer sorting task, statements were presented in random order. The SROM-sf was administered only to the first cohort of the university student sample. Gibbs recommends eliminating cases that have inconsistent preference patterns, whereas, in the computer version used for this study, participants were required to correct inconsistent answering at the end of the session. For the Russian samples, the tasks were administered in small groups. For the sorting task, paper cards with printed statements were presented in a fixed order.

Analysis

For an important part of the analysis, correspondence analysis was relied upon. This is an optimal scaling technique used to quantify the rankings of Categories 1 through 9 optimally such that the average correlation between Kohlberg's ordering of the categories and the participants' ordering of the categories is maximal. Correspondence analysis (available in SPSS Categories) finds the maximal correlation between the quantified categorical row variable and the quantified categorical column variable in a two-way contingency table (Kendall & Stuart, 1979). The application of correspondence analysis used for this study is described later.

For each participant, an ordering of the nine categories for the Kohlberg stages was collected. Such an ordering is represented in a matrix of nine columns indexed by *i*, representing the stages of Kohlberg, and nine rows indexed by *j*, representing the ordering of the statements given by the participant. If the element of the matrix for person *k* is represented by $x_{ij}(k)$, then $x_{ij}(k) = 1$ if participant *k* sets stage *i* at rank *j*, and otherwise $x_{ij}(k) = 0$. If the theoretically derived ranking of categories by Kohlberg is identical to the ranking given by participant *k*, only the diagonal of the matrix is filled with elements equal to 1.

For ease of exposition, orderings made for only four stages are used as a small example. Assume that for some individual the stages are rank ordered as 1, 3, 4, 2. Then this yields a 4 × 4 matrix (see Table 2A).

Assume that there are 50 individuals and each individual yields a matrix. These 50 matrices are summed so that the quantifications obtained are restricted to be equal over all individuals (the property that multiple correspondence analysis and correspondence analysis are closely related was used; see Gifi, 1990). Assume that we then obtain the matrix in Table 2B.

Notice that 47 of the 50 individuals correctly scored the first stage in the first place, but that switches occurred particularly often for Stages 3 and 4. A correspondence analysis of this 4 × 4 table would yield optimal quantifications for the columns (stages as ordered by Kohlberg) and the rows (stages as ordered by individuals), and these quantifications would be optimal in the sense that they provide the maximal correlation that is possible. Possible discrepancies, however, between quantifications for rows and column would be hard to interpret, and they are therefore restricted to be identical. To accomplish this, the matrix is made symmetric by adding to the matrix its transpose (Table 2C).

The correspondence analysis of this matrix provides optimal quantifications—1.66, .44, .56, and .66 for Stages 1 through 4, respectively. This reveals that, over all individuals, Stage 1 is much better distinguished from the other stages than Stages 2, 3, and 4, although on average 2, 3, and 4 seem to be clearly ordered. The reason is that Stage 1 is almost always ordered correctly, and, therefore, the diagonal value for Stage 1 is very high. These quantifications give a maximized correlation of .92, which can be interpreted as the average maximal correlation given that the quantifications are identical over all individuals. Notice that this high correlation is obtained mainly by quantifying Stage 1 in an opposite way from Stages 2, 3, and 4.

Correspondence analysis reveals two types of information: (1) the quantifications obtained for the rankings, and (2) the maximized correlation obtained by quantifying the rankings of the categories in this way. It should be noted that the quantifications obtained relate to each participant, and that, therefore, the maximized correlation can be interpreted as an average maximized correlation under the restriction that the quantifications are identical for each participant. These correlations can be seen as an index for the effect size. To assess the reliability of the quantifications thus obtained and of the average correlations, a bootstrapping procedure was employed (Hastie & Tibshirani, 1993). By randomly drawing 500 bootstrap samples, 95% confidence intervals were computed.

Table 2 Examples for Constructing Tables for Multiple Correspondence Analysis

A: 1 Participant				B: 50 Participants				C: Matrix B Added to B′			
1	0	0	0	47	2	1	0	94	3	3	0
0	0	1	0	1	39	5	5	3	78	12	7
0	0	0	1	2	7	24	17	3	12	48	37
0	1	0	0	0	2	20	28	0	7	37	56

RESULTS

Results for Dutch University Students

The average SROM-sf score for the first of the two cohorts of Dutch university student samples was 325.5 (*SD* = 18, *range* = 256–371). According to normative sample data provided by Gibbs et al. (1984), this is a normal sample, although it should be noted that the score may be reduced by a ceiling effect resulting from the use of only Stage 1 through 4 statements. The correlation of age with SROM-sf score was not significant, $r(272) = .08$, $p = .208$. For the second cohort, no SROM-sf data were collected.

For the first cohort, the distribution of the ranks assigned to each stage are shown in Figure 1. The series axis of Figure 1 pertains to statements 1, 1/2, 2, 2/3, 3, 3/4, 4, 4/5, and 5. The category axis stands for the rank assigned by the participants. The value axis shows the frequencies. The second cohort had very similar distributions; therefore, detailed results are given only for the first cohort. A global comparison between the cohorts follows.

