Volume 22, issue 2, April–June 2007 ISSN 0885-2014

ELSEVIER

COGNITIVE DEVELOPMENT

DEANNA KUHN
EDITOR

The Official Journal of
The Jean Piaget Society for
the Study of Knowledge
and Development

Available online at www.sciencedirect.com
ScienceDirect

Cognitive Development 22 (2007) 213–229

COGNITIVE DEVELOPMENT

A cross-cultural validation of stage development: A Rasch re-analysis of longitudinal socio-moral reasoning data

Jan Boom [a,*], Hans Wouters [a,1], Monika Keller [b]

[a] *Universiteit Utrecht, Utrecht, The Netherlands*
[b] *Max-Planck-Institute for Human Development, Berlin, Germany*

Abstract

Kohlberg's characterization of moral development as displaying an invariant hierarchical order of structurally consistent stages is losing ground. However, by applying Rasch analysis, Dawson recently gave new interpretation and support to his characterization of stage development. Using Rasch models, we replicated and strengthened her findings in a re-analysis of three sets of longitudinal socio-moral reasoning data collected in Iceland. A new application of Rasch analysis provided support for upward development. Our results supported Kohlberg's characterization of stage development and the cross-cultural stability of Dawson's findings that were exclusively based on US samples. We conclude that proposals to replace Kohlberg's characterization of moral development are premature.

Keywords: Socio-moral development; Rasch Analysis; Hierarchical integration; Structural wholeness

1. Introduction

Recently, the classical debate around stage development has received a new methodological impetus. Although Kohlberg's (1984) model is seminal in the field of moral development, many nowadays regard its basic assumptions as untenable (cf. Krebs & Denton, 2005). However, Dawson and colleagues gave new interpretation to these assumptions of stage development and offered empirical support by applying Rasch analysis. Our study intends to replicate and strengthen these results. This is important, because their findings have not yet been replicated by others. Moreover,

[*] Corresponding author at: Department of Developmental Psychology, Universiteit Utrecht, Heidelberglaan 1, 3508 TC Utrecht, The Netherlands.

E-mail address: J.Boom@fss.uu.nl (J. Boom).

[1] Present address: Academic Medical Center, University of Amsterdam, The Netherlands.

0885-2014/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogdev.2006.10.005

this study aims to connect their findings to other supportive findings of stage development (Boom, Brugman, & Van der Heijden, 2001; Keller & Wood, 1989; cf. Lourenço & Machado, 1996) by applying Rasch analysis to previously analyzed data by Keller and coworkers (Edelstein, Keller, & Schröder, 1990; Keller, 1996; Keller & Wood, 1989).

1.1. *Classical characteristics of structural development*

In Kohlberg's characterization of development that originates in Piaget's work (Flavell, 1972; Piaget, 1960), moral development progresses (1) through an invariant hierarchical order of (2) structurally consistent stages. Furthermore, (3) development is upwards and (4) progresses similarly across cultures (Colby & Kohlberg, 1987; Colby, Kohlberg, Gibbs, & Lieberman, 1983; Kohlberg, 1984).

1.1.1. *Hierarchical and invariant order*

Stages are considered to result from qualitative reorganizations of thought. Each stage integrates the previous stage leading to an invariant order of stages. This means that each individual has to pass through each stage (Colby & Kohlberg, 1987; Selman, 1980). Kohlberg and collaborators provided careful theoretical analysis of interview protocols as support for this characterization of stage development. Empirical support was typically indirect: in Colby and Kohlberg's original studies, the number of upward progressions exceeded the number of regressions, and the number of regressions was small enough to justify their attribution to measurement error (Colby & Kohlberg, 1987; Colby et al., 1983). Together with the infrequent occurrence of stage skipping, Colby and Kohlberg took this as support for the invariant and hierarchical order of the stages.

1.1.2. *Structural consistency*

Stages are considered to be structured wholes. In Colby and Kohlberg's view a "form of reasoning can be abstracted from the content of an individual's response to a variety of situations" (Colby & Kohlberg, 1987, p. 77). Consistent with this notion, Colby and Kohlberg report in their original studies that the majority of participants' scores reflected a modal stage although almost every participant received scores at two adjacent stages (Colby & Kohlberg, 1987; Colby et al., 1983). Note that Colby and Kohlberg did not assume every judgment to be based on one exclusive stage structure.

1.1.3. *Upward development*

Colby and Kohlberg expected the individual developmental process to be upwards, provided that the environmental conditions were normal. Upward development follows from the invariant hierarchical ordering of the stages and results in a step-by-step attainment of successive stages with neither regressions nor skipping of stages.

1.1.4. *Cross-cultural consistency*

Colby and Kohlberg expected invariant hierarchical order and structural consistency to be valid across a wide range of different cultural and societal contexts and Colby and Kohlberg (1987), Colby et al. (1983) claimed this to be the case.

Such a Piagetian characterization of development has been abandoned in the field of cognitive development (Brainerd, 1978). However, in the field of moral development the debate is going on. Some theorists have criticized Kohlberg's claims as being too strong (Krebs & Denton, 2005; Rest, Narvaez, Bebeau, & Thoma, 1999). Others have supported Kohlberg's characterization of moral development. Straightforward support for stage hierarchy comes from Boom et al. (2001) who report a substantial correlation between Kohlberg's stage order and the average stage ordering by the participants in their study. Further support for upward development in longitudinal analyses with smaller intervals between the waves than in Colby and Kohlberg's studies is provided by Keller et al. (Keller, 1986, 1996; Keller & Wood, 1989) Cross-cultural consistency was further supported in a Chinese sample (Fang, Fang, Keller, Edelstein & Kehle, 2003).

In contrast, structured wholeness has turned out to be a much more controversial stage characteristic. Recently, *the extent* to which there is structural consistency has been investigated (Krebs, Denton, Vermeulen, Carpendale, & Bush, 1991; Teo, Becker, & Edelstein, 1995; Walker, Gustafson, & Hennig, 2001). Berkowitz and Keller (1994) showed that, although the attainment of stages varied across the issues, the stages were sequentially ordered within issues (content aspects) of social and moral development. Moreover, heterogeneity of stage usage decreased over time. Although this characterization softens the structured wholeness of the stages, it preserves the idea of unity in development.

Perhaps the most persistent problem in settling the issues is methodological: it has been difficult to provide convincing evidence for or against the classical characteristics of structural development, perhaps due to the qualitative nature of the data (interviews) and the large range of possible stages. However, the recent use of Rasch models for evaluating the classical characteristics of structural development seems promising.

