International Journal of Behavioral Development
2002, 26 (2), 154-166
http://www.tandf.co.uk/journals/pp/01650254.html

DOI: 10.1080/01650250042000645

New tools, new insights: Kohlberg's moral judgement stages revisited

Theo Linda Dawson
University of California at Berkeley, USA

In this paper, four sets of data, collected by four different research teams over a period of 30 years are examined. Common item equating, which yielded correlations from .94 to .97 across datasets, was employed to justify pooling the data for a new analysis. Probabilistic conjoint measurement (Rasch analysis) was used to model the results. The detailed analysis of these pooled data confirms results reported in previous research about the ordered acquisition of moral stages and the relationship between moral stages and age, education, and **sex.** New findings include: (1) empirical evidence that transitions between "childhood" and "adult" stages of development involve similar mechanisms; *(2)* support for the notion of stages as qualitatively distinct modes of reasoning that display properties consistent with a notion of *structure d'ensemble;* and **(3)** evidence of a stage between Kohlberg's stages **3** and 4. Consistent with reports from earlier research, the relationship between age and moral development is curvilinear. The relationship between educational attainment and moral development is linear, suggesting that educational environments have an equivalent impact across the course of development. Older males have slightly higher scores than older females after age and education are taken into account (accounting for 0.3% of the variance in moral ability).

During the 1970s and 1980s researchers applied Piagetian principles to the study of reasoning outside the logicomathematical domain (for examples, see Armon, 1984; Kegan, 1982; Selman, 1980a). Much of this research was inspired by Kohlberg's seminal work (summarised in Colby & Kohlberg, 1987a) on the development of moral judgement. Although this research of Kohlberg and his colleagues generally supported: (1) the ordered acquisition of moral stages as defined in his sequence (Armon & Dawson, 1997; Nisan & Kohlberg, 1982; Snarey, Reimer, & Kohlberg, 1985; Walker, 1982); and (2) the absence of statistically significant reversals in the direction of development over time (Armon & Dawson, 1997; Nisan & Kohlberg, 1982; Snarey et al., 1985; Walker, 1982), postulates of *(3) structured* wholeness[1]—a global tendency for individuals to employ a single organisational structure to reasoning in the moral domain—and (4) *universality* were not as uniformly supported.

Ordered acquisition and a lack of reversals in moral development have been demonstrated employing both longitudinal and cross-sectional methods. The longitudinal evidence is compelling. The predicted sequence of stage acquisition with no stage-skipping and no statistically significant reversals were found in Kohlberg's original longitudinal study of New England schoolboys (Colby & Kohlberg, 1987a), in Walker's longitudinal study of Canadian children and their parents (1989), in Nisan's and Kohlberg's (1982) longitudinal study of city and country dwelling Turkish children, and in Snarey's longitudinal study of Israeli kibbutz residents (Snarey et al., 1985). In Armon's lifespan longitudinal study of middle class Americans (1984; Armon & Dawson, 1997) the only statistically significant reversal ($\frac{1}{2}$ stage) occurred in a 72-year-old respondent.

An additional, though weaker, source of evidence for the sequential acquisition of moral judgement stages is the relationship between moral stage and age. Age and moral stage are strongly correlated in childhood and adolescence. For example, Armon and Dawson (1997) report that through adolescence the relationship between age and moral stage is linear ($r = .88$). However, this relationship weakens in early and middle adulthood ($r = .61$)

Strong correlations between educational attainment and stage also provide support for the sequentiality of moral judgement stages. According to Kohlberg (1969), an important prerequisite of moral development is direct and repeated experience with moral conflict in social contexts. Formal education has been identified as a potential source of this kind of sociomoral experience, and several researchers have reported a moderate to strong positive relationship between educational attainment and stage of moral reasoning (e.g., Armon, 1984; Colby & Kohlberg, 1987a; Markoulis, 1989; Walker, 1986). The distribution of educational attainment by moral stage is linear and fan-shaped (Armon & Dawson, 1997), indicating that this relationship, like the relationship

[1] The terms "structured whole" and "*structure d'ensemble*" are used here to refer to continuity of reasoning within the moral domain. For a discussion of global versus domain-specific interpretations of *structure d'ensemble,* see Lourenco and Machado (1996), Smith (1993), **Vyuk** (1981).

Correspondence should be addressed to Dr Theo L. Dawson, University of California at Berkeley, Graduate School of Education, CD, Tolman Hall, Berkeley, CA 94720-1670, USA.

The author wishes to thank Larry Walker, Cheryl Armon, and Michael Commons for the use of their data. Thanks also to Ann Colby and the Murray Research Center at Radcliff College, for the use of Kohlberg's data. Appreciation is also due to Trevor Bond for introducing me to the Rasch model, to Mark Wilson for teaching me how to put it to work, and to Mark Wilson, Cheryl Armon, Karen Draney, W.P. Fisher, and three anonymous reviewers for their critical remarks on earlier drafts of this paper. The project was supported, in part, by a grant from the Spencer Foundation. The data presented, the statements made, and the views expressed are solely the responsibility of the author.

between age and stage, becomes less deterministic as the number of years of education increases. However, the relationship between educational attainment and moral stage can be described as linear rather than curvilinear, as is the case with age and moral stage.

The notion of *structured wholeness* (Piaget's *structure d'ensemble)* suffered when individual performances within and across the six issues in the Standard Issue Scoring Manual (SISM) (Colby & Kohlberg, 1987b) were frequently found to span more than two stages (Fischer & Bidell, 1998). Similarly, although cross-cultural studies generally supported invariant sequence and the absence of reversals (e.g., Nisan & Kohlberg, 1982; Snarey et al., 1985), claims of universality were comprised when notable differences across cultures were found in both conceptual content and highest stage attainment (Nisan & Kohlberg, 1982; Snarey et al., 1985). These cultural differences are particularly troubling in the light of two features of Kohlberg's method and theory: (1) the stages are partially defined in terms of particular philosophical content; and (2) each successive stage is considered not only to be more differentiated and integrated, but more philosophically adequate than any preceding stage (for a critique, see Puka, 1991). Gilligan's (1982) claim that men's moral reasoning is privileged over women's in Kohlberg's system, dealt a serious blow to cognitive developmental research in the moral domain, despite considerable evidence, including results presented here, that moral stage scores for women and men are distributed similarly once educational attainment has been taken into account (Armon & Dawson, 1997; Walker, 1984).

One originally unanticipated finding from moral development research employing the Kohlberg's instrument is that moral development continues into adulthood (Armon & Dawson, 1997; Colby & Kohlberg, 1987a; Nucci & Pascarella, 1987). In fact, an originally unanticipated finding from research employing Kohlberg's Standard Issue Scoring System **(SISS),** is that the highest stages of moral reasoning do not generally appear until well into adulthood. Two independently conducted longitudinal studies, Kohlberg's original 20-year study of approximately 60 males (Colby & Kohlberg, 1987a), and Armon's 12-year lifespan study of 43 respondents, ranging in age from *5* to 86 (Armon & Dawson, 1997), provide compelling evidence for "adult" moral reasoning stages. Adult forms of reasoning have also been identified in other howledge domains (Armon, 1984, 1993; Dawson, 1998; King & Kitchener, 1994). The highest measured stages of moral reasoning, stages 4 (consolidated formal operations) and *5* (post-formal operations), are rarely identified in the performances of individuals without some post-secondary education. Walker (1986), Markoulis (1989), and Armon (1984) found stage 4 reasoning only among adults who had obtained some college education, and in Armon's (1984) and Kohlberg and colleague's (Colby & Kohlberg, 1987a; Kohlberg & Higgins, 1984) studies, stage *5* performances were found only in individuals with at least some graduate work. Nucci and Pascarella (1987) report similar findings in their review of research on the relationship between college and the development of moral reasoning.

The discovery of "adult" stages raises the question of whether stage transitions during childhood are analytically and empirically analogous to stage transitions in adulthood. In other words, are adulthood stages, particularly, the "post-conventional" or "postformal" stage, 5, really stages? Although the present project does not address the analytical question (for this, see Commons, Trudeau, Stein, Richards, & Krause, 1998), the modelling methods employed here permit exploration of the empirical question by examining: (1) the unidimensionality of the latent trait, moral stage; and (2) the pattern of stage transitions along the moral development continuum.

