ELSEVIER

Cognitive Development 18 (2003) 61–78

COGNITIVE DEVELOPMENT

Domain-general and domain-specific developmental assessments: do they measure the same thing?

Theo L. Dawson [a,*], YiYu Xie [b], Mark Wilson [b]

[a] *Cognitive Science, Hampshire College, Amherst, MA 01020, USA*
[b] *Graduate School of Education, University of California at Berkeley, Berkeley, CA, USA*

Received 1 May 2002; received in revised form 1 October 2002; accepted 1 October 2002

Abstract

The concept of epistemological development is useful in psychological assessment only insofar as instruments can be designed to measure it consistently, reliably, and without bias. In the psychosocial domain, most traditional stage assessment systems rely on a process of matching concepts in a scoring manual generated from a limited number of construction cases, and thus suffer from bias introduced by an over-dependence on particular content. In contrast, the Hierarchical Complexity Scoring System (HCSS) employs criteria for assessing the hierarchical complexity of texts that are independent of specific conceptual content. This paper examines whether the HCSS and a conventional stage assessment system, Kohlberg's Standard Issue Scoring System (SISS), measure the same dimension of performance. We scored 378 moral judgment interviews with both scoring systems. We then conducted a multidimensional partial credit analysis to determine the extent to which the two scoring systems assess the same dimension of performance. The disattenuated correlation between performance estimates on the SISS and HCSS is .92. Based on this and other evidence, we conclude that a single latent trait — hierarchical complexity — is the predominant dimension assessed by the two systems.

Keywords: Cognitive development; Stage theory; Hierarchical complexity; Developmental assessment; Multidimensional partial credit model; Random coefficients multinomial logit model; Linear transformation

* Corresponding author. Tel.: +1-413-247-9115.
E-mail address: tdawson@hampshire.edu (T.L. Dawson).

0885-2014/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0885-2014(02)00162-4

During the latter half of the 20th century a number of developmental stage assessment systems were introduced (Armon, 1984; Colby & Kohlberg, 1987b; Kitchener & King, 1990; Rest, 1975). These scoring systems are based on cognitive developmental stage theories that are grounded in the notion of hierarchical integration (Piaget, 1985). The stages, referred to here as *orders of hierarchical complexity* (*complexity orders*), represent a series of systematic reorganizations of thought structures.

Most stage-scoring systems are domain specific, in the sense that they are designed to assess developmental stage in a single content domain. For example, Kohlberg's Standard Issue Scoring System (SISS; Colby & Kohlberg, 1987a, 1987b) was developed to measure cognitive development in the moral domain. However, efforts have also been made to establish a general method of assessment that focuses specifically on the hierarchical complexity of performance and thus can be applied in more than one domain (Case, 1985; Commons, Richards, Ruf, Armstrong-Roche, & Bretzius, 1984; Fischer, 1980). The Hierarchical Complexity Scoring System (HCSS; Commons, Danaher, Miller, & Dawson, 2000; Dawson, 2002a; Dawson, Commons, Wilson, & Xie, 1999), which is based on Commons' General Stage Model (Commons, Trudeau, Stein, Richards, & Krause, 1998) and Fischer's (1980) Skill Theory, is one such system.

In this paper, we employ the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) to compare the SISS and HCSS as developmental assessment systems. We address two main questions: (1) Do domain-based and generalized scoring systems measure the same dimension of performance? (2) Do these scoring systems produce results that are in keeping with the postulates of developmental stage theory?

The notion of developmental stages leads to certain empirical expectations. First of all, developmental stages are built one upon the other in the sense that the construction of a subsequent stage requires the elements and operations of the previous stage. Consequently, each is logically more difficult than its predecessor. This means that development should proceed from one stage to the next in an invariant sequence with no skipping. There is a large body of longitudinal evidence supporting sequentiality in the acquisition of stages of development (Armon & Dawson, 1997; Case, Okamoto, Henderson, & McKeough, 1993; Colby, Kohlberg, Gibbs, & Lieberman, 1983; Dawson, 1997; Fischer & Bullock, 1981; Kitchener, King, Wood, & Davison, 1989; Snarey, Reimer, & Kohlberg, 1985; Walker, 1982).

However, evidence of sequentiality does not provide adequate support for the existence of developmental stages. Each complexity order in a hierarchical complexity sequence is generally defined by a set of internally consistent formal properties. In Piaget's model, these are said to constitute a *structure d'ensemble*, or *structured wholeness*. In more recent formulations of developmental stages, the processes and elements of a given stage are more often spoken of as the processes and elements of a dynamic, complex system, and stage change is thought of as the transformation of a system of this kind into another that is more hierarchically complex (Fischer & Bidell, 1998; Stevens, 2000; van Geert, 2000). This transformative

process leads to the expectation that, at least within individual knowledge domains, reasoning on familiar problems should either be transitional from one stage of reasoning to the next or predominantly at a single stage. Evidence consistent with this pattern is more commonly found when assessments are based on domain-general stage criteria than when they are based on domain-specific stage criteria (Dawson, 1998, 2002c; Dawson, Commons, & Wilson, submitted for publication).

