States of Moral Judgment Development: Applying Item Response Theory to Defining Issues Test Data

Thijs van den Enden – Utrecht University, The Netherlands
Jan Boom – Utrecht University, The Netherlands
Daniel Brugman – Utrecht University, The Netherlands
Stephen Thoma – University of Alabama

Citation of published version:
Enden, T., Boom, J., Brugman, D., Thoma, S. (2018): States of Moral Judgment Development: Applying Item Response Theory to Defining Issues Test Data. *Journal of Moral Education.* 48(4).

DOI: https://doi.org/10.1080/03057240.2018.1540973

JOURNAL OF MORAL EDUCATION
2019, VOL. 48, NO. 4, 423–438
https://doi.org/10.1080/03057240.2018.1540973

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Stages of moral judgment development: Applying item response theory to Defining Issues Test data

Thijs van den Enden [a*], Jan Boom [a], Daniel Brugman [a] and Stephen Thoma [b]

[a]Department of Developmental Psychology, Utrecht University, Utrecht, The Netherlands; [b]Department of Educational Psychology, University of Alabama, College of Education, Tuscaloosa, AL, USA

ABSTRACT

The Defining Issues Test (DIT) has been the dominant measure of moral development. The DIT has its roots in Kohlberg's original stage theory of moral judgment development and asks respondents to rank a set of stage typed statements in order of importance on six stories. However, the question to what extent the DIT-data match the underlying stage model was never addressed with a statistical model. Therefore, we applied item response theory (IRT) to a large data set (55,319 cases). We found that the *ordering* of the stages as extracted from the raw data fitted the ordering in the underlying stage model good. Furthermore, *difficulty* differences of stages across the stories were found and their magnitude and location were visualized. These findings are compatible with the notion of one latent moral developmental dimension and lend support to the hundreds of studies that have used the DIT-1 and by implication support the renewed DIT-2.

KEYWORDS

moral judgment; stage; Defining Issues Test (DIT); item response theory (IRT)

Throughout life people develop in their moral judgment ability, i.e., 'the capacity to make decisions and judgments which are moral (i.e., based on internal moral principles) and to act in accordance with such judgments' (Kohlberg, 1964, p. 425). The instrument used most often to measure moral judgment ability is the Defining Issues Test (DIT) (Rest, Cooper, Coder, Masanz, & Anderson, 1974). The present study investigates whether the data of a very large data set of the DIT match the underlying stage model of the instrument.

The DIT was originally based on Kohlberg's theory of moral judgment development (Kohlberg, 1958, 1969), in which people progress through six stages of moral judgment (see Table 1, based on Rest, 1979). Each stage represents a different way of thinking about moral issues. People are supposed to progress through these stages as they gradually shift from lower to more complex forms of moral reasoning (Thoma & Dong, 2014). According to Rest, Narvaez, Bebeau, and Thoma (1999) 'development is, in part, the more frequent and reliable use of higher stage thinking … using less of the lower stages and more of the higher stages of thinking' (pp. 55–56). In other words, the DIT reflects a probabilistic stage model in which the probability of reasoning according to a higher stage of moral judgment increases when moral judgment ability increases.

CONTACT Thijs van den Enden thijsvdenden@gmail.com Heidelberglaan 1, H221, Utrecht 3584CS, the Netherlands

*Present address: Movisie, Catharijnesingel 47, 3511 GC, Utrecht, the Netherlands

Table 1. Stages of moral judgment.

Stage	Description
1	Focus on justice in terms of power and obedience; the right thing to do is what leads to the least punishment.
2	Focus on direct or indirect satisfaction of one's own needs. Fulfilling the needs of others is based on principles of reciprocity; if I am good to someone else, this person will later also be good to me.
3	Focus on what is approved by (significant) others. Intentions become important in judging behavior and therefore role taking is necessary.
4	Focus on fulfilling one's duty in society, obeying rules and showing respect for authority; laws are justified by means of maintaining social order itself, not the underlying principles or rights they are to protect.
5	Focus on individual autonomy, rights, personal values and opinions. Laws have a social utility and can be changed by social contract.
6	Focus on abstract universal principles such as the categorical imperative. Although these can be derived from other sources, individuals actively have to choose to live upon such a principle themselves.

To assess an individual's stage of moral judgment, Rest and colleagues devised the DIT-1 (Rest, 1979, 1986; Rest et al., 1974); a multiple-choice questionnaire for individuals from 14 years and older. The DIT-1 consists of six moral issues (stories) followed by 12 possible considerations or statements (items) per issue, most of which represent a specific stage of moral judgment (although no statements were formulated for stage 1, because the subjects were considered too advanced for this stage). For every story, respondents *rate* the importance of all the items and subsequently *rank* which four items they find most important. These responses result in several outcome measures (see Rest et al., 1974).

As it is easy to administer, the DIT-1 became the most widely used instrument to measure moral judgment, with a database of more than 200.000 records. Furthermore, it has been used in about 1300 studies predicting various outcomes, such as participating in volunteering activities (Van Goethem et al., 2012), job performance (Duckett & Ryden, 1994) and delinquency (Stams et al., 2006).