The present project has been undertaken in an effort to readdress some of the issues outlined here by pooling and reanalysing the data from four Kohlbergian studies, Kohlberg's (Colby & Kohlberg, 1987a) study of schoolboys; Armon's (Armon & Dawson, 1997) lifespan study; Commons' (Commons et al., 1989a) study of MENSA members; and Walker's (1989) longitudinal study of schoolchildren and their parents. In a departure from meta-analytic techniques, I employ probabilistic conjoint measurement models (for an overview, see Kingma & Van den Boss, 1988), demonstrating that all four of these studies assess the same dimension of ability (moral stage) to an extent that justifies combining their data for further analysis. Then, using related psychometric techniques, these data are examined for evidence of invariant stage sequence, *structure d'ensemble,* unidimensionality, and education, age, and sex effects. Pooling the data not only increases the statistical power for analyses, but provides a lifespan dataset from a broad population with few age gaps. This makes the overall model of moral development presented here more compelling and lends additional credence to earlier evidence about the relationship of moral stage to age, education, and sex.

The intention here is to explore the extent to which results from studies employing Kohlberg's instrument support the postulates of his theory, and to re-examine relationships between moral reasoning stage and age, sex, and educational attainment. It is not an attempt to resurrect the Kohlbergian research enterprise. This examination reveals flaws in the SISS as well as strengths. The major difference between this analysis and meta-analysis is that here we return to the original data, employing sophisticated modelling tools that were unavailable when these studies were conducted. This makes it possible to look at the data from new and revealing perspectives.

Method

Data

The pooled dataset consists of 996 estimable cases, comprising 620 males and 376 females between the ages of 5 and 86 (M = 32, SD = 16). Educational attainment is between 0 and 21 years (M = 13, SD = **5).** Some educational attainment and age data are missing. Participants are predominantly Caucasian and middle class.

The data for all of these studies were collected and analysed according to criteria in the *Standard Issue Scoring Manual* (Colby & Kohlberg, 1987a, b). Within these guidelines, however, the method of data collection differed across studies. Original data for Kohlberg's, Armon's, and Walker's studies were predominantly from live, audiotaped, and transcribed interviews, whereas data for Commons' study were written. Kohlberg, Commons, Walker, and Armon supervised the scoring of all interviews from their respective projects. Participants in the Kohlberg, Commons, Walker, and Armon

Table 1
Age range, interview formats, and coders across four studies of the development of moral reasoning and evaluative reasoning about the good

	Age range of sample	*Form of Administration*	*Coder*
Armon (n = 147)	5–86	Live interview	Armon
Commons (n = 149)	18–83	Written	Armon
Walker (n = 472)	6–53	Live interview	Walker
Kohlberg (n = 196)	10–36	Live interview	Kohlberg

studies were New England schoolboys, adult MENSA members, Canadian churchgoers and their children, and a convenience sample of predominantly middle class Americans, respectively. The age range of participants also differed across studies. A summary of the similarities and differences in data collection is shown in Table 1.

An additional difference between studies is that Kohlberg's, Armon's, and Walker's are longitudinal while Commons' is not.[2] Kohlberg's sample was tested on six different occasions at 4-year intervals. Armon's sample was tested on four different occasions at 4-year intervals, and Walker's sample was tested on two different occasions at 2-year intervals. All of the analyses in this report are conducted on the pooled longitudinal and cross-sectional data. When test times are separated by relatively long intervals, problems with independence and sample-size overestimation that can be introduced with this practice are avoided (Willett, 1989). The ns reported above and in the remainder of this paper, unless otherwise indicated, include each respondent at each test time. In order to eliminate concerns about the possible introduction of error with this approach, all analyses were also run separately on the data for each test time. The trends found at each test time were consistent with the trends reported for the pooled sample, with no exceptions.

In all of the studies, subjects were scored for their stage of performance in up to six categories of moral judgement (also referred to as issues or items): (1) life; (2) law; (3) conscience; (4) punishment; (5) contract; and (6) authority.[3] The range of scores includes 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0, each of which represents a stage or half-stage in Kohlberg's scheme. Half-stage scores can come about in two ways: (1) they can represent a mix of performances at adjacent stages; or (2) they can be scored as transitional by employing criteria in the scoring manual. Some subjects received scores on only a subset of issues. Moral judgement interviews are structured around the judgements and justifications that are spontaneously generated by participants in response to moral dilemmas and a series of structured probe questions about life, law, conscience, punishment, contract, and authority issues as they relate to these questions. The content of any given interview may or may not address all of the moral issues, and probe questions vary somewhat, depending on the responses of participants. Because of this, and because there are no apparent patterns in the distribution of missing responses, absent responses are treated as missing at random.

Analyses

A procedure from psychometrics, called *common item equating* (Kelderman, 1986), makes it possible to examine whether an individual instrument performs similarly across studies. If the instrument functions consistently, data from multiple studies can be pooled and analysed in a common frame of reference. Fortunately, many developmental studies use the same instruments to assess developmental level. The body of research in which the development of moral judgement has been assessed with Kohlberg's Moral Judgment Interview (MJI; Colby & Kohlberg, 1987a,b) is a case in point.

At least four potential problems arise when data from several developmental studies are pooled into a single analysis. First, the samples may not be from the same population; second, raters may not score similarly enough; third, the instrument may not be administered in the same way; and fourth, different portions of an instrument may be used across studies, resulting in blocks of missing data. These problems are addressed by Rasch's models for measurement (Andrich, 1988; Rasch, 1980), most commonly applied in educational and psychological testing. These models can be used to evaluate sample and rater effects and are robust with respect to missing data, although measurement error is reduced and estimate precision enhanced by more complete data. A primary requirement of these methods, when applied to the context of pooling results across studies, is that all respondents (within and across samples) are tested on at least a subset of common items; thus the term, *"common item equating"*. In the case of the MJI, each respondent must have received a stage score on at least one of six moral issues.

Although they are well known in psychometric circles, Rasch's models for measurement have been employed by cognitive developmentalists only recently (Andrich & Styles, 1994; Bond, 1994; Bond & Bunting, 1995; Dawson, 1998, 2000; Draney, 1996; Hautamäki, 1989; Muller, Sokol, & Overton, 1999; Noelting, Coudé, & Rousseau, 1995; Wilson, 1989). One area of application for these models is the examination of behaviour on measures intended to capture hierarchies of difficulty, which makes them highly suitable for developmental applications. Rasch's models test the extent to which data meet the requirement that performances and items (or levels of items) form an invariant hierarchical sequence (within probabilistic constraints) along a single continuum (Andrich, 1989).

In their raw ordinal form, little can be said about the amount of difficulty associated with transitions between stage scores. However, when participants are ordered by the likelihood that they will perform at a given stage, the persons whose raw scores are high will be closer to the top of the developmental continuum, and the persons whose raw scores are lower will be closer to the bottom of the continuum. Rasch's models convert these likelihoods into distinct quantitative estimates of: (1) item difficulty; and (2) person ability,

[2] We are presently examining the longitudinal results of the combined data from Armon's and Kohlberg's studies with a hierarchical linear modelling approach.

[3] The method for obtaining these scores requires the calculation of a weighted average score from all performances on a particular moral issue in an interview. I have chosen to use these weighted average scores rather than the raw scores, because the latter are unavailable in some cases.

expressed in the same equal-interval metric, giving meaning to the distances between estimates. The common metric along which both stage difficulty and respondent ability estimates are arranged is referred to as a logit scale, in reference to the log-odds unit employed (Wright & Masters, 1982). In the analyses presented here, the mean item difficulty is set at 0. The logit range is from −7 to 8.

The distance between logits has a probabilistic meaning. In the present case, an ability estimate for a given individual means that the probability of that individual performing accurately on an item at the same level is 50%. There is a 73% probability that the same individual will perform accurately on an item whose difficulty estimate is one logit easier, an 88% probability that he/she will perform accurately on an item whose difficulty estimate is two logits easier, and a 95% probability that he/she will perform accurately on an item whose difficulty estimate is three logits easier. The same relationships apply, only in reverse, for items that are one, two, and three logits harder. (For more on Rasch's models, see Andrich, 1988; Masters, 1982.)