1. The SISS

Kohlberg's SISS (Colby & Kohlberg, 1987b) is one of the best known stage-scoring systems. Kohlberg and his colleagues employed what they called a *bootstrapping* process to define moral judgment stages and construct a series of increasingly reliable scoring systems. Their final scoring system, the SISS, was constructed by analyzing seven sets of interviews from Kohlberg's original longitudinal study. Each of these sets of interviews included assessments from six test times, separated by 4-year intervals. Each performance was initially assigned a global stage score employing criteria from an earlier scoring system. Then, individual responses to each dilemma provided the basis for examples (criterion judgments) in the scoring manual.

The Standard Issue Moral Judgment Interview is made up of three forms, each of which is composed of three hypothetical moral dilemmas and several standard questions per dilemma. Each question is designed to probe respondents' reasoning on one or more of six moral issues — life, law, conscience, punishment, contract, and authority.

To assess moral development with the SISS, the researcher administers a set of moral judgment interviews, transcribes these, identifies each moral argument addressing one of six moral themes (life, law, conscience, punishment, authority, and contract), then employs the scoring manual to match each identified argument with a criterion judgment in the manual. These criterion judgments are intended to be structural in the sense that they reflect a particular socio-moral perspective or operative level, but they are expressed in terms of the content of interviews in the construction sample.

Kohlberg and his colleagues (Kohlberg & Candee, 1984) describe three periods of development in the moral domain: *preconventional*, *conventional*, and *postconventional*. Each of these three periods is subdivided into two stages so that Kohlberg's model comprises six stages of moral development. The Standard Issue Scoring Manual (SISM; Colby & Kohlberg, 1987b) provides scoring criteria for stages 1–5.

A recent scaling analysis of 996 moral judgment interviews scored with the SISS provides evidence for the specified developmental sequence and some evidence of structured wholeness in development from moral stages 3–5 (Dawson, 2002b). The instrument did not perform as well at the lower stages as it did at the higher stages. The failure of the SISS adequately to assess the lower stages has been

addressed by several researchers (Damon, 1977; Dawson, 2002b, in press; Keller, Eckensberger, & von Rosen, 1989; Kuhn, 1976).

2. The HCSS

An alternative approach to investigating the development of reasoning is to apply a generalized method of assessment to the hierarchical complexity of performances. The HCSS lays out explicit general criteria for determining developmental stage of performance in any domain of knowledge. Dawson's version of the HCSS (Dawson, 2002a), employed here, provides scoring criteria for eight complexity orders: (1) sensorimotor (Commons' sentential stage), (2) single representations (Commons' preoperational stage), (3) representational mappings (Commons' primary stage), (4) representational systems (Commons' concrete stage), (5) single abstractions (Commons' abstract stage), (6) abstract mappings (Commons' formal stage), (7) abstract systems (Commons' systematic stage), and (8) single principles/axioms (Commons' metasystematic stage).

When assessing the hierarchical complexity of a text the rater attends to two manifestations of hierarchical complexity. The first is the hierarchical order of abstraction of the concepts employed in its arguments, and the second is the most complex logical structure of its arguments. Appendix A briefly describes the six complexity orders identified in the data employed in the analyses that follow. For more detailed descriptions of these complexity orders, see Dawson and Gabrielian (in press).

The considerable differences between the HCSS and the SISS raise the question of whether the HCSS and SISS predominantly assess the same dimension of performance. Existing evidence indicates that orders of hierarchical complexity, as identified with the HCSS, are the same latent dimension of ability assessed with the SISS (Dawson, 2002c). However, this paper presents the first direct comparison of the two different scoring systems.