The average for the straightforward Pearson correlation between the ordering provided by Kohlberg and the ordering given by the participants was .49 for set 1, .64 for set 2, .64 for set 3, and .59 averaged over all three sets. A good fit with the theoretically predicted sequence would be represented by peaks on the diagonal in Figure 1. Such a pattern was found for the lower statements. It might be asked whether there are alternative orderings that would result in a better fit for the sample. Correspondence analysis provides the best possible fit by computing the optimal quantifications of the rankings. These optimal quantifications of the rankings led to maximized average correlations of .57 for Set 1, .80 for Set 2, .74 for Set 3, and .68 for the combined model. Confidence intervals were computed by bootstrapping. These results are summarized in Table 3.

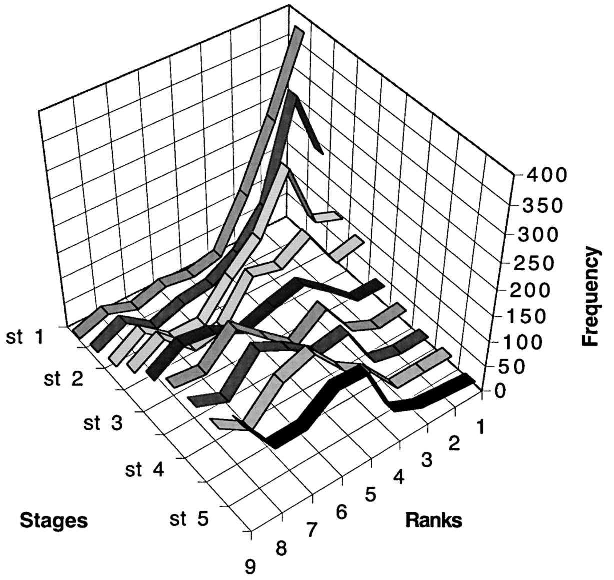

Figure 1 Cohort 1998 of Dutch undergraduates. Distribution of frequencies of assigning ranks to each stage.

Table 3 Average Correlations for Three Sets of Statements and for Sets Combined for Cohort 1998 of Dutch Undergraduates and for Grade 8 of Russian High School Students

	Stealing	Promise	Saving Life	Combined
Cohort 1998				
Straightforward correlation	.49	.64	.64	.59
SD	.30	.23	.27	.20
Maximized correlation	.57	.80	.74	.68
Confidence intervals[a]	.539–.611	.773–.827	.714–.777	.661–.707
Grade 8				
Straightforward correlation	.25	.39	.38	.34
SD	.35	.30	.34	.20
Maximized correlation	.35	.83	.73	.58
Confidence intervals[a]	.288–.399	.772–.886	.654–.800	.528–.619

[a] Confidence intervals based on 500 bootstraps.

The quantifications for each rank per set and for the combined model are shown in Figure 2. Quantifications reflect linear distance. They are chosen so that they sum to zero. The values for the quantifications have no intrinsic meaning; the relative distance between these quantifications is what matters, and therefore they can be linearly transformed.

Figures 2A through 2C show a clear increase in the quantifications for the main stages up to Stage 4. In Figure 2A, some irregularities obtained for the highest stages were also reflected in the lower average correlation for this set of statements. In Figures 2B and 2C, the expected leveling off can be seen.

Figure 2D gives the results of the three sets combined. Combining the three sets as described in the Method section is not the same as taking the averages of the quantifications of the three separate sets. In this case, the improvement of the maximized average correlation was modest (it went up from .59 to .68). With regard to the first hypothesis, the best fitting quantifications for this combined model display a smooth increase for the Kohlberg stage ordering with substage 4/5 showing a clear deviation. With regard to the second hypothesis, the stage ordering of the combined set displays the expected leveling off for the higher stages.

The correlation between the SROM-sf score and the individual index of the MJST (correlation per participant) was modest, $r(272) = .42$, $p = .000$. The correlation of age with individual index of the MJST was not significant for this sample, $r(275) = -.07$, $p = .252$.

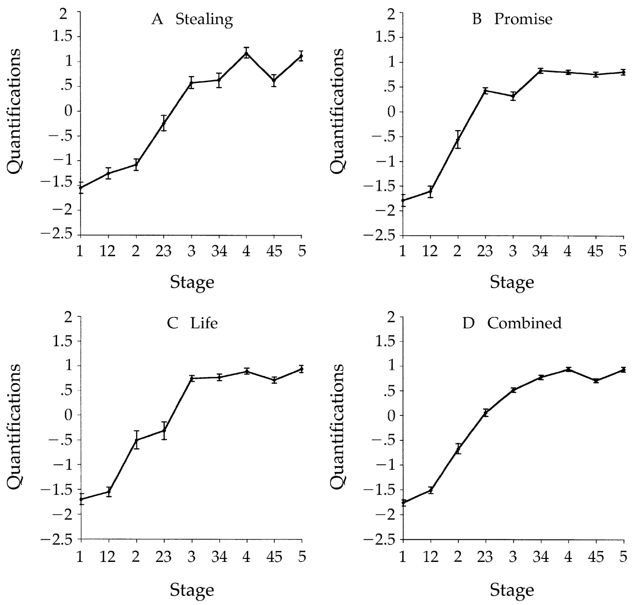

Figure 2 Cohort 1998 of Dutch undergraduates. Means and 95% confidence intervals for quantifications of the rankings found by a bootstrapping procedure, for three sets (A–C), and for all sets combined (D). Intermediate stages, as used by Colby and Kohlberg (1987), are indicated by combining the adjacent stages numbers. The values of the quantifications have no intrinsic meaning.