The logit estimates of item difficulty and person ability are but one of the statistics essential to measurement. Reliability and validity assessments require: (1) that item and person ability estimates be associated with an error term, which makes it possible to establish confidence intervals for all item and person ability estimates; and (2) one or more model fit statistics, so both items and persons can be examined for their conformity with the requirements of the model. Two types of fit statistics are included in the following analysis, outfit and infit. Fit statistics are used to assess whether a given performance (or item) is consistent with other performances (or items). They are based on the difference between observed and expected performances. Outfit statistics are based solely on the difference between observed and expected scores. In calculating infit statistics, however, extreme persons or items are downweighted. In most applications, the weighted infit statistics are more useful for assessing fit, because they are not affected by outliers. Infits (or outfits) near 1 are desirable. *t*-Values are calculated to assess the significance of both positive and negative divergences from 1. Interpretation of fit statistics is demonstrated below, in the results of the analysis.

The *partial credit model* (Masters, 1982, 1994), designed for items with more than two hierarchical categories, is employed here. Analyses were conducted with the computer program, Quest (Adams & Khoo, 1993). In keeping with the original formulation of the Rasch model, Quest treats person parameters as fixed effects. It has been argued that this limitation of the model restricts the generalisability of the results of Rasch analyses (Bartholomew & Knott, 1999), although the specific implications for research of the present kind are not entirely clear due to an apparent lack of published scholarly debate on this issue. Moreover, several researchers employ Quest and other software that treats person parameters as fixed effects to explore developmental constructs similar to those examined here (e.g., Bergan, 1988; Muller et al., 1999). In any case, concerns about generalisability are minimised in the present project by the large size of the dataset and its heterogeneity (Canadian Christians, boys from New England private schools, MENSA members, and a convenience sample from all over the country), combined with the fact that separate analyses of the four original datasets produced results that were highly consistent with one another.

Results

In order to determine whether the SISM functions similarly in all four studies, each dataset is first modelled individually, and the moral stage-item difficulty estimates are correlated. Subsequently, the data from all four studies are pooled, and modelled with a single partial credit analysis. Patterns of performance are analysed in terms of Kohlberg's stage theory, and relationships between moral judgement stage and gender, educational attainment, and age are examined.

Individual analyses

Individual partial credit analyses of the data from each original study were conducted in order to determine whether patterns of performance across the four studies were similar enough to warrant pooling the data for a single analysis. Results from the individual analyses were similar in two ways. First, the patterns of both stage-item difficulties and person ability estimates for the individual analyses were similar to one another. Consequently, they were also very similar to patterns in the overall model of the pooled data (presented below). Second, the correlations among the stage-item difficulties for the four individual analyses were very high. Stage-item difficulty estimates for each stage of each of the six moral issues were calculated and compared across the four studies. Despite differences in the samples, data collection, and raters, the stage-item difficulty estimates were strongly correlated (rs = .94–.98), as shown in Table 2. Correlations of this magnitude are a strong indication that the SISS functioned similarly enough across these studies to warrant pooling their data into a single analysis.

Pooled analysis

Item analysis. The infit and outfit statistics for all of the stage-item difficulty estimates were considered to fit the model if *t*-scores were smaller than 2.0. Table 3 shows the fit statistics and standard errors for each of the stage-item difficulty estimates in the analysis. All of the infit *t*s and outfit *t*s are well below 2.0. In fact, most are negative. Note, however, that the infit *t*s for the law and punishment issues are less than −2.0. There is less random variation in performances on these items than expected by the model. This is referred to as overfit. It means that individuals who have an estimated person ability higher or lower than the difficulty of a given level of an item—say, for example, level 3—are very unlikely to have been awarded a score at that level of the item. In this particular analysis, this overfit reflects a pattern of performance that is consistent with the notion that within a given domain, reasoning forms a *structure d'ensemble.* For the law and punishment items, individuals with person ability estimates

Table 2
Correlations among stage-item difficulty estimates for four moral development studies

	Armon	*Walker*	*Kohlberg*
Commons	.9429	.9696	.9482
Armon		.9824	.9816
Walker			.9830

Table *3*
Fit statistics for stage estimates (n = 996)

			Stage thresholds (standard errors below)											
Name	*Score*	*Max.*	1.5	2	2.5	3	3.5	4	4.5	5	*Injit* (MS_e)	*Outfit* (MS_e)	*Injit* (t)	*Outfit* (t)
1. **Life**	4274	7352	−7.31	−5.70	−2.54	−1.47	0.90	2.90	4.73	6.59	0.92	0.92	−1.7	−1.3
			1.03	0.62	0.24	0.22	0.16	0.17	0.23	0.33				
2. Law	4068	6984	−5.13	−3.84	−1.88	−0.84	0.75	2.28	4.83	6.76	0.89	0.90	−2.2	−1.7
			0.41	0.32	0.24	0.22	0.14	0.17	0.22	0.36				
3. **Conscience**	3765	6368	−5.88	−5.13	−2.51	−1.36	0.98	2.62	5.02	6.58	0.95	0.94	−1.1	−0.9
			0.70	0.58	0.26	0.24	0.18	0.16	0.25	0.36				
4. **Punishment**	4042	5908	−4.56	−3.41	−2.14	−1.17	0.10	2.23	5.79		0.81	0.84	−3.7	−2.6
			0.34	0.31	0.26	0.25	0.20	0.14	0.28					
5. **Contract**	4279	7336	−5.75	−5.05	3.11	−1.18	0.90	2.62	5.65	6.79	0.99	1.00	−0.1	0.0
			0.59	0.51	0.29	0.20	0.15	0.17	0.32	0.42				
6. **Authority**	3391	5672	−4.69	−3.89	−2.80	−1.83	0.70	2.87	4.91	6.10	1.01	1.01	0.3	0.1
			0.44	0.39	0.31	0.26	0.19	0.19	0.27	0.32				
Mean			*0.00*								0.93	0.94	−1.4	−1.1
SD			*0.30*								0.07	0.06	1.5	1.1

that reflect a high probability of performance at a given stage, say stage 3, are very likely to have been awarded a stage 3 score on the law and punishment issues, whereas individuals with person ability estimates that reflect a high probability of performance at stages **1,** 2, 4, or **5** are very unlikely to have been awarded a stage 3 score on the law and punishment issues. In a sense, from the perspective of the model, the pattern of performance on these items is "too good to be true". However, from the perspective of stage theory, this is an expected pattern of performance.

The map of person ability estimates and stage-item difficulty estimates shown in Figure 1 provides further information about performance on items. On the far left of the figure is the logit scale. It spans **−7.0** to +8.0 logits. To the right of the logit scale are the person ability estimates, each of which is represented by a **I**, 0, or **X.** To the far right are the stage-item difficulty estimates (Thurstone thresholds, Masters, 1982), each of which is labelled with its issue and stage. Wide, pale-grey bands highlight estimates for full stages 2, 3, 4, and 5. Note that the item difficulties for each stage or half-stage tend to cluster together at around the same ability level, with some overlap between 2.5 and 3.0, and a great deal of overlap between 1.5 and 2.0. When 95% confidence intervals for each of the stage-item difficulty estimates are calculated from the standard errors shown in Table *3,* areas in which there is no overlap of confidence intervals appear at the 3.0/3.5, 3.5/4.0, and 4.0/4.5 transitions. These are represented with narrow grey bands. There are no similar gaps between 1.5 and 2.0, 2.0 and 2.5, **2.5** and **3.0,** and 4.5 and 5.0.