3. Method

3.1. Data

For the purpose of a direct comparison, a random sample of 378 interviews was selected from a large Kohlbergian moral reasoning database and rescored with the HCSS. These interviews were collected between 1955 and 1990 by five different groups of researchers (Armon & Dawson, 1997; Berkowitz, Guerra, & Nucci, 1991; Colby et al., 1983; Ullian, 1977; Walker, 1989). The ages of participants range from 5 to 86 and are distributed as shown in Table 1. The population sampled is diverse, representing a wide range of socioeconomic and ethnic groups from Berkowitz's sample of working class children and their parents to Kohlberg's

Table 1
Age distribution by study

Age	Study					Total count
	Armon	Berkowitz	Kohlberg	Ullian	Walker	
6	0	0	0	6	1	7
7–8	2	0	0	7	7	16
9–10	1	0	9	5	4	19
11–12	5	0	0	0	2	7
13–14	4	5	17	0	2	28
15–16	3	18	6	0	6	33
17–18	5	22	14	0	4	45
19–20	5	11	8	0	0	24
21–25	6	0	15	0	0	21
26–30	2	0	11	0	1	14
31–35	10	6	8	0	1	25
36–40	8	28	4	0	13	53
41–45	5	35	0	0	6	46
46–50	1	11	0	0	6	18
51–55	1	4	0	0	4	9
56–86	10	0	0	0	3	13
Total	68	140	92	18	60	378

sample of school-boys. It is not possible to report a consistent account of these, however, because discrepant reporting methods were employed in the various studies (see individual studies for details). All interviews were recorded and transcribed as specified in the SISM. Respondents received either Form A, Form B, or Forms A, B, and C of the SISS, depending on the study in which they participated.

All interviews were scored as specified in Kohlberg's SISM (Colby & Kohlberg, 1987b) by the original research teams. The system assigns stage scores in half-stage increments across six moral issues — life, law, conscience, punishment, contract, and authority — as follows: First, each time an issue is identified, complete with a judgment and a justification for that judgment, it is given a stage score. Then, a weighted average score (weighted toward the higher stage responses) for each issue is calculated from the accumulated stage scores on each issue as described in Colby and Kohlberg (1987a). Unfortunately, these weighted average scores obscure some of the variation within performances, but we were unable to obtain the raw scores for most of the datasets. Thus, there are at most six scores for each individual (one for each of the issues). Full and half stages 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 were identified for each of the moral issues.

Overall, after examining test–retest, alternate form, and inter-rater reliability for the SISS, Colby and Kohlberg (1987b) concluded that Standard Issue scores are accurate within about one-third of a stage, with inter-rater correlations of .92–.98, and alternate form correlations of .83–.98.

Hierarchical complexity scoring was conducted by the first author as described in Dawson (2002a). The text segments (protocols) scored by the hierarchical com-

plexity rater were responses up to 19 standard probe questions posed by interviewers administering Form A (the Heinz and Joe dilemmas) of the MJI. Consequently, the scoring unit for the HCSS is different from the scoring unit for the SISS in several ways. First, only the Joe and Heinz interviews were scored with the HCSS. Second, the scoring was not issue scoring. Rather, the scoring units were the complete responses to standard probes, with one score awarded for each response. Third, no half-stage scores were awarded. All protocols received a full-stage score. If a performance appeared to be transitional between complexity orders, the higher complexity order score was awarded. We award the higher score rather than the lower score, because verbal performances of this kind tend to lag behind competence (Fischer & Bidell, 1998; Kitchener, Lynch, Fischer, & Wood, 1993), so that any evidence of the structures of a given complexity order are likely to reflect some competence at that complexity order.

Scoring with the HCSS involves identifying both the highest hierarchical order of abstraction (HOA) and most complex form of logic in text performances. A protocol is considered to be at a given complexity order if its elements embody the hierarchical order of abstraction of that complexity order, and the complexity of its logical structure meet the formal requirements of that complexity order. For example, a child might say, "It is worse for a father to break a promise (than a son) because he is older and knows not to lie." The order of abstraction here is second order representations — *promise* and *lie*.[1] The logical structure is a concrete system — if father is both older and knows not to lie, it is worse for him to break a promise than a son who is younger and may not know not to lie.

In these data, six complexity orders were identified: representational mappings, representational systems, single abstractions, abstract mappings, abstract systems, and single principles/axioms. Ideally, a protocol should represent a complete argument on a given topic. Fragmentary arguments are usually treated as unscorable, because they tend to be down-scored (scored at a lower complexity order than the modal complexity order of the performance). However, because this results in loss of data we chose to score fragmentary protocols if adjacent protocols in a given text provided enough information to aid in their interpretation. This meant that the rater had access to the entire interview when scoring. This is an accepted practice in this type of research (Armon, 1984; Colby & Kohlberg, 1987a).

Correlations among scores of four independent raters on a subset of 112 randomly selected cases ranged from .95 to .98. Agreement rates ranged from 80 to 97% within half a complexity order and from 98 to 100% within a full complexity order. This equals or exceeds inter-rater agreements commonly reported in this field (Armon, 1984; Colby & Kohlberg, 1987a). Table 2 shows the expected

[1] *Promise* and *lie* are second order representations because their underlying concepts constitute arguments about relations between first order representations. For example, for a child to construct the notion of a lie as an intentional untruth, the child has to coordinate conceptions of *true versus not true* with intention. Notions of truth and intention (*on purpose*) appear for the first time at the preoperational complexity order.