1.2. Rasch analysis

Following Dawson and colleagues, the current study employs Rasch models to examine stage development. In Rasch analysis, both participants and items can be quantitatively arranged on a common interval scale (a logit scale). Specifically, with Rasch analysis person abilities and item difficulties are estimated (Bond & Fox, 2001; Van der Linden & Hambleton, 1997). The difficulty of an item is based on the proportion of persons who respond correctly to that item. If most people respond correctly to an item, it will be scaled as easy, whereas if few people respond correctly, the item will be estimated as difficult. Similarly, the person abilities are based on the proportions of items that are correctly responded to by those persons. Persons high in ability will get a high proportion of items correct, while the reverse is true for individuals with low ability. The likelihood that a person will correctly respond to an item depends on the difficulty of that item relative to the ability of that particular person (Wright & Masters, 1982). If the person ability is less than the difficulty of an item, the likelihood to provide a correct response to this item is low. If the person ability increases, the likelihood to respond correctly becomes larger and the likelihood to respond correctly to a more difficult item also increases. These properties facilitate a straightforward employment of certain advanced Rasch models to examine stage development. Such Rasch models enable us to arrange stages and individuals on a common scale, respectively, on the basis of difficulty and ability estimates. In the context of moral development, and with more than two possible stage scores per issue, the terms ability and difficulty are perhaps a bit awkward, nevertheless, we continue to use them for sake of brevity.

The Rasch analyses by Dawson et al. provide support for Kohlberg's stage characteristics (Dawson, 2002a, 2002b; Dawson-Tunik, Commons, Wilson, & Fischer, 2005). In regard to invariant hierarchical order, Dawson et al. report a cumulative ordering of stage difficulties that is consistent with the stage order as proposed by Kohlberg. In support of structural consistency, Dawson (2002a, 2002b) found that reasoning responses reflecting a same stage across different aspects of moral reasoning differed only minor in difficulty. Conversely, there were larger differences in difficulty between reasoning responses representing different stages. Upward development was supported by Dawson's analysis in which ability estimates were predicted with the log of age as the independent variable.

1.3. Hypotheses of this study

(1) *Hierarchical and invariant order*: we expect that the appropriate Rasch models will reveal that the stage difficulties are clearly separated on the logit scale and in the order proposed by Kohlberg. (2) *Structural consistency*: in regard to different content aspects of moral reasoning, we expect reasoning reflecting the same stage to be equally difficult. (3) *Upward development*: we expect that a model reflecting upward development will be supported in favor of a model reflecting longitudinal stability. (4) *Cross-cultural consistency*: we expect to replicate and strengthen Dawson et al.'s American sample results with our re-analysis of Icelandic data.

2. Method

2.1. Participants

This re-analysis is based on the longitudinal data collected in Iceland by Edelstein et al. (1990). Samples were drawn from urban (Reykjavik) and rural Iceland. The sample size differed slightly for the three interviews as well as the number of assessments and ages at which these occurred. Interview 1 was administered to 99 urban participants and 54 rural participants at the ages of 9, 12, and 15 years. Interview 2 was administered to 102 urban participants and 60 rural participants at the ages of 12 and 15 years. Interview 3 was administered to 99 urban participants and of 56 rural participants at the age of 7, 9, 12 and 15 years (see Berkowitz & Keller, 1994; Edelstein et al., 1990; Keller, 1996; Keller & Wood, 1989).

2.2. Interviews and dilemmas

The interviews assess stage development in the socio-moral domain according to Selman's and Kohlberg's theories. According to Selman (1980), social perspective taking refers to the *descriptive* social understanding of the relationship between one's own and others' perspectives. Kohlberg (1984) regarded social perspective taking as the necessary but not sufficient basis of *prescriptive* moral judgments. Both social perspective taking and moral judgments are involved in the understanding of the specific facets of the personal and the social world including its normative demands (Keller & Edelstein, 1991). Descriptive and prescriptive reasoning can be seen as closely intertwined in reasoning about actions, persons and relations and the normative expectations governing interaction. In the socio-moral domain, the process of perspective-differentiation and coordination forms the structural basis of the different stages of development in descriptive-social perspective taking and prescriptive-moral reasoning. Each of the three interviews consists of several content aspects (issues) that are assessed with probe questions.

2.2.1. Interview 1

"*Concepts of own friendship*" was adapted from Selman (1980). Unlike Selman's interview that assessed concepts of friendship with a *hypothetical* friendship dilemma, Edelstein et al. (1990) interview assessed concepts of children's *own* friendships. Six issues were used, (see Appendix A).

2.2.2. *Interview 2*

Kohlberg's "*Judy dilemma*" assessed participants' moral reasoning in a situation of truth-telling to mother versus sister-loyalty in a conflict in which the daughter disobeyed the mother (Colby et al., 1987). The dilemma was modified (1) by giving the characters Icelandic names and (2) by including the mother's direct confronting question to the protagonist where the sister *actually* is. This manipulation puts more decisional pressure on the participants. Moral reasoning was assessed with nine issues (see Appendix B).

2.2.3. *Interview 3*

"*Moral Sensibility*" assessed participants' socio-moral reasoning about friendship with a dilemma adapted from Selman (1980). Its central conflict was the contradiction of friendship obligations with the protagonist's *own* hedonistic or altruistic concerns. Specifically, the protagonist had to decide whether to keep a promise of meeting an old friend or to go to the movie with a child, who was new in the class. It was left open to the child whether he or she focused on the hedonistic or morally relevant aspects of the problem. Moral sensibility was assessed with 11 issues (see Appendix C).

2.3. *Procedure and coding of the interviews*

Trained interviewers interviewed individual participants. Icelandic-English bilinguals transcribed and translated the interviews in English. Trained scorers assigned stage scores to participants' reasoning responses on the basis of different manuals, one for each interview. Interview 1 was scored by an adapted version (Essen, Keller and Mönnig 1987) of Selman's original manual (Keller & Edelstein, 1991; Selman & Jaquette, 1977). Interview 2 was scored with and adapted version (Keller, Eckensberger, & Rosen, 1989; Rosen, 1987) of the Standard Issue Scoring Manual (Colby & Kohlberg, 1987). Interview 3 was scored with a manual by Keller (1996) that is partly consistent with the other manuals. In particular, this last manual redefines pre-conventional morality (stages 1 and 2) by including empathy and relationship concerns (Keller, 1996; Keller et al., 1989).

Full and transitional half-stage scores (for interview 2 and 3) were defined for each of the issues (content aspects) separately. This procedure enabled a precise comparison of the issues, which is central to our data analyses. Inter-rater reliabilities varied, but were minimally 80% and mostly around 90% of full agreement in each interview (Keller, 1996). If participants responded to an issue with two or more answers, the highest stage response was recorded.