Although the instrument was widely used, its language and dilemmas became somewhat outdated. For instance, one of the stories referred to the Vietnam War as if it was a current event. Therefore, in 1999, a newer version of the instrument was presented: the DIT-2 (Rest, Narvaez, Thoma, & Bebeau, 1999). This instrument contains five newly written stories, equivalent to five of the stories of the DIT-1. The sixth story (Webster; see the Method section for a summary of the DIT-1 stories) is not included in the DIT-2. According to the authors, 'we shortened the test … when we found that one dilemma of the DIT-1 was not contributing as much to validity as were the other dilemmas' (Rest et al., 1999, p. 647). Furthermore, the DIT-2 uses a new way of scoring (the N2-index; Rest, Thoma, Narvaez, & Bebeau, 1997) and introduces a few new checks for participant reliability. Although the DIT-2 uses schema theory to describe moral judgment development, the instrument is still based on stage-typed statements like the DIT-1 (Rest et al., 1999). The DIT-2 is still widely used in research (see e.g., Cáceda, Andrew James, Gutman, & Kilts, 2015; Corcoran & O'Flaherty, 2016; Sorensen, Miller, & Cabe, 2017).

Several studies were conducted over the years to investigate the reliability and validity of both the DIT-1 and the DIT-2 (see e.g., Rest et al., 1999; Thoma & Dong, 2014). However, it was never formally tested whether the data of these instruments actually match the underlying probabilistic stage model of moral judgment development. Furthermore, it was

questioned whether items representing the same stage of moral judgment on different issues are equally difficult. Specifically, Rest (1979) himself noted that the probability of rating or ranking items representing the same stage as important differed across the stories. For instance, respondents tended to prefer stage 4 items less often for the Heinz and the Drug dilemma than for the other dilemmas, i.e., it is 'harder' to reason according to stage 4 for this dilemma. Although this may indicate that moral judgment development differs systematically across stories, the magnitude of such systematic differences has never been modeled precisely. In the present study, we will do so by using the principles of item response theory (IRT).

Classical test theory and item response theory

For analyzing test data, two main psychometric frameworks can be used: classical test theory (CTT) and IRT. In short, CTT relies on a 'true score theory' in which an observed score on a test or item is composed of a 'true score' and a certain amount of measurement error. Techniques often used in the social sciences, such as regression analyses and analysis of variance, are based on CTT (Embretson & Reise, 2000). However, since estimates of item difficulty and respondent ability depend on each other (see e.g., Magno, 2009), CTT estimates are strongly sample dependent. That is, the estimated ability of a respondent depends upon the difficulty of the items and vice versa; each one must always be viewed in the context of the other. In contrast to CTT, IRT separately estimates the difficulty distribution of the items and the ability distribution of the respondents and uses a common scale for both (Embretson & Reise, 2000). These estimates are based on the answering *patterns* of the respondents. For instance, on a math test, all students will give the correct answer to the easiest question, while only the most able students will answer the most difficult question correctly. It is possible to determine which questions are most difficult and which respondents are most able by looking at these answering patterns. In such a way, the difficulty of the items and the ability of the respondents can be determined even when background characteristics such as age or education are unavailable.

Present study

To date, all published studies investigating the data of the DIT relied on methods derived from CTT. Because such methods do not consider the difficulty of items separately from sample characteristics, these studies can only provide indirect evidence supporting or challenging specific components of the presupposed stage model of moral development theory. By applying several multi-category models derived from IRT (described further in the Method section; see also Embretson & Reise, 2000), we were able to evaluate whether answering according to higher stages of moral judgment development is indeed more difficult than answering according to lower stages. To do so, we investigated whether the answering patterns of respondents on the different issues match the presupposed stage model of moral judgment development. More specifically, this study aimed to:

(1) investigate whether the implicit stage order in the data fits the assumed stage order on which both the DIT-1 and the DIT-2 are based (2-3-4-5-6);
(2) model possible differences in difficulty of the stages over the six stories of the DIT-1; and
(3) interpret these differences on the underlying scale of moral judgment ability (facilitated with graphical representations).

Such an examination and precise modelling of the concurrence between the data of the DIT and the theory underlying the DIT has not been performed before. However, such an examination is indispensable for the interpretation of the results of the hundreds of studies that are using and have used the DIT-1 and the renewed DIT-2, which is based on the same underlying stage model of moral judgment development.

Method

Participants

For this study, we used a very large data set from the DIT-1. Available for our analyses were all data collected between 1988 and 2014 by the University of Alabama's Centre for the Study of Ethical Development: (1) from the full six-story form of the DIT-1; (2) collected from native speakers of the English language (mostly American and Canadian); and (3) meeting the requirements of a 'purged' sample. The DIT manual recommends using purged samples for all statistical analyses, because they produce clearer and more meaningful results (Rest, 1986). The purged sample excludes those cases that do not pass all DIT validity checks[1] (Rest, 1986), which are designed to identify subjects who did not fill out the questionnaire seriously or understand its instructions properly. When less than 60% of the respondents of a sample passed the checks, the entire sample was excluded. In total, about 15% of all the cases collected by the University of Alabama's Centre for the Study of Ethical Development was excluded based on the validity checks, resulting in a data set of 55,319 cases available for analysis. No differences were found between 'purged' and accepted respondents in gender or race, but younger participants failed the validity checks at a slightly higher rate (Thoma, personal communication, 9 January 2018).