Table 2
The relationship between complexity orders and Kohlbergian moral stages

Moral stage	Complexity order
1	Representational mappings (primary)
1.5	
2	Representational systems (concrete)
2.5	Single abstractions (abstract)
3	Abstract mappings (formal)
3.5	
4	Abstract systems (systematic)
4.5	
5	Single principles/axioms (metasystematic)

relationship between HCSS and SISS stages as determined by Dawson (in press) in her comparison of scoring criteria across the two systems.

3.2. Model

Members of the Rasch family of item response models were employed. Unidimensional and multidimensional partial credit analyses were conducted with *Conquest* (Wu, Adams, & Wilson, 1998) to compare the two sets of scores. Rasch models are increasingly employed in analyzing cognitive developmental data (Bond & Bunting, 1995; Bond & Fox, 2001; Dawson, 1998, 2002b; Dawson et al., submitted for publication; Demetriou, Efklides, Papadaki, Papantoniou, & Economou, 1993; Draney, 1996; Müller, Sokol, & Overton, 1999). These models are designed specifically to examine hierarchies of person and item performance, displaying both person proficiency and item difficulty estimates along a single interval scale (logit scale) under a probabilistic function. In addition, they can be employed to test the extent to which items or scores conform to a theoretically specified hierarchical sequence. A central tenet of stage theory is that cognitive abilities develop in a specified sequence, making the statistical tests implemented in a Rasch analysis especially relevant to understanding stage data. The detailed information about item functioning and individual performances provided by the software makes it possible to simultaneously examine group and individual effects.

The underlying model of the *Conquest* software is the random coefficients multinomial logit (RCML) model (Adams & Wilson, 1996). The RCML model is a generalized Rasch model that provides the flexibility of customizing models for particular test situations. Suppose θ is the latent variable and I is the total number of items, the probability of a response in category j of item i is modeled as

$$P(X_i = j; A, \vec{b}_i, \vec{\xi} \mid \theta) = \frac{\exp(b_{ij}\theta + \vec{a}'_{ij}\vec{\xi})}{\sum_{k=1}^{K_i}\exp(b_{ik}\theta + \vec{a}'_{ik}\vec{\xi})}, \tag{1}$$

where K_i is the total number of response categories in item i; $A = (\vec{a}'_{11}, \ldots, \vec{a}'_{1K_1}, \vec{a}'_{21}, \ldots, \vec{a}'_{2K_2}, \vec{a}'_{i1}, \ldots, \vec{a}'_{iK_i})'$, a design matrix of p columns, $\vec{a}_{ik}$; a

design vector in matrix A for $i = 1, \ldots, I, k = 1, \ldots, K_i$; $\vec{b}_i = (b_{i1}, \ldots, b_{iK_i})'$, a score vector of the response category from 1 to K_i for item I; and $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)'$, a vector of p free item parameters.

The score vector b_i provides the flexibility of non one-to-one mapping between the category and the score that is allocated to that category. It can be collected into a large vector $\vec{b} = (\vec{b}_{11}, \ldots, \vec{b}_{1K_1}, \vec{b}_{21}, \ldots, \vec{b}_{2K_2}, \ldots, \vec{b}_{i1}, \ldots, \vec{b}_{iK_i})'$, which allows different numbers of categories for different items. The vector of free parameters $\vec{\xi}$ and the design vector $\vec{a}_{ik}$, which is a linear combination of vector $\vec{\xi}$ determine how the model is specified and the design vector affords the possibilities of specifying customized models.

By extending the single latent variable to a D-dimensional latent space and collecting θ into a vector $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$, the following is the MRCML model:

$$P(X_i = j; A, B_i, \vec{\xi} \mid \vec{\theta}) = \frac{\exp(\vec{b}_{ij}' \vec{\theta} + \vec{a}_{ij}' \vec{\xi})}{\sum_{k=1}^{K_i} \exp(\vec{b}_{ik}' \vec{\theta} + \vec{a}_{ik}' \vec{\xi})}. \tag{2}$$

The scoring vectors $\vec{b}_{ik} = (b_{ik1}, b_{ik2}, \ldots, b_{ikD})'$ can be collected into a scoring matrix $B_i = (\vec{b}_{i1}', \vec{b}_{i2}', \ldots, \vec{b}_{iK_i}')'$ for item i. The distinction between Eqs. (2) and (1) is that b_{ij} and θ are scalars in Eq. (1) whereas these are vectors in Eq. (2).