2.4. *Analysis*

First we explain how we applied Rasch analysis to the data followed by how Rasch analysis supports stage development. Because we identified more than two stages, we employed Rasch models that are suited for polytomous items, i.e. items with multiple response categories (Bond & Fox, 2001; Ostini & Nering, 2006). In order to distinguish the response categories, these multiple response models estimate several difficulty parameters per issue. These difficulty parameters define the likelihood of responding according to a response category, given the person ability. This is illustrated in Figs. 1–3 where each Item Category Characteristic Curve (ICCC) represents

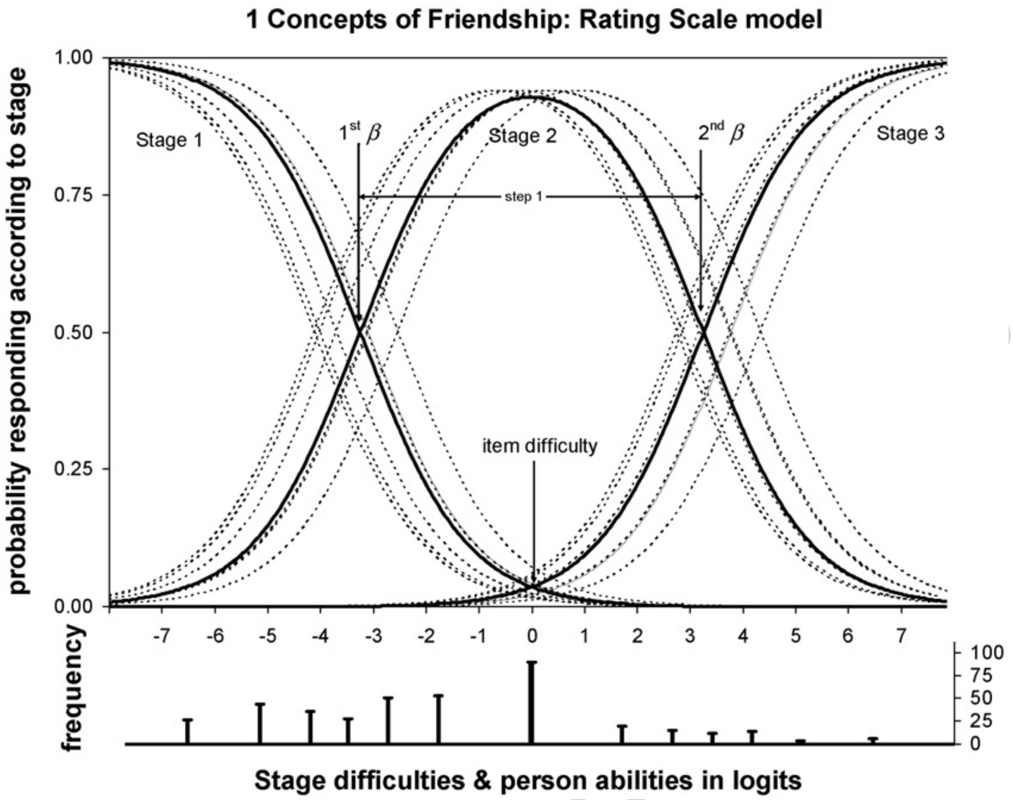

Fig. 1. ICCCs for interview 1, modeled with the Rating Scale model (light dotted curves) and the restricted Rating Scale model (heavy black curves), on the primary axis. Stage difficulty βs are at the intersections of ICCCs of two adjacent stages. The general difficulty for the item is at the intersection of the first and last ICCC. The differences between subsequent βs are referred to as steps. Distribution of abilities of participants (combined sample) is depicted on the secondary axis, using the same logit scale.

the likelihood of responding according to a certain response category (stage) as a function of the difficulty of that stage and person ability. The general shape (steepness) of the ICCCs follows from model assumptions but size (height) and location (along the x-axis) are completely defined by the issue difficulty parameters β and person ability parameters θ. As illustrated in Fig. 1, β is the value on the logit scale where two adjacent ICCCs intersect, therefore, if there are three stages, only two βs are needed to construct three curves.[1]

ConQuest (Wu, Adams, & Wilson, 1998) was used to estimate the stage difficulties with a Partial Credit Model, a normal Rating Scale model, and a restricted version of the Rating Scale

[1] For the interested reader we describe how the ICCCs are obtained (adapted from Verhelst, Glas, & Verstralen, 1995, p. 2). It is assumed that the response to issue i, denoted by X_i falls in the score range (0, 1, . . ., m_i) representing the stages in the sample (with the lowest stage indexed by $j=0$ and the highest by $j=m$). The probability of observing $X_i=j$ as a function of ability parameter θ is given by:

$$P(X_i = j|\theta) = \frac{\exp\left(j\theta - \sum_{g=1}^{j}\beta_{ig}\right)}{1 + \sum_{h=1}^{m_i}\exp\left(h\theta - \sum_{g=1}^{h}\beta_{ig}\right)} \quad (j = 0, \ldots, m_i) \tag{1}$$

where β_{ig}, $g = 1, \ldots, m_i$ are the difficulty parameters of issue i. Subsequently, the parameter estimates can for each stage j, be transformed to probability responding according to stage j (see Eq. (1)). Plotting X_i for various θ in one graph depicts each ICCC representing the estimated probability responding according to a stage, as a function of the stage difficulty β and the person ability θ (see Fig. 1).