Although, regrettably, the purged sample contained no background variables,[2] some general background information concerning the total sample is known: It covers a full age range from young adolescence to adulthood (peaking around the college ages), is roughly balanced by gender and represents a broad segment of educational, religious, ethnic and occupational groups (Evens, 1993; Thoma, personal communication, 1 October 2014). Although the unavailability of background characteristics brings certain limitations (see Discussion section), this is not a major problem for the analyses, as stage difficulty is estimated separately from sample characteristics. We could not use a variety of IRT models (i.e., Rasch models; Bond & Fox, 2001; Embretson & Reise, 2000; Rasch, 1977) that claim complete separation between ability and difficulty estimates. However, our sample is very large and has a very large range in terms of age and educational level (and other respects), so we are confident our results are to a large extent sample independent.

Measures

The DIT-1 consists of six short moral stories; briefly:[3]

(1) Heinz and the Drug: Should Heinz steal a drug to save the life of his dying wife, if he cannot pay the unfairly high price the druggist is asking for it?
(2) Escaped Prisoner: Should a neighbor who recognizes an escaped convict turn him over to the police, even though he built up an honest life after his escape from prison?
(3) Newspaper: Is a high school principle right to stop publishing the school newspaper when he thinks that its content undermines school rules and principles?
(4) Doctor's Dilemma: Should a doctor give a terminally ill patient so many painkillers that she would die, if the patient requests him so?
(5) Webster: Is it justifiable for a business owner (Mr. Webster) to not hire someone from a racial minority, because he might lose customers who are racially biased?
(6) Student Take-Over: When the president of the university refuses to stop an army training program the university professors voted against, are students justified to take over a university building as a protest?

Every story is followed by 12 statements (items) concerning the specific moral issue. Of the 72 statements in total, 62 represent a certain stage of moral judgment (Stage 2: 5 items, Stage 3: 17 items, Stage 4: 19 items, Stage 5: 16 items, Stage 6: 5 items). The other 10 statements are either nonsensical (Meaningless) or reflect strong opinions against laws and social order (Anti-establishment), and were not taken into account in the present study. For every story, respondents first rated all 12 items on a 5-point scale ranging from 1 (*no importance*) to 5 (*great importance*) and subsequently ranked which four items they found most important. These ratings and rankings are the basis for several outcome measures, of which the P-score is most often used. A P-score is calculated on the basis of assigning points to ranking data: 4 points to the stage of the highest ranked item to 1 point to the stage of the fourth ranked item. The P-score is calculated by dividing the total number of points across the six dilemmas for post-conventional items (hence the 'P') by the total number of points (60 when there are no missings). P-scores can range from 0% to 95% (not 100% because not every dilemma has four possible postconventional items).

Determining one stage score per respondent per story

To investigate whether answering patterns of respondents on the six stories of the DIT are consistent with the underlying stage model of moral judgment development, we used multi-category IRT (Embretson & Reise, 2000) techniques, which have already been successfully used to draw conclusions about the underlying stage model of moral development on interview data (Boom, Wouters, & Keller, 2007; Dawson, 2002). As explained above, these techniques can be used even when no background variables are available because the estimates of item difficulty are estimated separately from the samples from which they are obtained. However, they require that respondents are classified as belonging to one stage of reasoning for each story. Unfortunately, the raw

DIT data consist of a set of ratings and rankings of stage-typed items, and not of assigned stage scores. The available indices (e.g., the P-score and the N2-index) were also insufficient for IRT, because these represent aggregated scores of multiple stages instead of one stage score per story. Therefore, we had to determine a stage score per respondent per story ourselves. A simple solution would be to choose the stage which was (generally) rated or ranked highest for a story. However, in this way we would discard information (other ratings and rankings), and not use the possibility to estimate stage scores more reliably by taking into account order information. Therefore, we decided to estimate one stage score per story for every respondent based on their answering *patterns*. Because the P-score, which is the most often used outcome index of the DIT-1, is based on the rankings, we also decided to base our stage score on the rankings of the respondents. The exact procedure followed will be described now. To start with, we assigned 4 points to the stage of the highest ranked item, 3 points to the next and so on up to 1 point to the stage of the fourth ranked item. This is the same procedure which is followed in the computation of the P-score. For instance, the following rankings [Rank 1: Stage 4 item, Rank 2: Stage 3 item, Rank 3: Stage 2 item, Rank 4: Stage 3 item] of a respondent on one of the stories would result in the point distribution [Stage 2: 2 points, Stage 3: 4 points, Stage 4: 4 points, Stage 5: 0 points, Stage 6: 0 points]. Then, we computed a so-called 'relative ranking' per stage per story for every respondent. Relative rankings were computed by dividing the total amount of points for a certain stage (e.g., stage 4) by the total amount of points for that story (10 when there are no missing data and when no Meaningless or Anti-establishment items were ranked). This resulted in a pattern of five relative rankings; one for each stage in the DIT-1. For instance, the point distribution above would result in the following in the relative rankings [Stage 2: 0.2, Stage 3: 0.4, Stage 4: 0.4, Stage 5: 0, Stage 6: 0]. Second, we determined the stage of moral judgment that matched this pattern of 5 relative rankings best. We did so by determining the 'peak' of the distribution of these relative rankings. Each pattern—one per story, six per respondent—was compared to multiple binomial distributions[4] using the statistical program R (R Core Team, 2013), to see which binomial distribution fitted the pattern of relative rankings best. Subsequently, the stage score closest to the peak of this best fitting binomial distribution was picked as someone's estimated stage of moral judgment for that story. Thus, every respondent was assigned one stage score per story—six in total as the DIT consists of six stories. In the abovementioned example, the respondent would be assigned to stage 3 for this story, because the peak of the best fitting binomial distribution was closest to this the score of 3. As could be seen from this (odd) example, this is not necessarily the stage of the item ranked as most important. This makes sense since, in this example, the items ranked second, third and fourth were all from lower developmental stages. The stage scores based on the rankings were found to be significantly related to the P-scores of the same story (range r = .671—.747, all $p < 0.01$).