4. Analysis

Item frequency statistics show that for some of the moral issues, there are no or low responses in some stages, usually the two lowest and sometimes the highest stages. Some recoding was done before the data were read into *Conquest*. A whole digit number was assigned to every possible response category starting from 0. In order to facilitate the analysis some recoding of the data was done before they were read into Conquest. Table 3 shows the recoding schema. A whole digit number was assigned to every possible response category starting from 0. Item frequency statistics show that for some of the moral issues, there are no or low responses in some stages, usually the two lowest and sometimes the highest stages. Because estimation of stage parameters cannot be made accurately if only one or two respondents have been awarded a given score, a decision was made to recode those stage scores as missing. When an item had missing data at the lowest stages, scores for the higher stages were shifted downward accordingly.

As mentioned above, the advantage of using the scoring matrix B in the MRCML model is that it allows different numbers of categories for different items. This is exactly the case in the recoded dataset. For example, the moral issue of life scored with SISS has only one response at stage 1. After recoding, there is no stage 1 and 0 represents the transitional stage from 1 to 2 and so forth. By applying the scoring scalar (1 2 3 4 5 6 7 8) to the observed scalar (0 1 2 3 4 5 6 7) for that one item, the interpretation of stages is

Table 3
Stage score recoding scheme

SISS		HCSS	
Original	Recoded	Original	Recoded
1	0	Representational mappings	0
1.5	1		
2	2	Representational systems	1
2.5	3	Single abstractions	2
3	4	Abstract mappings	3
3.5	5		
4	6	Abstract systems	4
4.5	7		
5	8	Single principles/axioms	5

restored and consistent across items. This two-step procedure of recoding data and using the score matrix solves the problem of missing categories in responses, which is a common phenomenon in cognitive developmental interview datasets.

The multidimensional partial credit model is implemented by specifying the score matrix B and the design matrix A. We assume that the two scoring systems measure two dimensions of latent performance in order to test the extent to which the two dimensions are related. Thus, the score scalars for the six moral issues scored with the SISS are related to the first dimension and the score scalars for the 19 questions from the Joe and Heinz dilemmas scored with HCSS are related to the second dimension. This is called a between-item multidimensional model since every item is related to single dimension (Adams et al., 1997). The partial credit model assumes that step difficulties of moving from one response category to another vary across the items and this is specified in the design matrix. It was necessary to employ a partial credit rather than a rating scale model, which assumes step difficulties to be constant across items, because there are missing lower categories for some items and the number of steps is not the same from item to item.

4.1. Multidimensional result

In this analysis, the mean item difficulty for each dimension is set to 0 as an identifying constraint. Altogether, 124 parameters were estimated, including 2 and 3 parameters for the mean and variance–covariance matrix of the person distributions, respectively, and 119 for item location and step difficulties. The correlation, which is disattenuated (Adams et al., 1997), between the two dimensions calculated from the variance–covariance matrix is .92, indicating that the two scoring systems, to a large extent, assess the same dimension of performance. The estimated means of the person distributions on the two dimensions are 0.055 and 0.311,

and the estimated variances are 3.397 and 8.627, respectively. This model has a likelihood deviance (−2 loglikelihood, G^2) of 8881.918. Since the two scoring system have different scales and the items are estimated on single dimension only, there is little meaning in comparing the means and spreads of the two person distributions without any direct link between the scales.

4.2. Unidimensional result

A unidimensional partial credit analysis was also conducted. In this analysis, the mean of all 25 items (6 with SISS and 19 with HCSS) is set to 0. A total of 122 parameters were estimated, including one for the mean, one for the variance of the person distribution, and 120 item parameters. The estimated mean and variance of the person distribution are 0.161 and 6.126. This model has a likelihood deviance (−2 loglikelihood, G^2) of 9122.579. We can compare the fit of this model to that of the previous one, using the change in the likelihood ratio χ^2. The difference in the likelihood deviance between the unidimensional and multidimensional models is 240.66 and the difference in the number of parameters is 2. Distributed as a chi-square with 2 degrees of freedom, the difference in the deviance is statistically significant at .01 level. This indicates that the multidimensional model fits the data better than the unidimensional one, even though the correlation between the two scoring systems is extremely high.

Now that we have gathered some evidence that the two-dimensional model explains a statistically significantly greater amount of variance in the data than the unidimensional model, we would like to examine in detail the person and item performances under the multidimensional analysis. However, there is one problem left to be solved. An equating method between the two scoring systems should be established, otherwise, no direct comparison can be made between the results obtained on different scales. Recall that the SISS has a scale of 9 points whereas the HCSS has a scale of 6 points. It is possible to recode the data on a single scale (either 9 or 6) and conduct another multidimensional analysis, but since we also have the unidimensional results in which all the parameters are estimated on a common scale, we decided to use that scale to rescale the multidimensional parameter estimates.