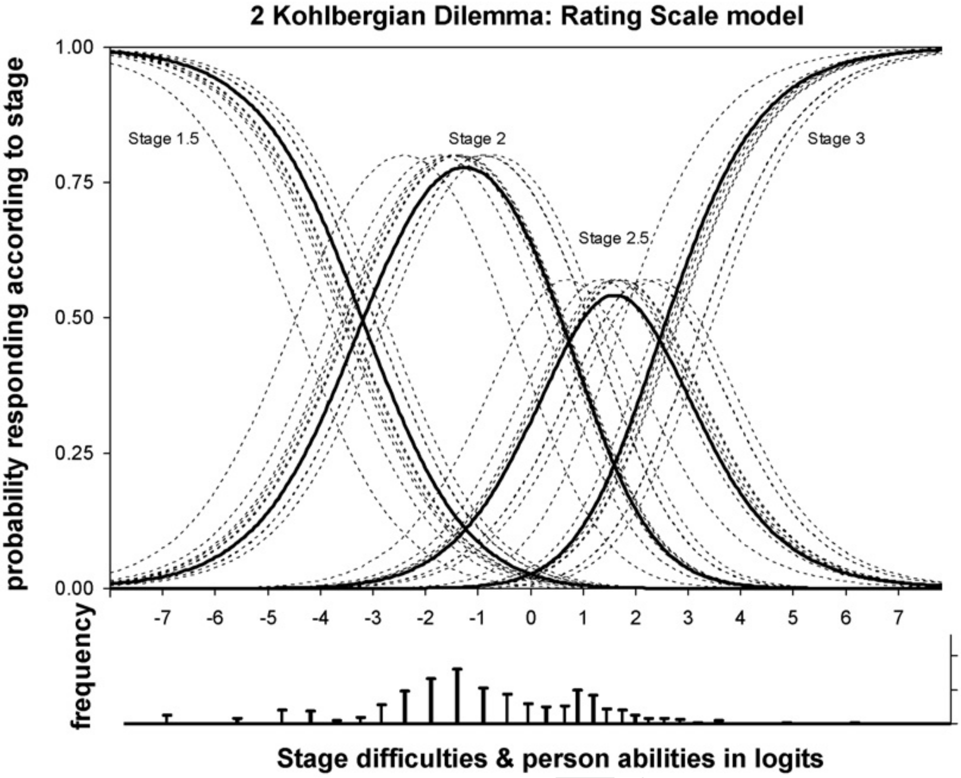

Fig. 2. ICCCs for interview 2, modeled with the Rating Scale model (light dotted curves) and the restricted Rating Scale model (heavy black curves), on the primary axis. Distribution of abilities of participants (combined sample) is depicted on the secondary axis, using the same logit scale.

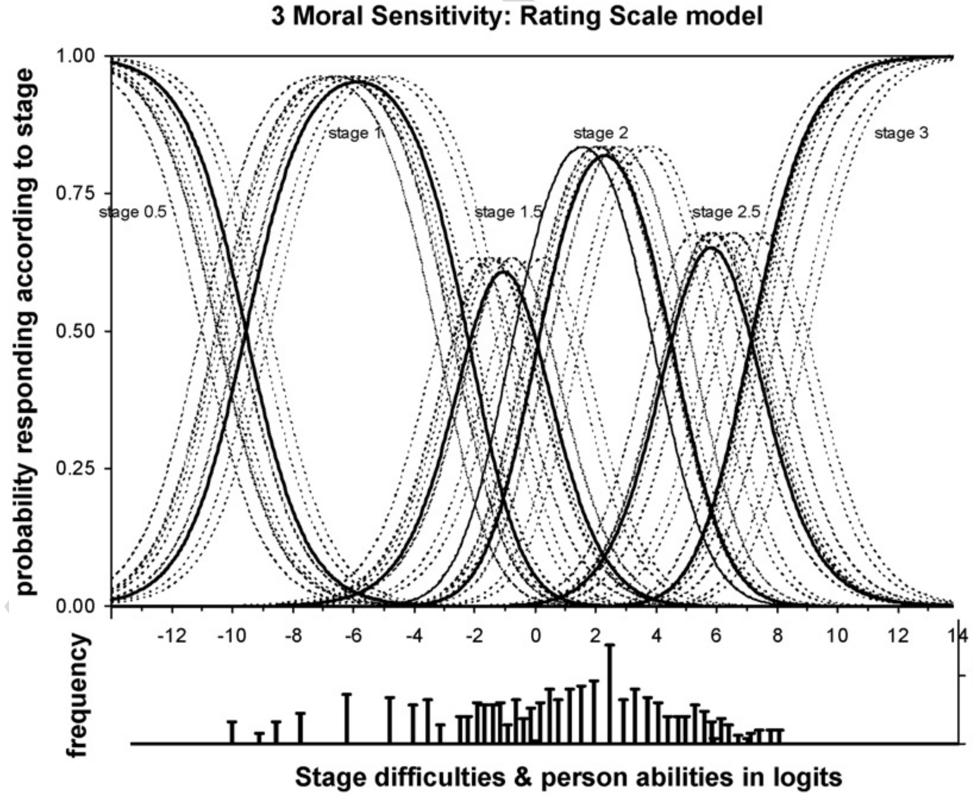

Fig. 3. ICCCs for interview 3, modeled with the Rating Scale model (light dotted curves) and the restricted Rating Scale model (heavy black curves), on the primary axis. Distribution of abilities of participants (combined sample) is depicted on the secondary axis, using the same logit scale.

model. To illustrate the differences between these models, assume that the value of each β is based on a general issue parameter and one or more steps (see Fig. 1). In The *Partial Credit Model* both the general issue parameter and the steps vary across each issue. As a result each issue has a different set of βs. This results in ICCCs that vary in location and size. The *Rating Scale model* is itself a restricted version of the Partial Credit Model (Andrich, 1978). The steps between adjacent βs are similar across the issues. However, the general issues parameters vary. The result is that ICCCs reflecting the same stage still vary in location, but not in size. In an attempt to further restrict the Rating Scale model, we fixed each general issue parameter at zero. This produced equal location and size in ICCCs reflecting the same stage, or in other words identical ICCCs.

We used Akaike's Information Criterion (AIC) to compare models that showed an acceptable fit to the data. As the Partial Credit Model estimates the difficulty estimates for each issue, it has more parameters than the other models. AIC corrects the goodness of fit of a model for the number of parameters in the model (Akaike, 1973). AIC takes this into account and reduces the possibility that the Partial Credit Model would have a better fit simply because it has more parameters. AIC is calculated as $-2\log L + 2\text{npar}$, where log is the natural logarithm, L is the likelihood, and npar is the number of parameters in the model. The smaller the AIC value, the better the fit.

2.4.1. *Hierarchical and invariant order*

Support for the hierarchical and invariant order of the stages follows from a good fit for the polytomous Rasch models. A good fit implies: (1) that the stages constitute one dimension along the logit scale; (2) a neat separation of the locations of the ICCCs along the logit scale and no extreme size differences (except for the first and the last one); (3) person abilities that are spread out roughly over the same range as the difficulties along the logit scale. Furthermore, models based on orderings that differ from Kohlberg's stage order (achieved by recoding the stages) should show a worse fit, small ICCCs and person ability estimates that lump together. Support for the hierarchical order implies support for invariant ordering and vice versa. Stage skipping is unlikely if the stages (ICCCs) are well separated as in the example in Fig. 1, because in normal development along the logit scale, each subsequent stage is the most likely stage for a particular range of ability levels. However, a small ICCC indicates a recessive stage that is likely to be skipped. We addressed hierarchical order with the restricted Rating Scale model because it is the most parsimonious model possible that still reflects the core of this hypothesis. For the assessment of hierarchical and invariant order we decided not to use the longitudinal structure of the data. Due to computational constraints, data from the next measurement occasions were treated as if they concerned older participants measured at the same occasion. At the same time we assumed that our violation of the assumption of independence of outcomes was probably not severe in these analyses (see Willet, 1989).