Analyses

Item response theory

Following the procedure described above, we determined six stage scores per respondent (one per story). These were used as input for IRT models, which estimate the

probability of choosing a certain answering category as a function of the difficulty of that answering category (b) and the latent ability of the respondent (θ), both expressed on the same scale. For our data, this reflects the probability of responding according to a certain stage (i.e., ranking items from this stage and adjacent stages as most important) as a function of the difficulty of that stage (the higher the stage, the more difficult it is) and the moral judgment ability of the respondent (the higher the stages that the respondent prefers, the more able he/she is). For instance, the probability that someone with a low moral judgment ability answers according to stage 2 (a 'low' stage) is higher than that of a person with a high moral judgment ability answering according to stage 2, whereas this person is more likely to endorse a 'high' stage. The latent moral judgment ability of the respondents in our study correlated significantly and strongly with the P-score of the person over the DIT as a whole (r = .848, p < .001).

As an illustration, Figure 1 graphically depicts the probability distributions of the rankings of the Doctor's Dilemma story. In this figure, the (standardized) latent moral judgment ability of the respondents (θ) is plotted on the X-axis, and the probability of rating according to a certain stage is plotted on the Y-axis (e.g., with an increasing moral judgment ability, the probability of rating according to higher stages increases). The difficulty (b) of the stages is expressed by thresholds demarcating the corresponding probability curves.

Three IRT models for more than two categories

First, to investigate the stage ordering within the DIT as a whole and the consistency of this stage ordering over the different stories of the DIT, we compared three different IRT models. In our baseline model (1) the difficulty of the stages (b) was set to be the same across the six stories of the DIT; in a second model (2) the distances between the

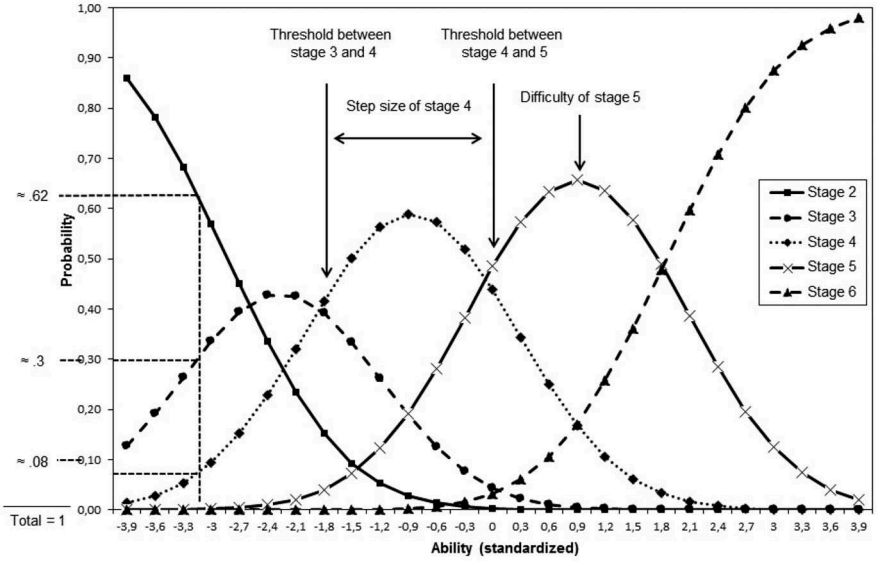

Figure 1. Probability curves of an unrestricted Graded Response Model based on the rankings of the items of the Doctor's Dilemma story.

difficulty of the stages (*b* stage x + 1 minus *b* stage x) was set to be equal across the stories; finally (3), we tested a model in which there were no constraints on difficulties of stages across the stories. In all models, the discrimination parameter (*a*) was set to be equal across the stories. In graphical terms, this refers to: (1) a model that looks exactly the same for all six stories of the DIT; (2) a model in which the distances between the curves were equal across the stories, but the whole set of curves could be shifted to the right or left of the figure per story; and (3) a model in which the size of the curves could also differ between the stories. Mplus version 7 (Muthén & Muthén, 2010) with a probit model and the Weighted Least Squares (WLSMV) as estimator was used for the IRT models. WLS is considered to be most reliable for categorical data (Brown, 2006). IRT-models in Mplus technically belong to the graded response model (GRM) framework (Samejima, 1969). Ostini and Nering (2006) discuss classification and terminology issues for polytomous GRM models. Our models 1 to 3 might be likened to normal ogive versions of a restricted Rating Scale GRM, a Rating Scale GRM and an unrestricted GRM (close to a generalized Partial Credit Model) respectively. See Boom et al. (2007) for a similar approach.