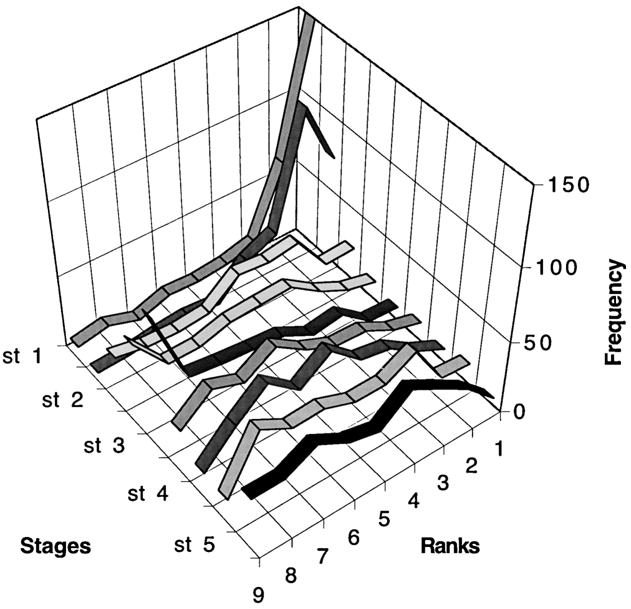

Figure 3 Grade 8 of Russian high school students. Distribution of frequencies of assigning ranks to each stage.

Results for Russian High School Students

The average SROM-sf score for the youngest Russian high school sample (grade 8) was 284.7 (*SD* = 28, *range* = 212–347). The average SROM-sf score for the older Russian high school sample was 298.7 (*SD* = 28, *range* = 224–361). Both were low, as expected, compared with the university student sample.

Detailed analysis is given only for the grade 8 sample. The grade 10 sample gives slightly different results that will be summarized later. The obtained frequency matrix of ranking scores for MJST statements for the grade 8 sample is summarized in Figure 3 (see description of Figure 1).

The average of the straightforward Pearson correlation between the ordering provided by Kohlberg and the ordering given by the participants was .25 for Set 1, .39 for Set 2, .38 for Set 3, and .34 overall. The optimal quantifications of the rankings led to maximized average correlations of .35 for Set 1, .83 for Set 2, .73 for Set 3, and .58 for the combined model. Confidence intervals for the mean were computed by bootstrapping. The results are summarized in Table 3. The quantifications for each rank per set and for the combined model are shown in Figure 4. The combined model gave a maximized average correlation of .58, which is much better than the straightforward average correlation of .34. All sets, however (Figure 4A to 4C) showed a significant increase in quantifications only for the stages up to stage 2, whereas Figure 4A had a strongly deviant pattern.

For the combined model, the increase for the Kohlberg stages was smooth up to Stage 3 and fell flat after that. With regard to the first hypothesis, support was weak because only the first three stages were involved. With regard to the second hypothesis, the expected leveling off was seen for all sets.

The correlation of age with SROM-sf score was not significant, *r*(87) = .18, *p* = .098. The correlation of age with MJST correlation per participant was not significant, *r*(103) = .05, *p* = .592. The correlation between SROM-sf score and correlation per participant was low, *r*(87) = .33, *p* = .002.

Results for Samples Combined

The samples for which we obtained SROM-sf data were first combined (all except cohort 99 of the university students) into one sample (*N* = 471, *M* = 19.9 years, *SD* = 7.0). The average SROM-sf score was 312 (*SD* = 29; 24 participants had no valid SROM-sf score). The correlation of age with SROM-sf score was low, *r*(447) = .38, *p* = .000. The correlation of age with the individual index of the MJST was low, *r*(471) = .25, *p* = .000. The correlation of the individual index of the MJST with

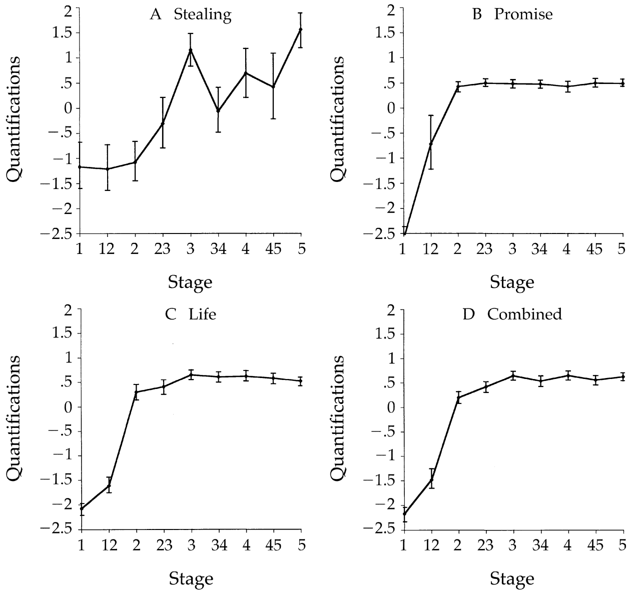

Figure 4 Grade 8 of Russian high school students. Means and 95% confidence intervals for quantifications of the rankings found by a bootstrapping procedure, for three sets (A–C), and for all sets combined (D). Intermediate stages, as used by Colby and Kohlberg (1987), are indicated by combining the adjacent stages' numbers. The values of the quantifications have no intrinsic meaning.