Gaps between confidence intervals of groups of stage-item difficulty estimates occur when individual performances are highly consistent. In this case, the gaps reflect the fact that a large percentage of individual performances are predominantly at a single stage across all six issues. This type of pattern is expected when learning involves the qualitative restructuring of knowledge rather than the simple additive accumulation of knowledge (Fischer & Bidell, 1998; Wilson, 1985). Similar evidence that stages are qualitatively distinct modes of reason-

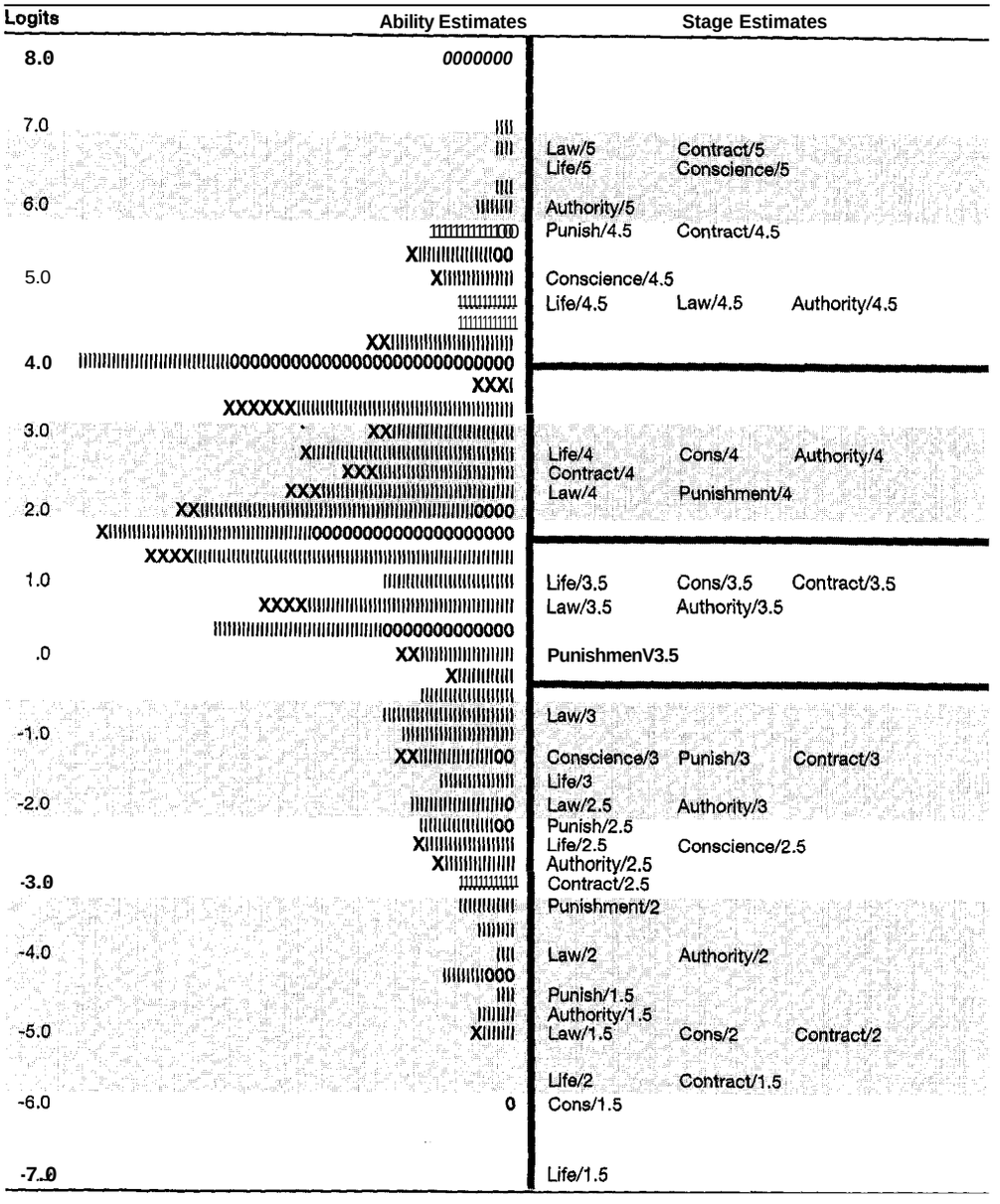

Figure 1. Map **of** person ability estimates and stage-item difficulty estimates *(n* = 996). Each I, **X,** or *0* = one case.

ing has been presented elsewhere (Dawson, 1998; Draney, 1996; Fischer, Hand, & Russel, 1984; Fischer & Kennedy, 1997; Hartelman, van der Maas, & Molenaar, 1998; Wilson, 1985).

In the present analysis, the distribution of stage-item difficulty estimates is complex. If Kohlberg's formulation of the stages is correct, a delay in development that would lead to gaps is expected following the consolidation of thinking at a given stage, and prior to any reorganisation at the following stage. Thus, we would expect to see gaps between full stage-item difficulty estimates and subsequent half stage-item difficulty estimates (the 2.0/2.5, 3.0/3.5, 4.0/4.5 transitions). Once new structures are available, it is expected that they will be relatively rapidly employed to restructure a range of knowledge, which means that we would expect smooth transitions, perhaps even some overlap of estimates, at 1.5/2.0, 2.513.0, 3.5/4.0, or 4.5/5.0. Such a pattern of smooth transitions and gaps is supportive of the cognitive-developmental notion of structured wholeness—that, at least within a given domain, reasoning should ''consolidate" at one stage before advancing to the subsequent stage (Kohlberg, 1969).

Although apparent between stage 3.0 and half-stage 3.5, and stage 4.0, and half-stage 4.5, statistically significant gaps are not seen at the 2.0/2.5 transition. The lack of a gap at the 2.0/2.5 transition may be due to any one (or a mixture) of four factors: (1) the smaller sample size in the 2.0/2.5 range; (2) a less reliable definition of the stages at this level; (3) more rater error at this level; or (4) less consistent reasoning at this level. Although the sample size is considerably smaller in this range than in the higher stage ranges, it should be noted that analyses of quite small samples sizes (140–200 cases) produce the same pattern seen here, with clear gaps at the higher stages, and no gaps at the lower stages—even when the number of respondents at the higher stages is fewer than the number of respondents in the present sample who are performing at the lower stages (for an example, see Dawson, 2000).

To determine whether patterns of performance appear less consistent at lower stages, the relationship between the range of stages represented in individual performances and ability estimates was examined. **A** hierarchical ANOVA revealed that the range of raw stage scores (from 0 to 2.5), increases somewhat as ability estimates decrease: $F(5, 984) = 7.294$, $p = .01$, $r = .19$. Although the effect size is small, this apparent decrease in consistency within individual performances may account, in part, for the overlap in stage-item difficulty estimates at the 2.0/2.5 transition. The reason for this trend is not clear, however.

In addition to the unexplained overlap in stage-item difficulty estimates at the 2.0/2.5 transition, there is a significant, unanticipated, gap at the 3.5/4.0 transition. This gap suggests that the transition from half-stage ***3.5*** to stage 4.0 is a move from one full stage to another, even though it is characterised in Kohlberg's model as a move from a transitional level to a full stage. Both Commons and his colleagues (Commons, Richards, with Ruf, Armstrong-Roche, & Bretzius, 1983; Commons et al., 1998) and Fischer et al. (1984) have proposed that there are two stages (abstract and formal), rather than one (Kohlberg's stage 3.0) between concrete operations (Kohlberg's stage 2.0) and systematic operations (Kohlberg's stage 4.0). In this formulation, Kohlberg's stage 3.0 is considered abstract or early *formal,* and his transitional level 3.5 is considered *formal.* The model presented in Figure 1 lends support to Commons' and Fischer's assertions.

If Kohlberg's half-stage stage 3.5 is accepted as a full stage, the pattern of stage-item difficulty estimates from stage 3.0 to stage 5.0 is remarkably consistent. Transitions from one full stage to another are marked by statistically significant gaps between stage-item difficulty estimates. Although this is not incontrovertible evidence that the transitions between both "adult" and "childhood" stages represent the same kind of qualitative change, it is, at the least, consistent with this thesis.

Person analysis. The overall person separation reliability for 126 nonextreme cases—cases with perfect scores and zero scores are not included in the estimation—is .93. The person separation reliability statistic is equivalent to Cronbach's alpha, and is based on the ratio of the variation in the mean squares (the standard deviation) to the error of measurement, also known as a signal/noise ratio (Wright & Masters, 1982). In this instance, a person separation reliability of .93 means that persons whose ability estimates are at a given stage can reliably be differentiated from persons whose ability estimates are as close as an adjacent stage. Standard errors for the person ability estimates range from 0.49 to 1.75 logits with a mean of 0.64.