2.4.2. *Structural consistency*

In Rasch analysis, the degree to which the ICCCs that represent the same stage are similar across different issues within an interview indicates the degree to which a stage is structurally consistent. At present it is unclear exactly *how* similar the ICCCs should be to claim structural consistency. We will use the term *full* structured wholeness to imply that the ICCCs have exactly the same shape, size and location across the issues. We use the term *moderate* structured wholeness to imply that there are small differences between the location and size across the issues. To test the prediction that "moderate structured wholeness" was more appropriate than "full structured wholeness", we compared the three models distinguished above. Moderate structured wholeness is reflected by βs being *approximately* equal over issues as in the Partial Credit Model and also

Table 1
Difficulty estimates (βs) for interview 1: concepts of own friendship

	Stages	
	1 and 2	2 and 3
Rating Scale model restricted		
Overall estimate (S.E.)	−3.25 (*.05*)	3.25[a]
Rating Scale model normal		
Average estimate (S.E.)	−3.44 (*.05*)	3.44
Standard deviation over issues	0.61	0.61
Partial Credit Model		
Issues		
1. Ideal friend	−3.20 (*.16*)	3.89
2. Motivation	−2.99 (*.16*)	3.37
3. Mechanism	−2.67 (*.16*)	4.60
4. Conflict resolution	−4.31 (*.16*)	3.21
5. Closeness/intimacy	−3.28 (*.15*)	2.38
6. Trust/reciprocity	−5.01 (*.14*)	4.01

Note: On the logit scale the βs are located at the intersections of two adjacent ICCCs (see Fig. 1). Standard errors (S.E.) are provided within parentheses. Horizontally read-off βs are significantly different under the criterion of Wright and Stone (1979, p. 95).

[a] The standard error for the last β in a row is redundant.

(but perhaps to a lesser degree) in the normal Rating Scale model. Full structured wholeness is reflected by βs being *exactly* equal over issues as is the case in our restricted version of the Rating Scale model.

2.4.3. Upward development

To assess upward development two models yielding two different longitudinal ability patterns are compared. In the first model, ability is estimated for each occasion. In the second model, ability is kept constant across occasions. A better goodness of fit for the first model supports upward development. For these analyses we focus on the person abilities and therefore use the data in their original longitudinal form. We chose a Multidimensional Rating Scale model that represented each longitudinal measurement occasion as a different dimension. In this way, we identified average ability for each measurement occasion.[2] To avoid Type I error, we fixed the means of the second model at a value with the lowest possible deviance ($-2\log L$) that we detected through a stepwise procedure.

3. Results

Tables 1–3 present sets of estimated βs for each interview. For the restricted Rating Scale model only one set of βs is estimated. For the normal Rating Scale model we report the average β of the first boundary, the average for the second boundary, etc. Differences between issues are captured by the standard deviations. Figs. 1–3 depict the normal Rating Scale models in more detail. For

[2] To avoid the confounding influence of different stage difficulties at different measurement occasions (e.g. instrumentation effects), we kept the stage difficulties constant across the measurement occasions. We achieved this by imputing for each dimension the difficulty estimates of the Rating Scale model of the combined measurement occasions.

Table 2
Difficulty estimates (βs) for interview 2: Kohlbergian dilemma

	Stages		
	≤1.5 and 2	2 and 2.5	2.5 and 3
Rating Scale model restricted			
Overall estimate (S.E.)	−3.19 (*.05*)	0.73 (*.05*)	2.45
Rating Scale model normal			
Average estimate (S.E.)	−3.44 (*.09*)	0.74 (*.09*)	2.70
Standard deviation over issues	0.52	0.52	0.52
Partial Credit Model			
Issues			
1. Moral reasoning (tell)	−2.39 (*.16*)	1.25 (*.18*)	2.73
2. Moral reasoning (not tell)	−3.15 (*.18*)	1.39 (*.18*)	3.06
3. Property reasoning	−3.55 (*.18*)	1.79 (*.20*)	2.80
4. Contract reasoning	−4.01 (*.19*)	1.00 (*.17*)	2.88
5a. Consequences for authority	−3.54 (*.18*)	0.42 (*.16*)	2.87
5b. Consequences for family	−3.53 (*.18*)	0.57 (*.16*)	3.26
6. Moral reasoning chosen action	−4.04 (*.20*)	0.08 (*.16*)	3.34
7. Affiliation reasoning	−3.31 (*.17*)	0.58 (*.16*)	2.41
8. Authority reasoning	−3.81 (*.14*)	−0.35 (*.13*)	1.28

Note: On the logit scale, the βs are located at the intersections of two adjacent ICCCs (see Fig. 1). Standard errors (S.E.) are provided within parentheses. Horizontally read-off βs are significantly different, see Table 1.

Table 3
Difficulty estimates (βs) for interview 3: moral sensibility

	Stages				
	0.5 and 1	1 and 1.5	1.5 and 2	2 and 2.5	2.5 and ≥3
Rating Scale model restricted					
Overall estimate (S.E.)	−9.56 (*.06*)	−2.17 (*.04*)	0.09 (*.04*)	4.50 (*.04*)	7.14
Rating Scale model normal					
Average estimate (S.E.)	−10.28 (*.09*)	−2.47 (*.08*)	−0.01 (*.08*)	4.62 (*.08*)	7.50
Standard deviation over issues	0.63	0.63	0.63	0.63	0.63
Partial Credit Model					
Issues					
1. Spontaneous def. of problem	−7.59 (*.24*)	−0.50 (*.22*)	1.06 (*.16*)	5.45 (*.17*)	9.34
2a. Reasons for choice “friend”	−11.21 (*.21*)	−3.20 (*.17*)	−0.45 (*.16*)	4.90 (*.15*)	7.98
2b. Reasons for choice “movie”	−13.08 (*.31*)	−0.98 (*.19*)	1.29 (*.15*)	6.50 (*.21*)	9.25
3. Moral evaluation of choice	−11.25 (*.21*)	−2.51 (*.16*)	−0.35 (*.17*)	4.71 (*.15*)	7.52
4. Perspective of protagonist	−10.44 (*.24*)	−2.93 (*.18*)	0.19 (*.17*)	5.82 (*.18*)	9.10
5. Perspective of friend	−11.99 (*.22*)	−3.20 (*.17*)	−0.21 (*.17*)	4.67 (*.15*)	8.44
6. Negotiation strategies	−12.12 (*.24*)	−2.62 (*.16*)	0.36 (*.16*)	5.11 (*.15*)	8.46
7. Balancing strategies	−12.58 (*.25*)	−1.89 (*.15*)	1.22 (*.14*)	5.40 (*.15*)	8.67
8. Sit. spec. confl. understanding	−11.76 (*.24*)	−3.26 (*.17*)	−0.45 (*.17*)	4.85 (*.15*)	8.87
9. Sit. spec. friend. understand	−10.26 (*.22*)	−3.53 (*.18*)	−0.31 (*.15*)	4.19 (*.16*)	6.82
10. Promise keeping	−8.33 (*.15*)	−2.27 (*.12*)	−0.22 (*.13*)	3.52 (*.14*)	5.80

Note: On the logit scale, the βs are located at the intersections of two adjacent ICCCs (see Fig. 1). Standard errors (S.E.) are provided within parentheses. Horizontally read-off βs are significantly different, see Table 1.