To interpret the results, we look at the (differences in) model fit for the different models. For evaluating model fit (comparative fit index [CFI], root mean square error of approximation [RMSEA]), we used criteria demarcated by Hu and Bentler (1999). A good model fit for any of the IRT-models would indicate that the proposed stage ordering (2-3-4-5-6) fits the DIT as a whole. Furthermore, a significant improvement in model fit (performed with chi-square difference tests) would indicate that a less restricted model fits the data better and that the difficulty of the stages differs systematically across the stories of the DIT. Finally, to interpret possible difficulty differences between stages within and between stories, we rely on a visualization of the probability curves of the stages over the six stories.

Furthermore, to investigate whether alternative stage orders would also—or better—fit the data, we first performed Homals analyses in SPSS (Meulman & Heiser, 2001) to identify which alternative stage orders might also fit the data (e.g., a reversal of stage 2 and 3). Subsequently, these alternative orders were examined by rerunning the procedures described in the subsections above, and comparing the resulting model fit with that of the final IRT model for the original stage order. For this comparison, we could only use absolute fit measures as the models were identical but the data (stage scores) were different. We chose the Bayesian Information Criterion (BIC) (Schwarz, 1978), which is well-suited for model comparison with large samples[5] (Aho, Derryberry, & Peterson, 2014). To obtain BIC, we had to use maximum likelihood (ML) estimation, which is acceptably reliable with five or more answering categories (Finney & DiStefano, 2006).

Results

Descriptives

There were huge differences in how often a certain item was ranked as important (ranging from 0.2% of the respondents [N = 86] ranking item 4 highest on the Heinz story to 22.5% [N = 12,472] ranking item 1 highest on the Webster story). Items

Table 2. Number of respondents assigned to a stage per story.

	Stage					
Story	2	3	4	5	6	Missing
Heinz and the Drug	5176	17,631	11,211	15,258	5981	62
Escaped prisoner	145	7269	29,090	18,558	164	93
Newspaper	1371	3088	23,054	26,741	1007	58
Doctor's Dilemma	229	4352	23,764	25,741	1146	87
Webster	1372	6528	32,621	14,127	592	79
Student Take-Over	849	5133	20,272	26,883	2070	112
Total	9142	44,001	140,012	127,308	10,960	491

representing stage 5 were ranked highest relatively most often, followed by items representing stage 4, 2, 6 and 3 (in that order). Table 2 indicates the number of respondents assigned to a specific stage for every story based on the rankings.

Item response theory models

Evaluation of stage ordering and difficulty differences across the stories

To see whether the stage ordering was correct and the difficulty of the stages was similar across the stories of the DIT, we compared the model fit of a restricted Rating Scale GRM, Rating Scale GRM and an unrestricted GRM. Model fits can be found in Table 3. The fit of subsequently less restricted models improved strongly and significantly. The improvement in model fit from the restricted Rating Scale GRM to the Rating Scale GRM (in which the distances between the curves were equal across the stories, but the whole set of curves could be shifted to the right or left of the figure per story) indicated that the items of different stories representing the same stage were not equally difficult. The improvement in fit from the Rating Scale GRM to the unrestricted GRM (in which the size of the curves could also vary across the stories) indicated that the step in ability between two stages (e.g., the distance between stage 3 and 5, hence the 'width' of stage 4) was also different across the stories. However, the fit of the least restricted model (unrestricted GRM) was good (with R^2 = .202), indicating that the proposed stage ordering (2-3-4-5-6) concurs with the DIT-data. Figure 2 depicts the curves of the final model.

In Figure 2, it can indeed be seen that the stage difficulty differs across the stories. Clearly, the probability curves of answering according to a certain stage of moral judgment on the six stories are not equal. However, the probability curves of the last three stories of the DIT (Doctor's Dilemma, Webster and Student Take-Over) are quite comparable. Furthermore, the differences of these stories in comparison with the Escaped Prisoner story can partially be explained by the fact that there are no items representing stage 2 in this story. Moreover, these four stories seem to discriminate quite well between the different stages of moral judgment development, as represented

Table 3. Fit statistics of item response theory models on stage scores of the DIT.

GRM Model	$\chi^2/\Delta\chi^2$	$df/\Delta df$	χ^2/df	CFI	RMSEA [90% CI]
Restricted Rating Scale	59,529.45***	34	1,750.87	.000	.178 [.177, .179]
Rating Scale	16,056.90***	5	3,211.38	.000	.165 [.163, .166]
Unrestricted	42,619.13***	15	2,841.28	.969	.033 [.031, .035]

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation, CI = confidence interval.
***$p < .001$.