The infit and outfit statistics for all person ability estimates were considered to fit the model if t-scores were greater than −2.0 or less than 2.0. Fit statistics lower than −2.0 indicate a greater than expected consistency within performances (overfit), whereas fit statistics higher than 2.0 indicate less consistency than expected (underfit). Both underfit and overfit are types of misfit, but are distinct in their implications,

In Figure 1, each case is represented by an **|**, **X,** or ***0.*** Performances that overfit the model are indicated with ***0.*** These performances are more consistent across issues than expected by the model. Seventy-eight of 119 performances with all issue scores at a single stage exhibit overfit. Forty-one of 95 cases with performances that spanned $1\frac{1}{2}$ or more stages exhibit underfit, because they are less consistent across issues than expected by the model. These are indicated with **X.**

Because Rasch models are probabilistic, a certain amount of ''noise'' or random variation is expected in the data. When the expected variation is not present, as is the case when many individuals perform at a single stage across all issues, these performances will overfit the specifications of the model.[4] However, performances of this kind are not problematic for stage theory, which expects a high level of consistency in the stage of reasoning exhibited by an individual in a given domain (Kohlberg, 1969). More problematic for stage theory are performances that span a wide range of stages—those that underfit the model. When misfit of this kind occurs, it is desirable to re-examine the original data to determine if coding errors were made or if there is evidence that these performances genuinely do not fit the expected pattern of response.

[4] Rasch's probabilistic models expect ability estimates to be more continuously distributed than they are in the present sample. The jagged, "toothy", quality of the ability distribution shown in Figure 1, accompanied as it is by a high degree of overfit, is a violation of the modelled measurement requirements. The fact that a pattern of performance that *is* in keeping with cognitive developmental theory shows up as a significant amount of overfit in a partial credit model points to a discontinuity between the model and both developmental theory and actual patterns of performance. This phenomenon has been observed elsewhere, and a model, which extends the Rasch model, has been developed to encompass the phenomenon (Draney, 1996; Wilson, 1989). Though promising, this model has not yet been formulated for the type of scored interview data employed here.

Unfortunately, the original interviews were not available for analysis, ***so*** this kind of evaluation was not possible.

The concentration of person ability estimates at the 4.0, 2.0, and 0.0 logit ranges, along with the general trend toward model overfit, indicate large subgroups of individuals who have a high probability of performing across all issues at stage 4.0, half-stage 3.5, or 3.0, respectively. For example, an individual whose ability estimate is 4.0 logits has a greater than 73% probability of performing at the stage 4 level on all moral issues, and less than a 27% probability of performing at the half-stage 4.5 level.[5]

Age, education, and sex effects

Correlations between moral reasoning ability and the age, educational attainment, and gender of participants are shown in Table 4.

Age. To further examine age, education, and sex effects, several multiple regression analyses were conducted. First, the relationship between moral ability estimates and age is examined. A logarithmic model provides the best fit, revealing a strong relationship between age and moral reasoning ability:

$$R = .75, F(1, 964) = 1244.06, p < .01,$$
$$\text{Moral ability estimate} = -9.69 + 7.64_{\text{logage}}.$$

In order to assess whether some stages in this model should be considered "adult" stages, the relationship between age and stage is examined in Table **5.** Stage assignment for this table was based on moral ability estimates as follows: stage 5.0 = *6.01 through 8.00,* stage 4.5 = *4.01 through 6.00,* stage 4.0 = 2.26 *through 4.00,* stage 3.5 = *0.01 through 2.25,* stage 3.0 = *−1.74 through 0,* stage 2.5 = *−2.99 through −1.75,* stage 2.0 = *−4.49 through −3.00,* stage 1.5 = *−7.00 through −4.50.* The minimum age at which any individual in this sample has at least a 50% probability of performing at stage 5.0 on any of the 6 moral issues is 25 [only 2 individuals below age 30 (10%) were in this group], with a mean age of 44, and although two individuals below age 21 (2%) had a 50% probability of performing at transition 4.5, the mean age at this level is 42. Only 3 individuals below age 21 (1.2%) had a 50% probability of performing at stage 4.0. Given that the minimum ages in this table can be said to represent minimum ages of acquisition, the results of this analysis support previous reports that stages 4.0, and 5.0, and transition 4.5 appear to occur rarely before adulthood.

Although there are no differences between males and females when sex and moral ability estimates are correlated directly, when sex is entered into a regression of moral ability estimates by the log of age, the curves for males and females are significantly different, as shown in Figure 2. Overall, males appear to perform at slightly higher levels than females, explaining about 1% of the variance in ability estimates. (In order to make the relationship between stage attainment and the ability estimates clearer, wide, horizontal, grey bands are included in Figures 2 and 3. These represent the approximate ranges for performances at Kohlbergian stages 2.0, 3.0, 4.0, and 5.0, as labelled on the right of each figure.) The multiple regression of the log of age and sex on the person ability estimates results in the following equation:

$$R = .76, F(2, 963) = 647.19, p < .01,$$
$$\text{Moral ability estimate} =$$
$$-9.63 + 7.74_{\text{logage}} - .53_{\text{sex}}\, t_{\text{logage}} = 35.96,$$
$$p < .01, t_{\text{sex}} = -4.75, p < .01.$$

The relationship represented in the above equation is complex. Table 6 shows the distribution of moral stage-item difficulty estimates by age and gender. (For a sense of where these standardised estimates fall on the stage continuum, consult Figure 2. Note that the difference in terms of actual stages are never more than $\frac{1}{4}$ of a stage.) The mean moral ability **(MAE)** estimates for males and females in each age group are shown on the right. For each age group, the estimates for the sex with the higher mean estimate are shown

Table 4
Correlations between moral reasoning ability and education, sex, and age

Education	*Age*	*Sex*
.7948	.6593	−.0212
(n = 929)	(n = 966)	(n = 987)
$p < .01$	$p < .01$	$p > .51$

Table 5
Stage attainment by age

Stage	*Valid cases*	*Min. age*	*Max. age*	*Mean*
5.0	19	25	66	44
4.5	99	17	83	42
4.0	244	18	86	40
3.5	350	13	72	35
3.0	120	8	58	19
2.5	65	7	18	12
2.0	49	6	17	10
1.5	19	5	14	8

Table 6
*Moral ability estimates **(MAE)** by age and sex*

	Sex	
Age group	*Male (Mean **MAE**)*	*Female (Mean MAE)*
5–9	**−3.32**	−4.03
10–14	−2.11	**−1.76**
15–19	0.08	**0.16**
20–24	1.22	**2.17**
25–29	2.05	**2.83**
30–34	**3.02**	2.67
35–39	**2.86**	1.91
4044	**2.17**	2.12
45–59	**3.17**	2.42
50–54	**3.25**	1.87
55–59	**3.53**	2.69
60–64	**4.07**	3.11
65–69	**3.28**	2.55
70–86	3.14	**3.30**

[5] Gibbs, Basinger, and Fuller (1992) report a similar finding employing their Sociomoral Reflection Instrument.

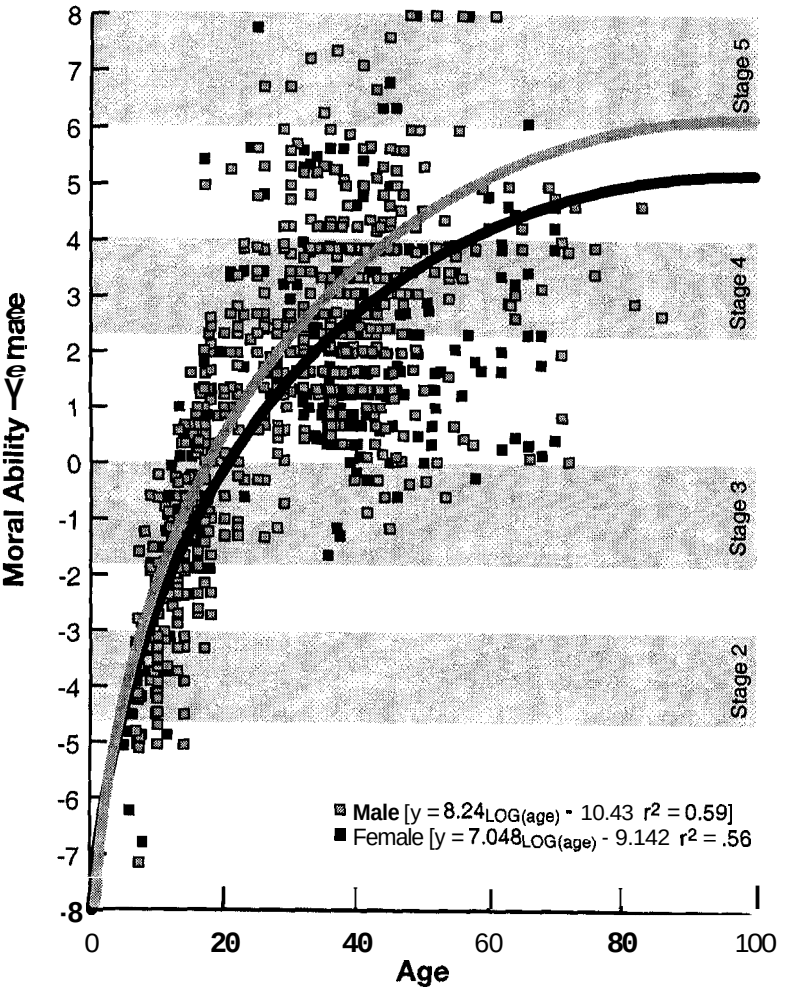

Figure 2. Regression of moral ability estimates with the log of age by sex (female = 376; male = 620).