SROM-sf score was modest, $r(447) = .51$, $p = .000$. The average of the straightforward Pearson correlation between the ordering provided by Kohlberg and the ordering given by these participants was .51. The optimal quantifications of the rankings led to a maximized average correlation of .63 for this combined model.

Instead of the naturally found differences for the existing groups, the SROM-sf data can also be used to create developmentally different groups. Three samples were created: participants with a SROM-sf score between 225 and 275 (Stage 2/3: $n = 45$, average of 15.1 years), between 275 and 325 (Stage 3: $n = 234$, average of 19.2 years), and between 325 and 375 (Stage 3/4: $n = 163$, average of 22.9 years). Only four participants had an SROM-sf score below 225 and only one had a score above 375, so considering more groups made no sense.

The average correlations were .33 for the Stage 2/3 sample, .48 for the Stage 3 sample, and .63 for the Stage 3/4 sample. Group had a significant effect, $F(2, 439) = 53.97, p < .000$ (see Table 4). Post-hoc comparisons confirmed that the average correlations were significantly different (Bonferroni).

The optimized correlations were .54 for the Stage 2/3 sample, .62 for the Stage 3 sample, and .70 for the Stage 3/4 sample. The quantifications after optimizing the average correlation are shown in Figure 5.

Table 4 Subdivision of All Participants with Valid SROM-sf Score (stratified by SROM-sf score) and Comparison of Average Correlations

	SROM-sf Score			
	<225, >375 (n = 5)	225–275 (n = 45)	276–325 (n = 234)	326–375 (n = 163)
Straightforward correlation		.33	.48	.62
SD		.24	.20	.16
Maximized correlation		.54	.62	.70
Confidence intervals[a]		.469–.607	.605–.658	.694–.746

Note: SROM-sf = Sociomoral Reflection Objective Measure–Short Form.

[a] Confidence intervals based on 500 bootstraps.

A change in pattern was seen, with the higher-stage group extending the increase in quantification to the higher stages. Because the quantifications sum

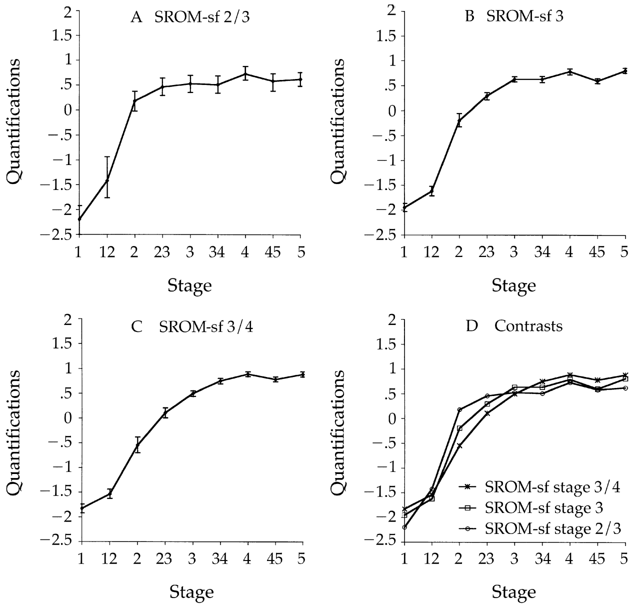

Figure 5 Subdivision of all participants with valid Sociomoral Reflection Objective Measure–Short Form (SROM-sf) score. (A–C) Three samples stratified by SROM-sf score. Means and 95% confidence intervals for quantifications of the rankings found by a bootstrapping procedure. (D) Contrasts of means between these three samples. Intermediate stages, as used by Colby and Kohlberg (1987), are indicated by combining the adjacent stages numbers. The values of the quantifications have no intrinsic meaning.

to zero, high quantifications must be compensated by low quantifications. Indeed, an interesting crossover was seen between the curves of the Stage 2/3 and Stage 3 groups and this crossover point was between items representing Stages 2/3 and 3. The crossover point for the Stage 3 and 3/4 groups was found precisely between items representing Stages 3 and 3/4.

Finally, all four samples were compared: the two cohorts of university students and the two grades of high school students (see Figure 6). The two university cohorts had almost the same pattern of quantifications (see Table 5).

The two high school samples also shared a similar type of pattern; however, the age difference was relatively larger here (see Table 5). Group had a significant effect, $F(3, 745) = 43.07$, $p < .000$, on the average MJST correlation for each of the four samples (see Table 5). Post-hoc comparisons confirmed that all four straightforward average correlations were significantly different (Bonferroni).