Table 4
AIC for the Partial Credit model and the 2 Rating Scale models for each of the 3 interviews

Modeltype	Interview 1	Interview 2	Interview 3
Rating Scale model restricted	3677	5343	14311
Rating Scale model normal	3568	5169	13452
Partial Credit Model	3490	5121	13210

Note: AIC is abbreviation of Akaike's Information Criterion: lower values represent better fit.

the Partial Credit Model a set of βs for each issue is estimated. The restricted Rating Scale model has the worst fit in all interviews, whereas the Partial Credit Model has the best fit, with the normal Rating Scale model in between (see Table 4). Although the precise person parameters are not needed for our hypothesis, a trustworthy model requires that the range of person ability estimates is roughly comparable to the range of the stage difficulties, which was confirmed in all three interviews (see lower parts of Figs. 1–3).

Hypothesis 1 (hierarchical order). Differences between the βs demarcating successive stages *within* an issue can be read off horizontally in the top row of Tables 1–3 for the restricted Rating Scale model. These tables correspond to the three interviews. In all interviews, stage 3 was the highest stage that was modeled. Interview 3 displays the largest range of stages, which is consistent with four measurements started at the youngest age. Interview 2 has the smallest range of stages, consistent with only two measurements started at an older age.

For each issue (in each interview), every model reports significant separateness of the difficulty parameters βs (see footnotes of Tables 1–3). However, since the size of the ICCCs is difficult to assess on basis of the βs alone, a graphical representation of the Rating Scale model is provided in Figs. 1–3.

To assess the invariance in terms of model fit, we varied the ordering of the stages. In the dataset we recoded the stage numbers to obtain orderings that contradicted Kohlberg's stage order. In the friendship dilemma, we identified only 3 stage levels resulting in: 1, 2, 3 (Kohlberg's order); 1, 3, 2; and finally 2, 1, 3 (contradicting stage orders). For the Kohlbergian Dilemma with 4 identified stage levels, there are 12 relevant possible alternative orderings. For the Moral sensibility interview with 6 categories there are 360 relevant possible orderings. For these final two interviews we limited ourselves to reversals of neighboring stage levels. The fit for all models with altered orderings was dramatically worse in all cases considered, compared with the restricted Rating Scale models in Table 4. The deviance was 700 higher after reversal of stages 2 and 3 for interview 1 and 803 higher when stages 1 and 2 were reversed. The deviance was 269 higher after reversal of stages 2.5 and 3 for interview 2 and 671 higher when stages 1.5 and 2 were reversed. The increase in deviance for the other reversals was even larger. For interview 3 the deviance was 524 higher after reversal of stages 2.5 and 3 for and 1179 higher when stages 0.5 and 1 were reversed and worse for all other reversals. In addition, in all cases, the βs of the reversed stages lumped together yielding small ICCCs with tops not reaching the 0.5 probability level (see Figs. 1–3).

Hypothesis 2 (structural consistency). In Tables 1–3 each column under the Partial Credit Model can be used to compare issue difficulty parameters β across the issues. Overall, the results for interview 1 provide the strongest support for structural consistency (Table 1). Across the issues, βs in each column are notably more alike than βs in a row. Hence, across the issues, the βs referring to the same stage boundary literally cluster together. For interview 2 (Table 2)

Table 5
Mean upward development across measurement occasions for longitudinal MultiDimensional Rating Scale model

Interview	Mean ability per age group			
	7	9	12	15
1. Concepts of own friendship		−3.87	−1.43	1.98
2. Kohlbergian Dilemma			−1.69	0.120
3. Moral Sensibility	−4.06	0.10	2.94	5.94

and interview 3 (Table 3), the same pattern appears, although less clear-cut. Table 4 reveals that the Partial Credit Model and the normal Rating Scale model both have, for each interview, a better goodness of fit (lower AIC) than the restricted Rating Scale model. Since β estimates are exactly similar across the issues in the restricted Rating Scale model, but vary in the Partial Credit Model and the normal Rating Scale model, this means that stages show differences across issues. However, these differences are relatively small in comparison with the differences of βs between successive stages within the issues.

Hypothesis 3 (upward development). For each of the three interviews, the longitudinal Rating Scale model with freely estimated mean abilities across the measurement occasions has a better fit (lower AIC) than the same model with fixed equal mean abilities: AIC is 3122 instead of 4337 for interview 1; 5073 instead of 5181 for interview 2; 11901 instead of 23836 for interview 3. The increase of mean ability for subsequent measurement occasions (see Table 5) supports upward development. Comparisons based on the Partial Credit Model lead to the same conclusions. Moreover, for all interviews the longitudinal model has a better fit than the combined model[3]: a drop of the AIC of 446 units for interview 1, 96 for interview 2, and 1551 for interview 3. This improvement of fit is largest for interview 3, the interview has the most (four) measurement occasions and smallest for interview 2 which only had 2 measurement occasions.

Hypothesis 4 (cross-cultural consistency). Finally, as hypothesized, we found cross-cultural consistency for Dawson et al.'s results. We replicated their main findings. For invariant hierarchical order (Hypothesis 1), however, we did not rely exclusively on the separateness of the stage difficulties. The likelihood distributions in Figs. 1–3 are more precise than the Thurstone variable maps offered by ConQuest. Moreover, we evaluated the fit of alternative orderings and found them to be wanting. For structural consistency (Hypothesis 2), our results are like the results reported in Fig. 2 from Dawson (2002a, p. 21), although she does not contrast full with moderate structured consistency as we did. Our approach to upward development provided converging results to Dawson's (2002b) regression analysis.

4. Discussion

With Rasch analysis, we examined the characteristics of structural development. The results supported the hierarchical order of the stages, their structural consistency within the interviews, and upward development. As these analyses were based on Icelandic data, there was cross-cultural stability of Dawson et al.'s results and methodology.

[3] All issue-parameters for the longitudinal models were imported from their unidimensional counterpart.