in bold. Although the males appear to have an advantage between ages 5–9 and 30–69, the females have the advantage from ages 10 to 29 and 70 to 86. One possible explanation for this complex pattern is cohort differences. It is plausible that older women did not have the same educational and lifesryle advantages afforded to men in their age cohort, whereas social change resulting from the women's movement of the 1960s and 1970s may have provided women in the younger cohort with more of these opportunities.

Educational attainment. Next, the relationship between ability estimates and educational attainment is examined. The multiple regression of educational attainment on ability estimates results in the following equation, in which individuals advance, on average, about $\frac{1}{2}$ stage for every four years of formal education:

$$R = .79, \mathrm{F}(1, 927) = 1590.12, p < .01,$$
$$\text{Moral ability estimate} = -4.33 + .42_{\mathrm{ed}}.$$

A scatterplot of this regression, with moral ability on the *y*-axis and educational attainment on the x-axis, shows a linear, but fan-shaped distribution of estimates is shown in Figure 3. The range of moral ability estimates increases with advances in educational attainment, indicating that the relationship between educational attainment and moral development weakens as years of educational attainment increase, though the overall slope appears to remain relatively constant. To examine this relationship further, a quadratic component was added to the regression to examine whether the effect of educational attainment declines as educational attainment increases. Although the quadratic component made a statistically significant contribution: $F(2, 926) = 839, p < .01$, it explained only an additional 1% of the variance in person ability estimates.

As shown in Table 7, in this sample, the minimum number of years of education required to achieve a 50% probability of performing at stage 5 on any issue was 15, or three years of post-secondary education. Only one person without a bachelor's degree *(5%)* performed at this level. Although the minimum number of years required to achieve a 50% probability of performing at the 4.5 level on any issue was

Table 7
Stage by educational attainment

Stage	*Valid cases*	*Min. ed.*	*Max. ed.*	*Mean*
5.0	18	15	21	19
4.5	96	11	21	17
4.0	232	9	21	17
3.5	334	7	21	14
3.0	116	1	19	10
2.5	65	2	18	7
2.0	48	1	15	4
1.5	19	1	9	3

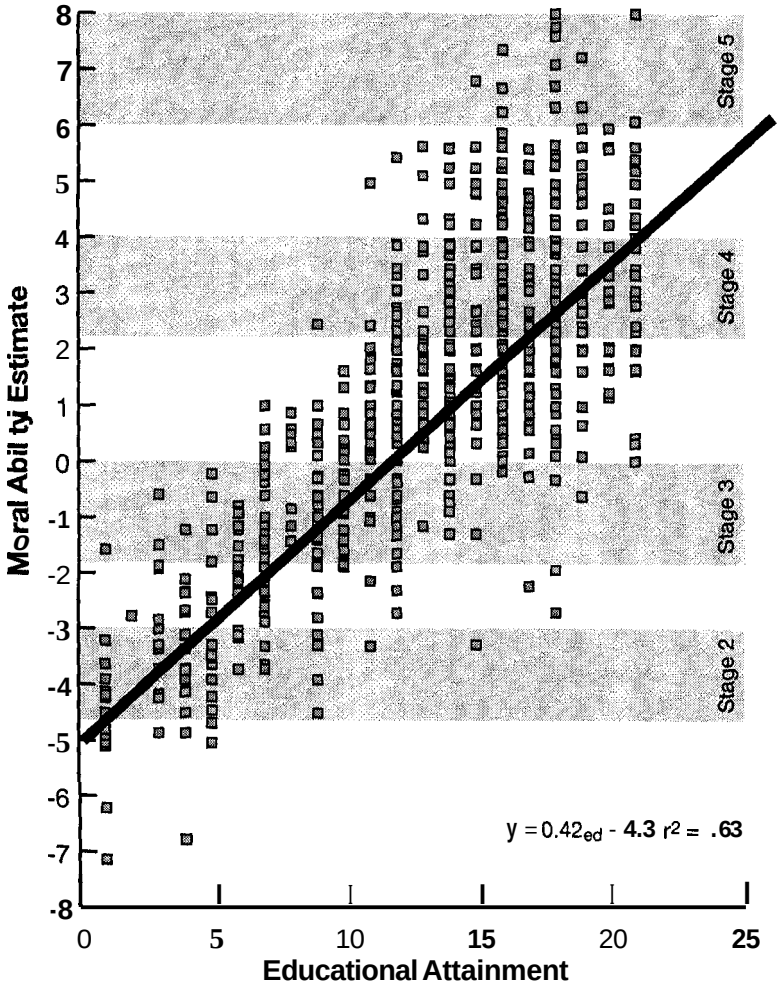

Figure 3. Regression of moral ability estimates with educational attainment ($n = 996$).

11, only two individuals with less than one year of college education (2%) performed at this level. Similarly, although the minimum number of years required to achieve a 50% probability of performing at the 4.0 level on any issue was 9, only two individuals with less than a high school diploma (1%) performed at this level.

Sex, age, and educational attainment. Adding sex to the regression of educational attainment on the moral ability estimates does not explain any additional variance. However, sex explains about **0.3%** of the variance when entered into a stepwise regression of moral ability estimate with education and age:

$$R = .81,\ F(3,\ 922) = 604.88,\ p < .01,$$
$$\text{Moral ability estimate} =$$
$$-6.97 + .28_{ed} + 3.22_{logage} - .28_{sex}\ t_{ed} = 14.86,$$
$$p < .01,\ t_{logage} = 9.02,\ p < .01,\ t_{sex} = -2.74,\ p < .01.$$

Clearly, education accounts for most of the variance (63%) in moral ability estimates. The log of age adds an additional *3%,* whereas sex contributes less than 0.3%. The reduction in the effect for sex, after education is taken into account, lends support to the argument that most, if not all of the sex difference in moral ability estimates is due to cohort effects rather than systematic biases in the scoring system or theoretical perspective.

Discussion

The analyses presented here brought together four sets of data, collected by four different research teams over a period of 30 years. The samples included a group of parents and their children, a diverse life-span sample, a group of MENSA members, and a group of private-school boys. Four groups of rates scored the data for stage using Kohlberg's Standard Issue Scoring Manual (SISM).

All of these differences between the datasets would interfere with attempts at comparison using conventional analytical methods. At best, a meta-analysis could be conducted, comparing statistical results from one sample to another, but there would be no way to assess just what was being compared. Rater agreement and consistency would have to be assumed, despite the fact that differences in interpretation and interview methods could easily vary in ways that would influence outcomes.

Exploring the datasets for fit to a probabilistic measurement model provided a basis for comparing results from these three studies. Despite their independent samples and execution, the stage scoring across the studies was congruent enough to result in very high correlations between stage-item difficulty estimates (.94–.98). Pooling the four datasets employed here was easily justified by these correlations.

The detailed analysis of these pooled data resulted in

interesting evidence that confirms results reported in previous research about the ordered acquisition of moral stages and the relationship between moral stages and age, education, and sex. This analysis also provides new support for the notion of stages as qualitatively distinct modes of reasoning that display properties consistent with a notion of *structure d'ensemble,* and reveals evidence of a stage, between Kohlberg's stages 3 and 4, that has not previously been revealed in analyses of Kohlbergian data.