DISCUSSION

Empirical evidence was found for the hierarchical nature of moral judgment. Participants sorted statements representing moral stages as defined in Kohlberg's theory according to expectations. Confirming the first hypothesis, the best-fitting quantifications for the ranks of statements representing Kohlbergian stages displayed a clear smooth increase up to Stage 4 for the university student samples and up to Stage 3 for the high school student samples. From full Stage 4 to full Stage 5 for the university cohorts, and from Stage 3 to Stage 5—including intermediate stages—for the high school grades, there was no significant change. The maximized averages of the individual correlations for all samples were reasonably high (from .58 to .68). Overall, no alternative ordering described the data better than the order proposed by Colby and Kohlberg (1987).

Confirming the second hypothesis, the higher stage statements were more difficult to rank than the lower stage ones. The differences between the quantifications (distances) for the adjacent stages tended to diminish for the higher stages for both the university and high school samples, at least for the combined sets. Moreover, it was shown that this effect was dependent on age and SROM-sf score. The higher the SROM-sf score (or age), the more the increase extended to higher stage statements.

It is worth noting that for the combined sets, the in-between stages received quantifications precisely in-between the quantifications of the adjoining full stages, with only one exception. The only real deviation from the expected pattern was encountered for intermediate Stage 4/5. No explanation is offered for this phenomenon; however, the dip found does offer a reminder that the status of this intermediate Stage 4/5 has been a matter of discussion over the past decades (see Kohlberg, 1984, p. 429). Originally, a Stage 4½ was proposed as a transitional "relativistic" orientation explaining an unexpected dip in the developmental pattern. Subsequently, it was claimed that cleaning the content of statements by focusing on the formal structure, which took a complete revision of the scoring manual (Kohlberg, Levine, & Hewer, 1983), eliminated the need for such deviant developmental patterns. It is concluded that the use of the in-between stages posed no further problems (besides Stage 4/5), and in fact strengthened the case.

Although support was found for the ranking of the stages as proposed by Kohlberg, four qualifications

Table 5 All Four Samples Compared for Average Correlations

	High School		University	
	Grade 8 (n = 103)	Grade 10 (n = 93)	Cohort 99 (n = 278)	Cohort 98 (n = 275)
Straightforward correlation	.34	.42	.53	.59
SD	.20	.16	.22	.20
Maximized correlation	.58	.66	.62	.68
Confidence intervals[a]	.528 –.620	.618–.700	.590–.650	.661–.707

[a] Confidence intervals based on 500 bootstraps.

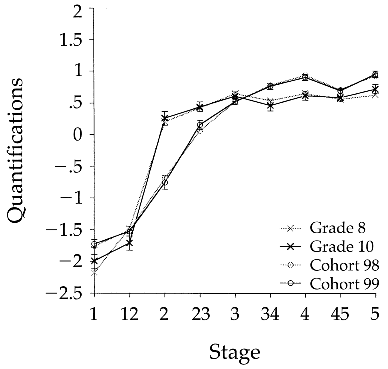

Figure 6 All four samples compared. For cohort 99 and grade 10, the means and 95% confidence intervals for quantifications of the rankings found by a bootstrapping procedure are shown. For cohort 98 and grade 8 only means are shown (see Figures 2 and 4 for confidence intervals).

are suggested by closer inspection. First, only 2 out of the 749 participants gave identical rankings over three sets. For each of the four samples, the averages of the straightforward individual correlations with the order proposed by Kohlberg were substantial but not at all perfect, and all had large standard deviations. Second, for the sample of high school students, the quantifications showed many more irregularities than for the university student sample. Although the data didn't suggest a clear alternative ordering of the stages, most items were not systematically distinguished by these high school students. Third, when sets of statements were analyzed separately, even stronger irregularities in ranking were obtained. Fourth, for the highest stages, there was only weak evidence for hierarchy.

With regard to the first point, the large individual variation combined with the strong effects found is striking. The very narrow bands for the confidence intervals of the means for the quantifications, and the fact that results were almost identical for comparable samples, points to high robustness. Nevertheless, a large amount of noise in the data on the individual level seems inevitable in research based on reactions to prototypical statements. How participants understand the statements will be influenced by many factors, not all relevant to moral development. Only large numbers of participants will reveal the looked-for effect of moral development.

With regard to the second point, it appears that for the less-educated and younger (and probably less-developed) sample of participants, consistent sorting was more difficult. The same is true for the less morally developed subgroups. Although the fit after optimization for the younger sample was quite good, at least for Sets 2 and 3, the associated pattern of quantifications makes clear that, from Stage 2/3 onward, the statements were effectively treated randomly. This led to nearly equal averages for each statement. These findings in fact support the postulated developmental effect, as discussed later.