4.3. Rescaled result

Twenty-five item location parameters in the unidimensional analysis were constrained to have a mean of 0. The estimated parameters were divided into groups by their scoring method, 6 in one group and 19 in the other. Means (μ_{uni}) and standard deviations (σ_{uni}) were calculated for each group. In the multidimensional analysis, the item location parameters were already separated by their dimension and constrained to have a mean of 0 on each dimension. Thus, after calculating the standard deviance (σ_{multi}) of these estimated parameters for the two dimensions, a linear transformation was performed to obtain rescaled parameter estimates. The

Table 4
Estimated means (S.E.) of the person distributions

Analysis	Dimension 1 (SISS)	Dimension 2 (HCSS)
Multidimensional analysis	0.055 (1.84)	0.311 (2.94)
Rescaled analysis	0.272 (2.50)	0.228 (2.67)

following is the transformation formula:

$$\xi_{\text{rescaled}} = \xi_{\text{multi}} \left(\frac{\sigma_{\text{uni}}}{\sigma_{\text{multi}}} \right) + \mu_{\text{uni}}. \tag{3}$$

The step parameters were rescaled using the same transformation.

A multidimensional analysis was thus performed by anchoring the item parameters at the transformed values. The correlation between the two dimensions is still .92. Table 4 shows the estimated means of the person distributions on the two dimensions obtained from this anlaysis compared to those from the previous multidimensional analysis.

Apparently, after rescaling, the two person distributions do not vary drastically from one other. The rescaled analysis has a likelihood deviance (−2 loglikelihood, G^2) of 9011.607 with 125 parameters in total (unconstrained). The difference in the likelihood deviance between this model and the unidimensional model is 110.97 and the difference in the number of parameters is 3. Distributed as a chi-square with 3 degrees of freedom, the difference in the deviance is statistically significant at .01 level. The rescaled model fits the data better than the unidimensional model.

Figs. 1 and 2 graphically present the rescaled results. The logit scale, along which both person performance and stage difficulty estimates are arranged, is shown in the leftmost column on both figures. Fig. 1 shows the person performance map, including the numbers of persons at each logit point for both dimensions. Fig. 2 shows the stage difficulty estimates, separated by item type. The six moral issues scored with the SISS are listed in the middle column and the questions scored with the HCSS are listed on the right. There are clear patterns of ordered stage difficulty estimates in the item levels, especially in the HCSS column. Some overlap of transitional and full stages are apparent in the SISS column. In addition, there are very clear separations between the stage difficulty estimates in the HCSS column. The separation is less obvious in the SISS column; in particular, there is some unexpected overlap at the lower stages with stage 2 estimates spanning four logits.

When item difficulty estimates cluster into clearly separated groups in a partial credit analysis, it is an indication that, in general, individuals are more likely to perform at their modal level, and less likely to perform at lower or higher levels than their modal level (Bond & Fox, 2001). The graphic results of this trend are bands of white space between groups of item estimates. When the individual statements of a large percentage of the respondents in a sample are scored predominantly at

Logit	# Dimension 1 (SISS) Distribution	# Dimension 2 (HCSS) Distribution
8.0	0	0
	2 XX	1 X
	0	1 X
	0	1 X
	0	0
6.0	2 XX	2 XX
	4 XXXX	6 XXXXXX
	7 XXXXXXX	3 XXX
	2 XX	5 XXXXX
	5 XXXXX	5 XXXXX
4.0	4 XXXX	9 XXXXXXXXX
	9 XXXXXXXXX	13 XXXXXXXXXXXXX
	18 XXXXXXXXXXXXXXXXXX	9 XXXXXXXXX
	11 XXXXXXXXXX	23 XXXXXXXXXXXXXXXXXXXXXX
	17 XXXXXXXXXXXXXXXX	19 XXXXXXXXXXXXXXXXXX
2.0	26 XXXXXXXXXXXXXXXXXXXXXXXXXX	20 XXXXXXXXXXXXXXXXXXX
	31 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	19 XXXXXXXXXXXXXXXXXX
	28 XXXXXXXXXXXXXXXXXXXXXXXXXX	21 XXXXXXXXXXXXXXXXXXXX
	29 XXXXXXXXXXXXXXXXXXXXXXXXXXXX	27 XXXXXXXXXXXXXXXXXXXXXXXXXX
	21 XXXXXXXXXXXXXXXXXXX	24 XXXXXXXXXXXXXXXXXXXXXXX
0.0	22 XXXXXXXXXXXXXXXXXXXX	31 XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	19 XXXXXXXXXXXXXXXXX	22 XXXXXXXXXXXXXXXXXXXX
	20 XXXXXXXXXXXXXXXXXX	19 XXXXXXXXXXXXXXXXX
	22 XXXXXXXXXXXXXXXXXXXX	8 XXXXXXX
-2.0	8 XXXXXXX	20 XXXXXXXXXXXXXXXXXX
	10 XXXXXXXXX	9 XXXXXXXX
	20 XXXXXXXXXXXXXXXXXX	9 XXXXXXXX
	6 XXXXX	5 XXXX
	11 XXXXXXXXX	9 XXXXXXXX
-4.0	6 XXXXX	13 XXXXXXXXXXX
	8 XXXXXXX	7 XXXXXX
	4 XXXX	2 XX
	2 XX	7 XXXXXX
	2 XX	3 XXX
-6.0	2 XX	2 XX
	0	4 XXXX
	0	0
	0	0
	0	0
-8.0	0	0