Hierarchical and invariant order was supported because the βs were never reversed and were well separated, in each issue, of each interview, in each variety of polytomous Rasch model that we considered. Note that βs that are reversed or relatively close to each other, only imply that the ICCC between them is small, but cannot affect the ordering of the ICCCs. Interestingly, half-stage ICCCs were consistently smaller in size in Figs. 1–3 which suggest that these half-stages were less often used by the participants and were more likely to be skipped than full stages. This is consistent with findings by Dawson (2002a, 2002b). Apparently, half-stage scores, which are difficult to reconcile with the idea of hierarchical integration, are less consistent as developmental markers than are whole stage scores. Nevertheless, the categories must order according to the integer values used to score the data. More formal support, therefore, followed from the fact that the scoring conform the theoretically proposed order of the stages consistently had a better fit than altered scorings that were inconsistent with the theoretical predicted stage order. In sum, there was ample support for the hierarchical order of the stages.

Structural consistency: In each of the three interviews moderate structured wholeness was supported. The βs representing the same stage boundaries clustered together as they were more similar *across* the issues than βs demarcating different stages *within* each issue. This was the case for both the Rating Scale model and for the Partial Credit Model. The ICCCs in Figs. 1–3 illustrate this clustering for the Rating Scale model. Full structured wholeness was not supported because the restricted Rating Scale model with exactly equal βs had a worse fit than both the Rating Scale model and the Partial Credit Model in which the βs can vary across the issues. These results are consistent with earlier findings (Berkowitz & Keller; 1994; Keller, 1996; Keller & Wood, 1989) and Dawsons analyses (Dawson, 2002a, 2002b; Dawson-Tunik et al., 2005). Our results confirm that these findings are not due to methodological artifacts. Future comparisons of these models with more or with less stringent models might provide a more complete picture of the extent to which there is structural consistency.

Upward development was supported in all interviews, with upward longitudinal ability patterns having a better goodness of fit than longitudinal ability patterns fixed across the measurement occasions. Our results strengthen Dawson's (2002b) finding based on regression analyses that age is a significant and strong predictor of moral ability estimates. The Multidimensional Rating Scale model enabled us to examine upward development whilst controlling for fluctuating difficulty estimates for the stages. This enabled us to cope with the problem of instrumentation, i.e. changes in the instrument from one measurement occasion to another confound participants' abilities. Nevertheless, the third interview posed the most problems for all hypotheses. The enormous improvement in the goodness of fit for the longitudinal model (the Multidimensional Rating Scale model) despite the fact that all issue difficulty parameters were the same, compared to the combined non-longitudinal models, suggests that it was perhaps not justified to assume that there would be no dependencies between measurement occasions (see Willet, 1989). From a theoretical point of view this is not necessarily problematic: it points to individual differences in growth. Although, it is in principle possible to combine growth modeling with Rasch analysis, such an analysis would require much more participants than even large interview based studies can offer.

Approaching stage development with Rasch Analysis has led to new or refined interpretations of the core characteristics of (stage) development and to new questions and answers. An improvement that became clear from both Dawson et al.'s results and even more so from our results is that Rasch analysis enables researchers to address the hierarchical order of the stages empirically, whereas previous studies only examined whether stage development progressed according to a model that postulated invariant hierarchical stage order and structural consistency. The key to this distinction is the independent estimation of the stage difficulties and person reasoning abilities, which is one of the important strengths of original Rasch analysis (Ostini & Nering, 2006). Difficulty estimates reflect the properties of the instrument and ability estimates reflect the developmental process because persons develop while the instrument does not (should be stable). By now it will be clear that our first two hypotheses concern the instrument whereas the third hypothesis (upward development) concerned the individual developmental process. However, invariant hierarchical and upward development are often mixed up, both conceptually and methodologically. Taken together, we collected ample support for stage development by rigorously comparing advanced Rasch models with each its distinct characteristics. Both our results and the results by Dawson et al. show that proposals to replace Kohlberg's characterization of moral development by reductionist views of morality are premature.

Acknowledgments

We would like to thank Wolfgang Edelstein and Eberhard Schröder for access to the data; Herbert Hoijtink, Judith Dubas, John Gibbs, and Daniel Brugman, for their encouragement and comments on earlier drafts of this article. The first and second author contributed equally to the paper. Parts of these results were presented at the 2005 meeting of the Jean Piaget Society, Vancouver, Canada. We would like to thank Applied Psychological Measurement for funding this presentation.

Appendix A

Concepts of friendship: the issues and their questions

Issues	Questions
1. Ideal friend	- Ideal friend What is pleasant/unpleasant/nice/boring about a girl/boy for you? What do you like about boys/girls? Is there something special about him/her that makes you want him/her as a friend?
2. Motivation	- Importance of friendship and Motives for friendship Do you think it is necessary to have a best/a very good friend? Why?/Why not? How do you think it would be to have no friends?
3. Mechanism	- Mechanism of friendship initiation Now let us imagine you came in a new class and would not know anyone: what would you do in order to become friend with someone? How would you find out whether someone is pleasant or boring? What is needed for boys/girls to become friends? What do you do if you want to get a friend?
4. Conflict resolution	- Quarreling in friendship relations and conflict solutions What are the kinds of things you quarrel about? How do you settle a quarrel? Which of you decides what is done? Which of you has to do something in order to become best friends again? Can you be best friends even though you quarrel? How come?

Issues	Questions
5. Closeness/intimacy	- Closeness and Intimacy in friendship relations How do you know you are best friends? What is the difference between a friend and a best friend? How can you be with your best friend that you cannot be with anyone else? What makes a friendship really close?
6. Trust/reciprocity	- Trust and Reciprocity in friendship relation What does it mean to trust your friend? How do you trust him/her? Is trust important in a friendship? Why?

Appendix B

Moral reasoning: the issues and their questions

Issues	Questions
1. Moral reasoning (telling truth)	(if decided to tell) Why is Lilja's decision the right one?
2. Moral reasoning (not telling truth)	(if decided not to tell) Why is Lilja's decision the right one?
3. Property reasoning	Is it important in Lilja's decision, that her sister earns the money by herself?
4. Contract reasoning	Is it important in Lilja's decision, that the mother promised Joná to go the concert? Why is it important in general to keep promises?
5. Consequences	
a. for authority	What are the consequences of this conflict for the parents' authority?
b. for family	What are the consequences of this conflict for the family?
6. Moral reasoning for chosen action	What are the most important reasons for Lilja's decision?
7. Affiliation reasoning	How should a sibling relationship be constituted in general?
8. Authority reasoning	How should a parent–children relationship be constituted in general?