With regard to the third point, for the university students, all sets showed a clear increase in quantifications up to Stage 4 for the full stages, but no further increase was found for the higher stages. The increases were all such that the confidence intervals were nonoverlapping. The intermediate stages showed some deviations from a smooth increase in each of the sets. In the high school sample, however, the pattern of quantifications for the first set deviated considerably from the hypothesized pattern. The average straightforward correlation with Kohlberg's ordering was rather low for this set, whereas the increase was very modest after optimization. Moreover, the confidence intervals were large. Therefore, it is concluded that the alternative ordering obtained by optimization, in this case, offered no real improvement over the standard ordering given by Kohlberg.

With regard to the fourth point, it is acknowledged that the evidence for a hierarchy in the higher stages in the data is weak (but see second hypothesis). A substantial sample of postconventional thinkers would not be expected in undergraduates and it may be difficult to find such a sample in natural circumstances (i.e., without selecting them first by using a written measure). Even in this sample, however, some progress beyond Stage 4 was noticeable, with the exception of Stage 4/5.

The detailed evidence for the second hypothesis is more varied. First, the distribution of the ranks assigned to each stage showed that for the lower stages the peaks in the distributions were as expected: for example, the Stage 1 statements were given rank 1 most frequently in all samples. The distributions were not single peaked for the higher stages, however. Second, the optimal quantifications of the rankings did not show a linear increase; the increase tended to diminish for the higher stages, for all four samples. Third, the correlation between SROM-sf score and i dividual correlations in the MJST was .51 overall. Participants with the higher SROM-sf score were better at ranking the statements according to the theoretically proposed ordering than participants with lower SROM-sf scores. Fourth, the quantifications for the samples stratified by SROM-sf score were looked at. It was assumed that the range of SROM-sf scores was a reasonable indicator of stage on group level. It was also assumed that the relatively large confidence intervals for the Stage 2/3 group were due to the smaller size of this group. Granted these assumptions, the fact that the increase in quantifications was large, or at least clear, up to their own stage (the stage of this group) but no further is interesting. The clearest differences were found around Stage 2 (and 2/3) with no overlap in the confidence intervals. This is consistent with the fact that the SROM-sf only uses items representing Stages 1 through 4. In the comparison of all four samples the same types of differences were seen, even more articulated, between the younger and less-educated samples and the older and more-educated samples.

Comparison with Previous Studies

The present investigation went beyond previous findings and included some technical improvements as well. In this study's sorting task, participants were asked to compare statements in terms of sophistication, whereas in most of the profile analyses discussed in the introduction, only importance ratings were used. Thus, the task directly appealed to the participants' notion of a hierarchy, whereas the profile analysis indirectly reconstructed participants' notion of a hierarchy. Differences with previous studies that used ratings obtained with the DIT follow from the fact that the DIT aims at assessing moral judgment in general (see Rest, Narvaez, Bebeau, & Thoma, 1999). Therefore, dilemmas and prototypical statements have to cover a broad range of topics and viewpoints (in Kohlberg's terminology: content). Moreover, the DIT was designed to have psychological reality so that only statements that were used frequently in pilot research (giving priority to psychological reality over designer counterbalancing) were included. In this study's sorting task, however, all statements have, as much as possible, the same type of reasoning, and all possible stages and intermediate stages are included, which maximizes the effect of the different structural logic's, and as such more directly addresses Kohlberg's hierarchical model. Compared with the few published studies that used score profiles (mostly of DIT ratings), many more participants and more categories (nine) were used in this study. Moreover, because the statements came directly from the definitive version of the scoring manual, they put Kohlberg's final theory directly to the test. In addition, a formal analysis was used and results were presented in a graphical form that conveyed more of the complexity found than a single index could.

Two other studies are closely related to this study with regard to the sorting procedure used. Rest (1973, p. 104) reported a ranking test with 47 participants who were requested to rank the six main stages as defined in Kohlberg's older theory. There were no clear positive results, no doubt because of small sample size. Walker et al. (1984) used paired comparisons of statements and found, as was found in this study, that participants valued the higher stage statements more. However, additional information was extracted from the sorting task in this study: the correlations pertaining to an individual profile, an index of how stages are separated, and an index of how well the Kohlbergian ordering fits the sample. Walker et al. (1984) determined each participant's own individual stage by using the MJI procedure and determined the highest level of understanding of statements according to criteria also derived from the MJI manual. They used the latter to create comparisons with and without sufficient understanding involved. Although interesting given this study's theoretical assumptions, Walker et al.'s example was not followed. Assessing own stage or highest stage of understanding in this way is a very demanding procedure that is difficult to replicate. More importantly, such an approach would presuppose what this study sought to question: the validity of the hierarchical sequence. One might object that this study's use of the different stage groups (as found by the SROM-sf) led to the same circularity; however, because the groups were used only to focus on the developmental effect and because a comparable shift in patterns was found for different educational levels also (see Figures 5 and 6), this objection does not hold.

Possible Objections and Alternative Explanations

Considerable differences between the different stories/sets were found. The best example is Item 5 in Figure 4A (representing Stage-3 reasoning). This item stacks out in a way not seen in the other figures. Notice that this was the only item in the MJST that appealed to authority and important persons. Perhaps this was a case of undesirable item-pull, but only so for the younger sample. Nevertheless, most of the items behaved as expected.