Moral development, as assessed by the SISM, is strongly related to both educational attainment and age. In keeping with findings from previous research, the relationship between age and moral development is curvilinear and fan-shaped, as shown in Figure 2. The relationship between educational attainment and moral development is linear, rather than curvilinear, suggesting that educational environments have an equivalent impact across the course of development. However, this relationship, too, is fan-shaped, suggesting that as we age the impact of education becomes more variable.

The analysis of the relationship of age and moral development also supports previous evidence that the higher stages of moral development are appropriately labelled "adult stages". Moreover, the model of development presented in Figure 1 suggests that these stages represent the same kind of qualitative shifts in modes of reasoning that take place at stages that predominate in childhood and adolescence. This adds support to an increasing body of evidence that characterisations of adulthood as a period of decline in mental abilities are narrow, if not incorrect. Notions of adult stages in particular, and adult development in general, raise interesting questions. First, how can adult stages be reconciled with theories that link stage change with childhood biological changes (e.g., Epstein, 1990)? And in a different vein, is it possible that some of the changes in cognition previously viewed as declines, such as evidence pointing to the "crystallisation" of intelligence, are better viewed as symptoms of higher order functioning? These and other issues are being explored in an increasing body of research into positive adult development (for examples, see Alexander & Langer, 1990; Commons et al., 1989b; Kohlberg & Higgins, 1984; Sinnott & Cavanaugh, 1991).

Only a weak relationship, accounting for less than 0.3% of the variance, was found between sex and stage after age and education were taken into account, with older males performing at slightly higher levels than older females. Walker (1984), in his meta-analysis of 79 studies of sex differences in moral reasoning development, found inconsistent evidence of differences in childhood, with males doing better in some studies and females doing better in others. Adult differences were more consistent, with males apparently doing better than females, but this effect disappeared when educational attainment was taken into account. The present analysis suggests that some effect of sex on moral ability remains after taking both education and age into account, but the effect is very small and nonsystematic, in that males and females appear to have the advantage at different ages. Gilligan (1982) challenges the universality of Kohlberg's moral stages on the basis of the assumption that males and females perform differently on the MJI. The preponderance of evidence strongly suggests otherwise.

The map of development in Figure 1 provides evidence of gaps between full-stages that support both the concept of an invariant hierarchical sequence in stage development, and the notion of stages as qualitatively distinct modes of reasoning that display properties consistent with a notion of *structure d'ensemble,* an idea that has been much debated in the literature (see, for example, Bidell & Fischer, 1992; Demetriou, Efklides, Papadaki, Papantoniou, & Economou, 1993; Kohlberg & Higgins, 1984; Turiel & Davidson, 1986). The gaps at the 3.0/3.5 and 4.0/4.5 transitions indicates that reasoning tends to consolidate at a given stage before progressing to the next stage. From stage 3 onward, individual stage-item difficulty estimates across life, law, conscience, punishment, contract, and authority issues tend to cluster within narrow ranges, about one logit in width, with statistically significant gaps between groups of full-stage and subsequent half-stage-item difficulty estimates. Keeping in mind that here we are looking at reasoning within a narrowly defined domain, this pattern supports the notion of stages as *structured **wholes,*** coherent systems of thought that tend toward consolidation at a given order of complexity until conditions are such that movement to the next order of complexity is possible. The absence of a gap between the estimates at the 2.0/2.5 transition violates this pattern. Further research must be conducted to determine whether this is the result of measurement error or differences in the nature of moral development at this level.

The distribution of stage-item difficulty estimates in clumps along the moral ability scale is supportive of the notion that stages represent qualitatively distinct modes of reasoning. Although stage-item difficulty estimates occur in clumps, person ability estimates can fall at any point on the ability scale. The fact that person ability estimates can fall at any point along the scale could be taken to support a cumulative model of learning. However, the pattern of these estimates is not smooth, as might be expected if learning can best be described as a cumulative rather than transformative process. Instead, the distribution of person ability estimates is "toothy". Though a given individual can perform at any point on the developmental continuum, more individuals are clustered at points where consolidated performances are likely than at points where mixed performances are likely. This distribution suggests that learning is not a smooth additive process, but a transformative one, in which one qualitatively distinct mode of reasoning is replaced by another qualitatively distinct mode of reasoning.

An interesting finding is the apparent existence of an additional stage between Kohlberg's stages 3 and 4 (see note 5). This is in keeping with assertions by both Fischer, Hand, and Russel (1984) and Commons (Commons et al., 1983, 1998) that the concrete stage (Kohlberg's stage 2) is followed by both an abstract stage (Kohlberg's stage 3) and a formal stage (Kohlberg's half-stage 3.5 or 3/4). Kohlberg's stages were initially modelled on Piagetian stages, and developed into their present form through a process of bootstrapping. Criteria for scoring at transitional levels were developed through the bootstrapping process, and these levels were never viewed as stages in their own right. That the criteria for 3.5 appear, to a large extent, to capture the formal stage as defined analytically by Commons and Fischer (though Fischer calls them levels rather than stages), and the criteria for stage 3 appear to capture the abstract stage, is a fortuitous "accident" of the bootstrapping method.

This analysis reveals considerably more about moral development than traditional methods, primarily by providing a means for estimating probabilistic, equal-interval item difficulties and person ability estimates. The Rasch family of measurement models have been used extensively in educa-

tional measurement and outcomes assessment. Their potential value in developmental research is enormous. They can be applied to many of the problems faced by developmental researchers. For instance, they can be used to: (1) construct developmental measures; **(2)** examine the construct validity of developmental measures; ***(3)*** calibrate developmental instruments; (4) examine the pooled results from studies that intentionally measure the same developmental construct; (5) compare different developmental scoring systems; and (6) contribute to the creation of universally recognised and accepted sample-free units of measurement (Fisher, 1994). In addition, as demonstrated here, they are an excellent tool for examining stage performance because of the rich information they provide about both individual performances of items and persons in combination with the information they provide about developmental trends.

Our understanding of developmental phenomena hinges, in part, on our ability to construct theoretical models of development and submit these to rigorous empirical examination. Shared understanding of development could be greatly enhanced by "common currencies" for the exchange of quantitative information (Fisher, 1994) such as the sample-free logit metric suggested by the results of the analysis presented in this paper. Until relatively recently, the practical difficulties surrounding developmental research, such **as** restrictions on sample size imposed by time and expense constraints, have made it difficult to devise and adequately test developmental instruments, particularly outside of the logico-mathematical domain. The rigorous but flexible measurement principles employed by Rasch's models permit us to simultaneously re-examine our theoretical constructs and instruments, and open the door to new insights.

Manuscript received December 1999
Revised manuscript received March 2000

References

Adams, R.J., & Khoo, S.-T. (1993). *Quest: The interactive test analysis system.* Victoria, Australia: Australian Council for Educational Research Ltd.

Alexander, C.N., & Langer, E.J. (Eds.) (1990). *Higher stages of human development: Perspectives on adult growth.* New York Oxford University Press.

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park, CA: Sage.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J.A. Keats, R. Taft, R.A. Heath, & S.H. Lovibond (Eds.), *Mathematical and theoretical systems* (pp. 7–16). North-Holland Amsterdam.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *Journal of Ealy Adolescence, 14,* 328–344.

Armon, C. (1984). Ideals of the good life and moral judgment: Ethical reasoning across the lifespan. In M. Commons, F. Richards, & C. Armond (Eds.), *Beyond formal operations: Vol. 1. Late adolescent and adult cognitive development.* New York Praeger.

Armon, C. (1993). Developmental conceptions of good work: A longitudinal study. In J. Demick & P.M. Miller (Eds.), *Development in the workplace* (pp. 21–37). Hillsdale, NJ: Erlbaum.

Armon, C., & Dawson, T.L. (1997). Developmental trajectories in moral reasoning across the lifespan. *Journal of Moral Education, 26,* 433–453.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis.* Oxford, UK. Oxford University Press.

Bergan, J.R. (1988). Latent variable techniques in measuring development. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 233–261). New York: Plenum.