Fig. 1. Rescaled map.

single stages (are consolidated at single stages), the bands of white space widen. The cognitive developmental postulate of structured wholeness predicts that once individuals have access to the structures and processes of a given order of complexity, they will tend to apply these systematically, at least within a given domain. Consequently, as noted above, it would be unlikely for individuals to demonstrate reasoning at more than two adjacent complexity orders within a particular domain of knowledge. Only two types of performances would be expected: those consolidated at a single stage, and those in transition from one stage to another (Dawson, 1998; Dawson et al., submitted for publication). In a partial credit analysis, large gaps between groups of item difficulty estimates reflect just such a pattern of performance.

Logit	SISS	HCSS
8.0		
	5	
	5	
		SP
	5 5	**SP SP**
6.0	**4.5 4.5**	**SP SP SP**
	5 4.5 4.5	**SP SP SP SP SP SP SP SP**
	4.5 4.5	**SP SP**
		SP SP
4.0		
	4 4	
	4	AS
	4 4	AS
2.0	4	AS AS AS AS
		AS AS AS
	3.5 3.5	AS AS AS AS AS AS AS AS
	3.5 3.5 3.5	
		AS
0.0		
	3.5	
	3	
	3	
	3	**AM**
-2.0	**3** 2.5	**AM**
	3 3 2.5 2.5	**AM AM AM AM AM AM AM AM**
	2.5	**AM AM AM AM AM AM AM AM AM AM**
	2.5	**AM**
	2	SA
-4.0	2.5 **2**	SA SA
		SA SA SA
	2 1.5 1.5	SA SA SA SA SA SA SA
	2 1.5	SA SA
	1.5	
-6.0		**RS RS**
	2 1.5	**RS RS RS**
-8.0		
	2	

RS = representational systems, SA = single abstractions, AM = abstract mappings, AS = abstract systems, SP = single principles/axioms

Fig. 2. Rescaled map.

The rescaled map in Fig. 2 also reveals that the HCSS is somewhat "easier" than the SISS. For example, the abstract mappings order is theoretically equivalent to a moral stage score of 3, but the band for this complexity order in the HCSS column is a bit below the band for stage 3 in the SISS column.

5. Discussion

The results from the direct comparison of the two different scoring systems ($r = .92$) reveal that, to a very great extent, the HCSS and the SISS measure the same latent dimension of ability: hierarchical complexity of performance. Though the multidimensional model provided a statistically significantly better fit than the unidimensional model, it is evident that both systems produce ordered stage-like estimates, and that particular complexity orders and moral stages are systematically associated with one another.

Second, individual performances display a high degree of consistency in their hierarchical complexity. This consistency in performances is more marked when performances are scored with the HCSS than with the SISS. The stage estimates produced with the HCSS are more clearly delineated than those with the SISS, with larger gaps between stages and no overlap. Further, detailed examination of individual performances indicates that the HCSS scores across the protocols of any one respondent rarely span more than two complexity orders.

Third, the estimated stages produced by the HCSS appear to be somewhat "easier" than those produced by the SISS. As the HCSS specifies that statements should be scored at the highest stage of hierarchical complexity evident in the statement, and that borderline cases should be assigned the higher stage whereas the SISS assigns borderline cases to a transitional stage, this "easiness" in the HCSS is what we expected. In addition, because no concept matching is necessary with the HCSS, the actual complexity order of an argument is less likely be obscured by the failure to find a matching argument in a scoring manual.