The more fundamental question is whether alternative explanations are possible (nonmoral or nondevelopmental) for the patterns found. For example, Shweder (1990) argued that justifications in moral reasoning tasks reflect accommodations to cultural knowledge (or social representations) rather than the developmental logic of justice. Emler and colleagues (Emler, Renwick, & Malone, 1983; Emler & Stace, 1999) have claimed that the differences between the higher Stages 4 and 5 do not reflect differences in development of the concept of justice, but rather reflect differences in political orientation (conservatives versus liberals). With the current samples, an alternative explanation concerning the higher stages as claimed by Emler and colleagues cannot be ruled out completely. The order proposed by Kohlberg, however, showed the most convincing fit of the data, including the dip for Stage 4½. The alternative viewpoints will have difficulty explaining the clear difference in pattern between the samples stratified for SROM-sf score and age.

Next, suppose that nonmoral ideas may have a *systematic* age/development effect that is almost indistinguishable from the expected effect based on moral development. Such systematic but nonmoral effects may be due to culture and context (e.g., type of content) or to language (e.g., verbal intelligence).

Attempts were made to minimize the effect of type of content. The classification of CJs in terms of norm, element, and issue was meant by Kohlberg to cover and neutralize unwanted content variation. In this study, norm and issue were held constant within sets of statements but there were differences in element. No systematic effect of differences in element on the quantifications was found, however.

Regarding verbal intelligence, a relationship between moral reasoning development and verbal intelligence is recognized within the Kohlbergian literature. More specifically, Kohlberg assumed that within the Piagetian framework a certain level of (verbal) intelligence is a prerequisite for a certain level of moral intelligence. As such, it sets a ceiling on the comprehension of moral statements. The Moral Judgment Sorting Task uses original statements from the Colby and Kohlberg manual. These statements may reflect verbal intelligence insofar as the statements in the Colby and Kohlberg manual reflect verbal intelligence; the statements in the manual also reflect verbal fluency insofar as verbal intelligence is a part of moral reasoning development. Using an interview method in a large normal population, the correlation was .30 at the highest. The statements used for the MJST in this study were controlled for verbal complexity and no significant correlation with stage order was found. It is concluded that the statements in the MJST are not corrupted more by verbal intelligence than the statements in the Colby and Kohlberg manual.

The highest relationship reported between a verbal intelligence measure and a sociomoral intelligence measure, as far as is known, can be found in Gibbs, Basinger, and Fuller (1992, p. 39). These researchers used a written productive measure of the Sociomoral Reflection Measure–Short Form (SRM-sf; range Stages 1–4). The correlation between the SRM-sf and the Verbal intelligence score (Quick Word Test) was .49. Nevertheless, these authors claim that their measure was sensitive for assessing moral reasoning development. Using the short-form interviews or comprehension written measures for assessing moral reasoning development constructed by Gibbs, the correlations reported between moral development and verbal intelligence within this specific population are in the range .30 to .40. Using Gibbs interview and comprehension measure in a population of normal high school students, correlations up to .30 might be expected. In sum, it seems unlikely that morally irrelevant aspects of the statements (e.g., linguistic features) influenced the results in a significant way, which is consistent with the fact that a nondecreasing trend was found in this study for the higher stages.

Limitations Concerning Individual Scoring

The evaluation of stage-prototypic statements is difficult to catch in one simple index. On the one hand, the results suggest that individuals should be characterized by a complete latent profile. Depending on the position one takes along the latent developmental continuum, the view on each of the possible stages does change in a nonlinear and asymmetric fashion. On the other hand, when the individual profile is based on just a few sets of statements—statements that are not interpreted by everyone in exactly the same way—an actual individual profile of ranks is not a very reliable indicator of the latent and presumably smoother profile of evaluation. First, although at the group-level the results were very robust, the low correlations with age within the samples was not expected. This could be due, however, to the homogeneous nature (both in terms of age and educational level) of the samples. Second, it is clear that the MJST individual score is measuring something similar to, but not exactly the same as, the SROM-sf. Note that identical MJST scores (the individual straightforward correlation with the Kohlberg ordering) can reflect different orderings of the statements. Perhaps other indexes might correlate better with the SROM-sf score and still other indexes of the MJST might correlate better with the P-index of the DIT; however, which MJST index (or combination of indexes) would be best for assessing moral development on the individual level is an empirical question not yet answered. In sum, more work on the individual index needs to be done.

CONCLUSION

The results lend plausibility to the interpretation of Kohlberg's model of moral development as a reconstruction of changes in the moral intuitions of subjects. The main conclusion is that the hierarchy of moral judgments is intuitively recognized by subjects at least up to their own current stage of thinking. Without, however, an older and more morally mature sample, it is not possible to draw stronger, more positive conclusions concerning the highest stages and the hierarchy. Nevertheless, the overall patterns found support the claim that the hierarchical structure of Kohlberg's stage theory of moral judgment is the corollary of Piagetian reflective abstraction.