Bidell, T.R., & Fischer, K.W. (1992). Beyond the stage debate: Action, structure, and variability in Piagetian theory and research. In R.J. Sternberg & C.A. Berg (Eds.), *Intellectual development* (pp. 100–140). New York Cambridge University Press.

Bond T.G. (1994). Piaget and measurement: II. Empirical validation of the Piagetian model. *Archives de Psychologie, 63,* 155–185.

Bond, T., & Bunting, E. (1995). Piaget and measurement: III. Reassessing the methode clinique. *Archives de Psychologie,* **63,** 231–255.

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

Colby, A., & Kohlberg, L. (1987b). *The measurement of moral judgment: Vol. 2. Standard issue scoring manual.* New York Cambridge University Press.

Commons, M., Armon, C., Kohlberg, L., Richards, F.A., Grotzer, T.A., & Sinnott, D. (Eds.) (1989b). *Adult development: Vol. 2. Models and methods in the study of adolescent and adult thought.* New York: Praeger.

Commons, M.L., Armon, C. Richards, F.A., Schrader, D.E., Farrell, E.W., Tappan, M.B., & Bauer, N.F. (1989a). A multidomain study of adult development. In D. Sinnott, F.A. Richards, & C. Armon (Eds.), *Adult development: Vol. 1. Comparisons and applications of developmental models* (pp. 33–56). New York: Praeger.

Commons, M.L., Richards, F.A., with Ruf, F.J., Armstrong-Roche, M., & Bretzius, S. (1983). A general model of stage theory. In M. Commons, F.A. Richards, & C. Armon (Eds.), *Beyond formal operations* (pp. 120–140). New York: Praeger.

Commons, M.L., Trudeau, E.J., Stein, S.A., Richards, S.A., & Krause, S.R. (1998). Hierarchical complexity of tasks show the existence of developmental stages. *Developmental Review, 8,* 237–278.

Dawson, T.L. (1998). *"A good education is ..." A life-span investigation of developmental and conceptual features of evaluative reasoning about education.* Doctoral dissertation, University of California at Berkeley, CA.

Dawson, T.L. (2000). Moral reasoning and evaluative reasoning about the good life. *Journal of Applied Measurement, 1,* 346–371.

Demetriou, A., Efklides, A., Papadaki, M., Papantoniou, G., & Economou, A. (1993). Structure and development of causal-experimental thought: From early adolescence to youth. *Developmental Psychology, 29,* 480–497.

Draney, K.L. (1996). *The polytomous Saltus model; A mixture model approach to the diagnosis of developmental differences.* Unpublished doctoral dissertation, University of California at Berkeley, CA.

Epstein, H.T. (1990). Stages in human mental growth. *Journal of Educational Psychology, 82,* 876–880.

Fischer, K.W., & Bidell, T.R. (1998). Dynamic development of psychological structures in action and thought. In W. Damon & R.M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (5th ed., pp. 467–561). New York: Wiley.

Fischer, K.W., Hand, H.H., & Russel, S. (1984). The development of abstractions in adolescence and adulthood. In M.L. Commons, F.A. Richards, & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development* (pp. 43–73). New York: Praeger.

Fischer, K.W., & Kennedy, B. (1997). Tools for analyzing the many shapes of development: The case of self-in-relationships in Korea. In K.A. Renninger & E. Amsel (Eds.), *Process of development* (pp. 117–152). Mahwah, NJ: Erlbaum.

Fisher, W.P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement* (pp. 36–72). Norwood, NJ: Ablex.

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

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development.* Cambridge, MA: Harvard University Press.

Hartelman, P.A., van der Maas, H.L.J., & Molenaar, P.C.M. (1998). Detecting and modeling transitions. *British Journal of Developmental Psychology, 16,* 97–122.

Hautamaki, J. (1989). The application of a Rasch model on Piagetian measures of stages of thinking. In P. Adley (Ed.), *Adolescent development and school science* (pp. 342–349). London: Falmer.

Kegan, R. (1982). *The evolving self: Problem and process in human development.* Cambridge, **MA:** Harvard University Press.

Kelderman, H. (1986). *Common item equating with the log-linear Rasch model.* Twentye, The Netherlands: Department of Education, University of Twente.

King, P.M., & Kitchener, K.S. (1994). *Developing reflective judgment.* San Francisco, CA: Jossey-Bass.

Kingma, J., & van den Boss, K.P. (1988). Unidimensional scales: New methods to analyze the sequences in concept development. *Genetic, Social, and General Psychological Monographs, 114,* 477–508.

Kohlberg, L. (1969). Stage and sequence: The cognitive-develpmentalapproach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago, IL: Rand McNally.

Kohlberg, **L.,** & Higgins, A. (1984). Continuities and discontinuities in childhood and adult development revisited—again. In L. Kohlberg (Ed.), *The psychology of moral development: The nature and validity of moral stages* (Vol. 2, pp. 426–497). San Francisco, CA: Jossey-Bass.

Lourenco, O., & Machado, A. (1996). In defence of Piaget's theory: A reply to 10 common criticisms. *Psychological Review,* 103, 143–164.

Markoulis, D. (1989). Postformal and postconventional reasoning in educationally advanced adults. *Journal of Genetic Psychology, 150,* 427–439.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G.N. (1994). Partial credit model. In T. **Husen** & T.N. Postlethwaite (Eds.), *The international encyclopedia of education* (pp. 4302–4307). London: Pergamon.

Muller, U., Sokol, B., & Overton, W.O. (1999). Developmental sequences in class reasoning and proportional reasoning. *Journal of Experimental Child Psychology, 74,* 69–106.

Nisan, M., & Kohlberg, L. (1982). Universality and variation in moral judgment: A longitudinal and cross-sectional study in Turkey. *Child Development,* **53,** 865–876.

Noelting, G., Coude, G., & Rousseau, J.P. (1995, June). *Rasch analysis applied to multi-domain tasks.* Paper presented at the Twenty-Fifth Annual Symposium of the Jean Piaget Society, Berkeley, CA.

Nucci, L., & Pascarella, E.T. (1987). The influence of college on moral development. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 3, pp. 271–326). New York: Agathon Press.

Puka, B. (1991). Toward the redevelopment of Kohlberg's theory: Preserving essential structure, removing controversial content. In W.M. Kurtines & J.L. Gewirtz (Eds.), *Handbook of moral behavior and development: Vol. 1. Theory* (pp. 373–393). Hillsdale, NJ: Erlbaum.

Rasch, G. (1980). *Probabilistic model for* **some** *intelligence and attainment tests.* Chicago, IL University of Chicago Press.

Selman, R.L. (1980). *The growth of interpersonal understanding: Developmental and clinical analyses.* New York: Academic Press.

Sinnott, J.D., & Cavanaugh, J.C. (Eds.) (1991). *Bridging paradigms: Positive development in adulthood and cognitive aging.* New York: Praeger.

Smith, L. (1993). Necessary knowledge: Piagetian perspectives on constructivism. Mahwah, NJ: Erlbaum.

Snarey, J.R., Reimer, J., & Kohlberg, L. (1985). Development of social-moral reasoning among Kibbutz adolescents: A longitudinal cross-cultural study. *Developmental Psychology, 21,* 3–17.

Turiel, E., & Davidson, P. (1986). Heterogeneity, inconsistency, and asynchrony in the development of cognitive structures. In I. Levin (Ed.), *Stage and structure: Reopening the debate* (pp. 106–143). Norwood, NJ: Ablex.

Vyuck, R. (1981). *Critique and overview of Piaget's genetic epistemology, 1965–1980.* New York: Academic Press.

Walker, L.J. (1982). The sequentiality of Kohlberg's stages of moral development. *Child Development, 53,* 1330–1336.

Walker, L.J. (1984). Sex differences in the development of moral reasoning: A critical review. *Child Development,* ***55,*** 677–691.

Walker, L.J. (1986). Experiential and cognitive sources of moral development in adulthood. *Human Development, 29,* 113–124.

Walker, L.J. (1989). A longitudinal study of moral reasoning. *Child Development, 60,* 157–166.

Willett, J.B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49,* 587–602.

Wilson, M. (1985). *Measuring stages of growth. A psychometric model of hierarchical development* (Occasional paper 29). Australian Council for Educational Research.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago, IL Mesa Press.