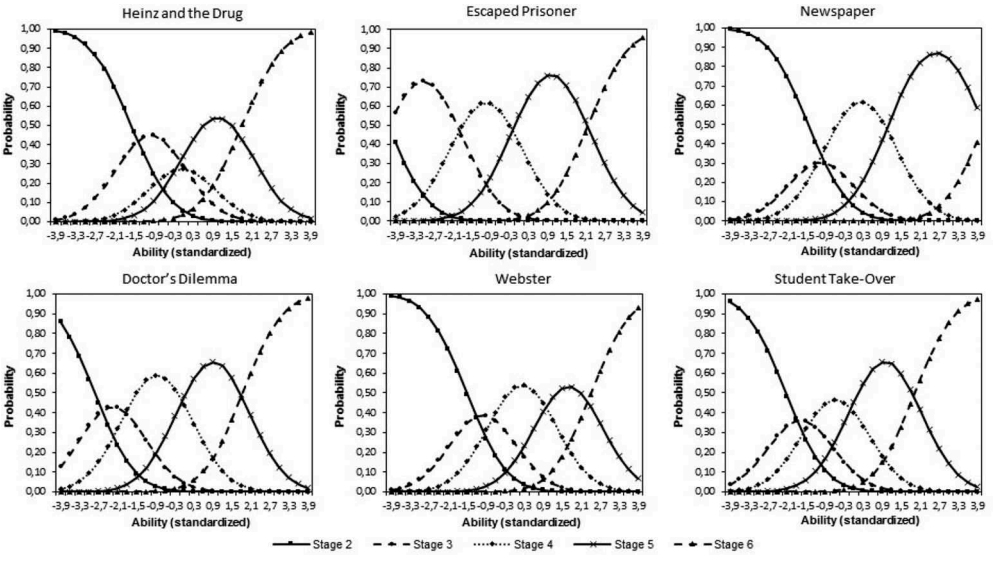

Figure 2. Probability curves of the unrestricted Graded Response Model based on the rankings of the items of the six stories of the DIT.

by the clear separation of the curves. It seems that the difficulty differences found in IRT analyses are especially due to problems with the discriminative power of two of the stories: Heinz and the Drug and Newspaper. More specifically, the Heinz and the Drug story is not able to distinguish stage 4 well from the other stages, while the Newspaper story is not able to distinguish stage 3 from the other stages.

Alternative stage orders

The Homals analyses suggested two alternative stage orders. The model fit (BIC) of these suggested alternative models is depicted in Table 4. A BIC-difference of 10 already indicates strong evidence in favor of the model with the lowest BIC-value (Raftery, 1995). As can be seen, both alternative stage orders fitted worse than the original model.

Table 4. Bayesian Information Criterion (BIC) of models with alternative stage orders.

Model	BIC
Original	727,312.03
3-2-4-5-6	762,379.41
2-3-***6-4-5***	780,680.64

Discussion

By applying multi-category IRT models on a very large data set of the DIT, we were able to evaluate the stage order of the underlying model of moral judgment development. In general, we found support for the concurrence between the assumed stage model of

moral judgment on the one hand, and the rankings of the stage-typed statements of the DIT on the other. Concerning the first aim of the study, we found that the proposed stage order fitted the data well. IRT techniques revealed that the fit of the final IRT model (unrestricted GRM, or generalized Partial Credit Model) was good. Additionally, a comparison of the proposed stage model with alternative stage models indicated that the proposed stage order was to be preferred over an alternative ordering of the stages. We also confirmed that the stage difficulty was not strictly consistent over the stories. A significant improvement in model fit of highly restricted to less restricted IRT-models indicated that the difficulty of the stages differed systematically across the stories; i.e., that the stages are not equally difficult in the different stories. This finding that moral judgment differs systematically across stories has been noted before (e.g., Rest, 1979) and is, as pointed out before, considered to be problematic for the interpretation of the results of the DIT, as this seems to imply that moral judgment abilities differ systematically according to the specific context (dilemma) that is provided. However, on the other hand, the IRT model is based on the assumption that each respondent is characterized by his or her latent ability, which is unique and the same over all six stories. This idea of a characteristic individual ability (moral judgment ability) is explicitly and unambiguously supported by our results, which seems to imply that there is a core moral judgment ability. Future research might perhaps find predictors for the systematic differences found between stories.

New in our approach is that we could disentangle the order of the stages (which fitted the data good) from difficulty differences (which were found), and visualize the magnitude and location of these differences by providing probability curves for each story separately but all sharing the same x-axis. The graphical representation in Figure 2 shows that it is harder to reason according to stage 4 for the Heinz and the Drug dilemma (already noted by Rest, 1979), and to stage 3 for the Newspaper dilemma (not noted before). In fact, these stages are never the most probable stage for any person estimate level; i.e., even for a person theoretically in stage 3, it would be more probable that this person would be placed in stage 2 or 4, based on the Newspaper story. This may also indicate ordering differences between the six stories of the DIT (maybe in the Newspaper story stage 3 is easier than stage 2), but we were not able to test these differences. Although for all stories, reasoning from all stages was used by at least some subjects, it would be interesting to investigate whether the consistency of the difficulty of the stages over the stories of the DIT could be improved by reformulating or removing items that were not considered to be important by most respondents. For now, we would argue to always use the full six-story version of the DIT-1 (or five-story version of the DIT-2) to find the most reliable results about the moral judgment ability of the respondents. In hindsight, it might be unfortunate that the Webster story, which is able to differentiate between different stages quite well in our analysis, was removed from the DIT-2. Our study indicates that it may differentiate better between respondents in different stages of development than the Heinz and the Drug or the Newspaper dilemma. With those stories, it is difficult to differentiate people in respectively stage 4 and 3 from people in adjacent stages.

Three kinds of strengths and limitations of the study should be noted. First, some are related to the data set used for analysis. A major strength is that our sample contains a large part of all DIT-1 questionnaires ever administered and covers a broad range of

people, making our results generalizable across a large part of the population of native speakers of the English language. Nevertheless, a major limitation is that our sample did not contain any background variables, i.e., the population cannot be delineated. Although we argue that it is a strength of the study that we could do the analyses despite the lack of age information, it would have been interesting to see how the stage ordering and difficulty differs with individual characteristics (e.g., age, education, region, religion or occupation) and how moral judgment ability estimated with IRT is related to other variables (e.g., prosocial behavior or political attitudes). Given the high correlation with the P-score this seems a promising line of further research. Furthermore, the sample also did not contain information about the year of administration of the measure. The validity of the measure might have changed over the cohorts.