Appendix C

Moral sensibility: the issues and their questions

Issues	Questions
1. Spontaneous definition of problem	"How the actor will decide? How will he/she choose?"
2a. Reasons for choice "friend"	If the child chose for 'friend': "Why he/she will decide so?"
2b. Reasons for choice "new child"	If the child chose for 'new child': "Why he/she will decide so?"
3. Moral evaluation of choice	"Is it a right decision? why is it right/not right?"
4. Perspective of protagonist (mean score)	"How will the protagonist feel after his/her decision?" "What will the protagonist think at the cinema? Why?" "What will the protagonist think at his friend's? Why?"

Issues	Questions
5. Perspective of friend/consequence—global (mean)	“What will/would the friend think/feel, if the protagonist decides to go/decided to go to the movie?” “How would his friend react? Would his friend think this is all right? Why/Why not?” “What are the consequences for the friendship? (if choice friend/movie)” “What will/would the new child think/feel, if the protagonist decides/decided to meet his friend?”
6. Negotiation strategies—global (mean)	“What would the actor have said to the friend, if he/she has gone to the movie?” “What would the actor have said to the friend, if he/she has gone to the friend?” “What would the actor have said to the new child, if he/she has gone to the friend/to the movie?” “What would the actor have said to peers, if he/she has gone to the friend/to the movie?”
7. Balancing strategies—negative consequences	“What could the actor do to balance the friendship with friend?” “What could he/she do to become friends all three together?”
8. Situation specific conflict understanding	Interviewer’s rating of the situation specific conflict understanding
9. Situation specific friendship understanding	Interviewer’s rating of the situation specific understanding of friendship
10. Concept of promise keeping	Interviewer’s rating with three questions including the question at moral evaluation

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–574.

Berkowitz, M. W., & Keller, M. (1994). Transitional processes in social cognitive development. *International Journal of Behavioral Development*, *17*, 447–467.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hillsdale, NJ: Erlbaum.

Boom, J., Brugman, D., & Van der Heijden, P. G. M. (2001). Hierarchical structure of moral stages assessed by a sorting task. *Child Development*, *72*, 535–548.

Brainerd, C. J. (1978). The stage question in cognitive-developmental theory. *The Behavioral and Brain Science*, *12*, 173–213.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment: Theoretical foundations and research validation: vol. 1*, Cambridge: Cambridge University Press.

Colby, A., Kohlberg, L., Gibbs, J., & Lieberman, M. (1983). A longitudinal study of moral judgment. *Monographs of the Society for Research in Child Development*, *48*, 1–124.

Colby, A., Kohlberg, L., Speicher, B., Hewer, A., Candee, D., Gibbs, J., et al. (1987). *The measurement of moral judgment: Standard Issue Scoring Manual (SISM): vol. 2*, Cambridge: Cambridge University Press.

Dawson, T. L. (2002a). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement*, *3*, 1–45.

Dawson, T. L. (2002b). New tools, new insights: Kohlberg’s moral judgment stages revisited. *International Journal of Behavioral Development*, *26*, 154–166.

Dawson-Tunik, T. L., Commons, M. L., Wilson, M., & Fischer, K. W. (2005). The shape of development. *European Journal of Developmental Psychology*, *2*, 163–195.

Edelstein, W., Keller, M., & Schröder, E. (1990). Child development and social structure: A longitudinal study of social structure. In P. Baltes, D. Featherman, & R. Lerner (Eds.), *Life-span development and behavior: 10*, (pp. 152–185). Hillsdale, NJ: Erlbaum.

Essen, C. v., Keller, M., & Mönnig, M. (1987). *Manual zur Entwicklung von Freundschaftsvorstellungen [Manual for scoring friendship reasoning]*. Berlin: Max Planck Institute for Human Development and Education.

Fang, G., Fang, F. X., Keller, M., Edelstein, W., & Kehle, T. J. (2003). Social moral reasoning in Chinese children: A developmental study. *Psychology in the Schools*, *40*, 125–138.

Flavell, J. H. (1972). An analysis of cognitive-developmental sequences. *Genetic Psychology Monographs*, *86*, 279–350.

Keller, M. (1986). Freundschaft und Moral: Zur Entwicklung der moralischen Sensibilität in Beziehungen [Friendship and morality: Towards the development of moral sensibility in relationships]. In H. Bertram (Ed.), *Gesellschaftlicher Zwang und Moralische Autonomie* (pp. 195–223). Frankfurt am Main: Suhrkamp.

Keller, M. (1996). *Moralische Sensibilität: Entwicklung in Freundschaft und Familie [Moral Sensibility: Development in friendship and family]*. Weinhein: Beltz, PVU.

Keller, M., Eckensberger, L. H., & Rosen, v. K. (1989). A critical note on the conception of preconventional morality: The case of stage 2 in Kohlberg's theory. *International Journal of Behavioral Development*, *12*, 57–69.

Keller, M., & Edelstein, W. (1991). The development of socio-moral meaning making: Domains, categories and perspective-taking. In W. Kurtines, & J. Gewirtz (Eds.), *Handbook of moral behavior and development: vol. 2*, (pp. 89–114). Hillsdale, NJ: Erlbaum.

Keller, M., & Wood, P. (1989). Development of friendship reasoning: A study of interindividual differences in intraindividual change. *Developmental Psychology*, *25*, 820–826.

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

Krebs, D. L., & Denton, K. L. (2005). Toward a more pragmatic approach to morality: A critical evaluation of Kohlberg's model. *Psychological Review*, *112*, 629.

Krebs, D. L., Denton, K. L., Vermeulen, S. C., Carpendale, J. I., & Bush, A. J. (1991). Structural flexibility of moral judgment. *Journal of Personality and Social Psychology*, *61*, 1012–1023.

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

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models: vol. 144*, Thousand Oaks: Sage.

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*. London: Tavistock.

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

Rosen, v. K. (1987). *Structure and content aspects of moral reasoning: A qualitative analysis*. Berlin: Max Planck Institute for Human Development and Education.

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

Selman, R. L., & Jaquette, D. (1977). *The development of interpersonal awareness*. Cambridge, MA: Harvard-Judge Baker Social Reasoning Project.

Teo, T., Becker, G., & Edelstein, W. (1995). Variability in structured wholeness: Context factors in L. Kohlberg's data on the development of moral judgment. *Merrill-Palmer-Quarterly*, *41*, 381–393.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: One Parameter Logistic Model* (*Version 3.0*). Arnhem, www.cito.nl.

Walker, L. J., Gustafson, P., & Hennig, K. H. (2001). The Consolidation/Transition Model (CTM) in moral reasoning development. *Developmental Psychology*, *37*, 187–197.

Willet, 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.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modelling software*. Melbourne, Australia: Australian Council for Educational Research.