ACKNOWLEDGMENTS

The authors thank Nico Brand, Harald Kunst, Geert Panhuysen, and all other faculty and students participating in DIPS2 for their cooperation in collecting the Dutch data. They thank Andrei Podolski and collaborators for collecting the Russian data. The authors are indebted to the Netherlands Organisation for Scientific Research (NWO) for its financial support of project 047.007.010.

ADDRESSES AND AFFILIATIONS

Corresponding author: Jan Boom, Universiteit Utrecht, Department of Developmental Psychology, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands; e-mail: j.boom@fss.uu.nl. Daniel Brugman and Peter G. M. van der Heijden are also at Universiteit Utrecht.

REFERENCES

Basinger, K. S., & Gibbs, J. C. (1987). Validation of the Sociomoral Reflection Objective Measure—Short Form. *Psychological Reports, 61*, 139–146.

Boom, J. (1991). Collective development and the learning paradox. *Human Development, 34*, 273–287.

Boom, J., & Molenaar, P. C. M. (1989). A developmental model of hierarchical stage structure in objective moral judgements. *Developmental Review, 9*, 133–145.

Brand, N., & Houx, P. J. (1992). MINDS: Toward a computerized test battery for health psychological and neuropsychological assessment. *Behavioral Research Methods, Instrumentation and Computers, 24*, 385–389.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment* (Vols. 1 & 2). Cambridge, MA: Cambridge University Press.

Davison, M. L., Robbins, S., & Swanson, D. B. (1978). Stage structure in objective moral judgments. *Developmental Psychology, 14*, 137–146.

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

Emler, N., & Stace, K. (1999). What does principled versus conventional moral reasoning convey to others about the politics and psychology of the reasoner? *European Journal of Psychology, 29*, 455–468.

Gibbs, J. C., Arnold, A. D., Morgan, R. L., Schwartz, E. S., Gavaghan, M. P., & Tappan, M. B. (1984). Construction and validation of a multiple-choice measure of moral reasoning. *Child Development, 55*, 527–536.

Gibbs, J. C., Basinger, K. S., & Fuller, D. (1992). *Moral maturity. Measuring the development of sociomoral reflection.* Hillsdale, NJ: Erlbaum.

Gibbs, J. C., Widaman, K. F., & Colby, A. (1982). Construction and validation of a simplified, group-administrable equivalent to the moral judgment interview. *Child Development, 53*, 895–910.

Gifi, A. (1990). *Non-linear multivariate analysis.* New York: Academic Press.

Habermas, J. (1984). Interpretative social science vs. hermeneuticism. In N. Haan, R. B. Bellah, P. Rabinow, & W. Sullivan (Eds.), *Social science as moral inquiry.* New York: Columbia University Press.

Habermas, J. (1990). Justice and solidarity: On the discussion concerning stage 6. In T. Wren (Ed.), *The moral domain: Essays in the ongoing discussion between philosophy and the social sciences* (pp. 224–254). Cambridge, MA: MIT press.

Hastie, T., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* London: Chapman and Hall.

Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics* (4th ed., Vol. 2). New York: Hafner.

Kohlberg, L. (1984). *Essays in moral development: Vol. 2. The psychology of moral development. The nature and validity of moral stages.* San Francisco: Harper & Row.

Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and a response to critics* [Contributions to Human Development, Vol. 10]. Basel: Karger.

Langford, P. E. (1997). Separating judicial from legislative reasoning in moral dilemma interviews. *Child Development, 68*, 1105–1116.

Piaget, J. (1960). The general problems of the psycho-biological development of the child. In J. M. Tanner & B. Inhelder (Eds.), *Discussions on child development* (pp. 3–27). London: Tavistock.

Piaget, J. (1977a). *Recherches sur l'abstraction réfléchissante: Vol. 1* [Inquiries into reflective abstraction: Vol. 1]. Paris: Presses Universitaires de France.

Piaget, J. (1977b). *Recherches sur l'abstraction réfléchissante: Vol. 2* [Inquiries into reflective abstraction: Vol. 2]. Paris: Presses Universitaires de France.

Puka, B. (1994). *Moral development: A compendium: Vol. 2. Fundamental research in moral development.* New York: Garland.

Rest, J. R. (1973). The hierarchical nature of moral judgment: A study of patterns of comprehension and preference of moral stages. *Journal of Personality, 41*, 86–109.

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

Rest, J. R., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking.* Mahwah, NJ: Erlbaum.

Rest, J. R., 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.

Rest, J. R., Thoma, S. J., Narvaez, D., & Bebeau, M. J. (1997). Alchemy and beyond: Indexing the Defining Issues Test. *Journal of Educational Psychology, 89*, 498–507.

Shweder, R. A. (1990). In defense of moral realism. *Child Development, 61*, 2060–2068.

Van Haaften, A. W., Korthals, M., & Wren, T. (Eds.). (1997). *Philosophy of development.* Dordrecht, The Netherlands: Kluwer Academic.

Walker, L. J. De Vries, B., & Bichard, S. L. (1984). The hierarchical nature of stages of moral development. *Developmental Psychology, 20*, 960–966.

Wren, T. (1990). *The moral domain: Essays in the ongoing discussion between philosophy and the social sciences.* Cambridge, MA: MIT Press.