Second, some strengths and limitations concern the DIT itself. The DIT is the most widely used and most respected measure of moral development. However, because the DIT contains no items representing stage 1, we were not able to test the whole stage model. Another problem, for the present study, was that the DIT offers no direct stage assignments to respondents. We could not use the available DIT-indices (P or N2-score) and had to rely on a transformation. This transformation was based on the idea that the distribution of relative rankings, per story per respondent, can be approached by a binomial distribution. One might argue that this presupposes that the stages are ordered in the way which we are testing in this study, giving rise to the objection of circularity. However, when we tested the fit of a model with a different stage order (with Homals analyses followed by IRT models), we applied the same principles already on an initial level with a less good fit as a result. More importantly, we would argue it is an important finding that—assuming the underlying model makes sense—the data fit this underlying model. Although this is perhaps not a fully objective test of the stage model, it does tell us much about the internal consistency of the DIT framework and instrument. We would argue that it is very informative for future research to investigate the internal consistency of other stage models of moral judgment development by applying the same techniques on data from other instruments, in particular for instruments that use more direct stage assignment. Similarly, it would be interesting to fit the moral schema model (Rest et al., 1999) on DIT-2 data.

Third, some strengths and limitations are due to characteristics of IRT. Because IRT models are cumulative, while the DIT is based on a preference model, we could not use the raw DIT items in IRT models. Therefore, we decided to use estimated stage scores based on the patterns of rankings, which are different from indices used in other validation studies. Nevertheless, the estimated stage scores did correlate strongly with the externally validated P-score (see Rest et al., 1999). Furthermore, the patterns of many participants did not clearly indicate one preferred stage per story (see also Cooper, 1972), inevitably resulting in a less reliable estimation of stage scores. However, unreliability of individual stage scores should be ruled out on the group level. Moreover, we showed that other stage orders fitted the data worse when the exact same procedures, being equally (un)reliable, were followed. Nevertheless, thanks to IRT analysis, the present study is unique in its ability to compare the data of the DIT with the assumed stage model of moral judgment development. To date, no study has evaluated this stage model as a whole and assessed differences between moral issues. For the first time, we were able to disentangle an evaluation of the ordering of the stages from differences in difficulty of the stages and visualize the magnitude and location of these differences. We hope this study to be part of a series (Boom et al., 2007; Dawson, 2002)

investigating the internal consistency of developmental stage models by looking at the concurrence between the ordering proposed by the theory and the ordering found in the data it produced. It might even be possible to make suggestions for refining theories or operationalizations within a framework, e.g., to devise a DIT-measure containing items and stories that show a better fit with the presupposed model. Further suggestions on improving the instrument could be derived by determining the complexity of the stage items using Fischer's Skill Theory (Fischer, 1980) or Commons and Richards' Model of Hierarchical Complexity (Commons et al., 2008, 2014).

In conclusion, this study lends support to the results of the hundreds of studies that were based on the DIT-1 and the renewed DIT-2 and to the underlying stage model, which was the most influential in thinking about moral development (Lapsley, 2006). The differential effect of stories on how likely it is to find certain moral issues important, confirms that a too strict stage concept should be replaced by the concept of a latent moral ability which is probabilistically related to stages. Practically, as some stories are less well able to discriminate between the different stories of the DIT, we argue that it is important to always use the full version of the DIT, and take possible story bias into account by interpreting the results of the hundreds of studies that use and have used the DIT-1 and the successive DIT-2 for investigating moral judgment, and also for new theory building on moral judgment in various contexts.

Notes

1. In the DIT, several items are included which are nonsensical (Meaningless items, e.g., 'Whether Heinz is a professional wrestler, or has considerable influence with professional wrestlers.'). The first validity check excludes respondents with total rank scores for Meaningless items higher than 8 (absolute) or 14% (relative)—indicating that the respondent evaluated items more on their complexity or pretentiousness than on their meaning. The second validity check excludes cases with extreme non-consistent scorings, i.e., who rank items as most important which were rated as unimportant and vice versa. The third validity check excludes respondents who rated more than 9 items the same (e.g., nine out of 12 items rated as 'not important') on more than two stories. In the final check, respondents who had missing data on more than 4 of the non-Meaningless items were excluded.
2. Background variables were systematically collected only from the introduction of the DIT-2 onwards.
3. Summarized by the first author. For the complete stories see Rest et al. (1974).
4. There are several reasons why the comparison with a binomial distribution is meaningful. First, binomial distributions only have one peak, just as we expect the distribution of relative rankings to have one peak (the 'modal' stage). Second, binomial distributions have a certain upper and lower boundary, just as the stages have a minimum of (stage) 2 and a maximum of (stage) 6. Third, the binomial distribution is well suited for small N, which matches the small number of scoring options (5 different stages). Finally, one can shift the curve of the binomial distribution by changing one parameter (the π), which makes it easy to determine the best fitting distribution. In this case, the distributions of rankings were compared with 101 binomial distributions, with π varying from 0.00 to 1.00 in steps of 0.01. The best fitting distribution was determined based on the lowest squared error between both.
5. Nevertheless, results were equal when Akaike Information Criterion (AIC) was used.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was conducted at Utrecht University.

Notes on contributors

Thijs van den Enden is a recently graduated research master student, focused on child and adolescent (moral) development.

Jan Boom is Associate Professor at the Department of Developmental Psychology at the Universiteit Utrecht. His long time interest has been in trying to understand development with a focus on cognitive (Piaget) and moral (Kohlberg) development. Originally his research was focused on the conceptual level; more recently the focus has been broadened to include the methodological and statistical level.

Daniel Brugman is Emeritus Professor in Developmental Aspects of Interventions in Youth Care at the Department of Psychology at Utrecht University, The Netherlands.

Stephen Thoma is University Professor and Program Coordinator of Educational Psychology at the University of Alabama. His specialty area is personality and social development in late adolescence and youth with a focus on moral judgment development.

ORCID

Thijs van den Enden http://orcid.org/0000-0001-9690-4933
Jan Boom http://orcid.org/0000-0003-4625-042X
Daniel Brugman http://orcid.org/0000-0002-0697-7678
Stephen Thoma http://orcid.org/0000-0002-1559-9920

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*, 631–636.

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

Boom, J., Wouters, H., & Keller, M. (2007). A cross-cultural validation of stage development: A Rasch re-analysis of longitudinal socio-moral reasoning data. *Cognitive Development*, *22*, 213–229.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Cáceda, R., Andrew James, G., Gutman, D. A., & Kilts, C. D. (2015). Organization of intrinsic functional brain connectivity predicts decisions to reciprocate social behavior. *Behavioural Brain Research*, *292*, 478–483.

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K., & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, *9*, 182–199. Retrieved from https://dareassociation.org/documents/Commons%20URM.pdf

Commons, M. L., Li, E. Y., Richardson, A. M., Gane-McCalla, R., Barker, C. D., & Tuladhar, C. T. (2014). Does the model of hierarchical complexity produce significant gaps between orders and are the orders equally spaced? *Journal of Applied Measurement*, *15*, 1–29.

Cooper, D. (1972). *The analysis of an objective measure of moral development* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.
Corcoran, R. P., & O'Flaherty, J. (2016). Examining the impact of prior academic achievement on moral reasoning development among college students: A growth curve analysis. *Journal of Moral Education*, *45*, 433–448.
Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral judgement stages revisited. *International Journal of Behavioral Development*, *26*, 154–166.
Duckett, L. J., & Ryden, M. B. (1994). Education for ethical nursing practice. In J. R. Rest & D. Narvaez (Eds.), *Moral development in the professions: Psychology and applied ethics* (pp. 51–70). Hillsdale, NJ: Erlbaum.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Evens, J. (1993). *Indexing moral judgment using multidimensional scaling* (Unpublished doctoral dissertation proposal). University of Minnesota, Minneapolis.
Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.
Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*, 477–531.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.
Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years ten to sixteen* (Unpublished doctoral dissertation). University of Chicago, Chicago.
Kohlberg, L. (1964). Development of moral character and moral ideology. In M. L. Hoffman & L. W. Hoffman (Eds.), *Review of child development research* (Vol. I, pp. 381–431). New York, NY: Russel Sage Foundation.
Kohlberg, L. (1969). Stage and sequence: The cognitive developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). New York, NY: Rand McNally & Company.
Lapsley, D. K. (2006). Moral stage theory. In J. Smetana & M. Killen (Eds.), *Handbook of moral development* (pp. 37–66). Mahwah, NJ: Erlbaum.
Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*, 1–11.
Meulman, J., & Heiser, W. (2001). *SPSS categories 11.0*. Chicago, IL: SPSS Inc.
Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angelos, CA: Muthén & Muthén.
Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.
R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.
Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements in symposium on scientific objectivity, Vedbaek, Mau 14-16, 1976. *Danish Year-Book of Philosophy*, *14*, 58–94. Retrieved from http://www.rasch.org/memo18.htm
Rest, J., Thoma, S. J., Narvaez, D., & Bebeau, M. J. (1997). Alchemy and beyond: Indexing the Defining Issues Test. *Journal of Educational Psychology*, *89*, 498–507.
Rest, J. R. (1979). *Development in judging moral issues*. Minneapolis, MN: University of Minnesota Press.
Rest, J. R. (1986). *DIT manual: Manual for the Defining Issues Test* (3rd ed.). Minneapolis, MN: Center for the Study of Ethical Development, University of Minnesota.

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

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

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology*, *91*, 644–659.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph no. 17*. Richmond, VA: Psychometric Society.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sorensen, D. P., Miller, S. E., & Cabe, K. L. (2017). Developing and measuring the impact of an accounting ethics course that is based on the moral philosophy of Adam Smith. *Journal of Business Ethics*, *140*, 175–191.

Stams, G. J., Brugman, D., Deković, M., van Rosmalen, L., van der Laan, P., & Gibbs, J. C. (2006). The moral judgment of juvenile delinquents: A meta-analysis. *Journal of Abnormal Child Psychology*, *34*, 697–713.

Thoma, S. J., & Dong, Y. (2014). The Defining Issues Test of moral judgment development. *Behavioral Development Bulletin*, *19*, 55–61.

Van Goethem, A. A. J., Van Hoof, A., Van Aken, M. A. G., Raaijmakers, Q. A. W., Boom, J., & Orobio de Castro, B. (2012). The role of adolescents' morality and identity in volunteering. Age and gender differences in a process model. *Journal of Adolescence*, *35*, 509–520.