Fourth, complexity order estimates are more orderly than moral stage estimates, particularly at the representational systems order (moral stage 2). Dawson (in press) argues that the lack of "orderliness" at Kohlberg's lower stages is due to the misspecification of some SISS scoring criteria at stages 1 and 2. She notes that several researchers have reported problems with the definition of Kohlberg's lower stages (Damon, 1977; Keller et al., 1989; Kuhn, 1976), and argues that Kohlberg's lower stage scoring criteria were based on performances that were too developmentally advanced to provide accurate data on lower-stage behavior. Kohlberg's youngest respondents were 10 years old, the modal age for the emergence of abstractions (Fischer & Bidell, 1998).

Overall, a domain-general developmental stage scoring system grants several advantages. First, the time and expense of producing a different scoring system for every domain of knowledge is not necessary, and the need for raters to go through

learning process of different systems is eliminated. Second, a domain-general scoring system makes it possible to conduct independent content analyses and stage assessments. Only when conceptual content and stage are assessed independently can we make meaningful cross-cultural, cross-gender, and cross-context comparisons of conceptual knowledge within developmental levels. Moreover, the independent assessment of conceptual content and hierarchical complexity works against the tendency for particular conceptual content to be conflated with developmental stage. In this way we avoid problems associated with defining stages in terms of the conceptual content produced by a limited number of respondents located in a particular time and place.

The MRCML model is a powerful tool for comparing developmental assessment systems. In concert with the RCML model, it has provided us with important insights into the functioning of the SISS and HCSS, allowing us to address important questions about similarities and differences between domain-dependent and domain-general developmental assessment systems.

Acknowledgments

This work would not have been possible without the donation of interview data by the Murray Research Center, Larry Walker, Cheryl Armon, Marvin Berkowitz, Michael Commons, and Peggy Drexler. The research reported in this paper was made possible in part by a grant from the Spencer Foundation. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

Appendix A. The representational mappings to single principles/axioms complexity orders

At the representational mappings order, the new concepts are referred to as second order representational sets. They coordinate or modify representational sets (the concepts constructed at the single representations order). The most complex logical structure of this complexity order is linear, coordinating one aspect of two or more representations.

At the representational systems order, the new concepts are third order representational sets. These coordinate elements of representational systems. The most complex logical structure of this complexity order is multivariate, coordinating multiple aspects of two or more representations.

At the single abstractions order, the new concepts are referred to as first order abstractions. These coordinate representational systems. The most complex logical structure of this complexity order identifies one aspect of a single abstraction.

At the abstract mappings order, the new concepts are referred to as second order abstractions. These coordinate or modify abstractions. The most complex

logical structure of this complexity order coordinates one aspect of two or more abstractions.

At the abstract systems order, the new concepts are referred to as third order abstractions. These coordinate elements of abstract systems. The most complex logical structure of this complexity order coordinates multiple aspects of two or more abstractions.

At the single principles/axioms order, the new concepts are referred to as first order principles. These coordinate abstract systems. The most complex logical structure of this complexity order identifies one aspect of a principle or axiom coordinating systems.

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Armon, C. (1984). *Ideals of the good life: Evaluative reasoning in children and adults*. Unpublished doctoral dissertation, Harvard, Boston.

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

Berkowitz, M. W., Guerra, N., & Nucci, L. (1991). Sociomoral development and drug and alcohol abuse. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (pp. 35–53). Hillsdale, NJ: Erlbaum.

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

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

Case, R. (1985). *Intellectual development: Birth to adulthood*. New York: Academic Press.

Case, R., Okamoto, Y., Henderson, B., & McKeough, A. (1993). Individual variability and consistency in cognitive development: New evidence for the existence of central conceptual structures. In R. Case & W. Edelstein (Eds.), *The new structuralism in cognitive development: Theory and research on individual pathways* (pp. 71–100). Basel, Switzerland: Karger.

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

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

Colby, A., Kohlberg, L., Gibbs, J., & Lieberman, M. (Eds.). (1983). *A longitudinal study of moral judgment* (Vol. 48).

Commons, M. L., Danaher, D., Miller, P. M., & Dawson, T. L. (2000, June). *The Hierarchical Complexity Scoring System: How to score anything*. Paper presented at the Annual meeting of the Society for Research in Adult Development, New York.

Commons, M. L., Richards, F. A., Ruf, F. J., Armstrong-Roche, M., & Bretzius, S. (1984). A general model of stage theory. In M. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations* (pp. 120–140). New York: Praeger.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, S. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review, 18*, 237–278.

Damon, W. (1977). Measurement and social development. *Counseling Psychologist, 6*, 13–15.

Dawson. (1997). *New tools, new insights: Kohlberg's moral reasoning stages revisited.* Paper presented at the Twenty-Seventh Annual Symposium of the Jean Piaget Society, Santa Monica, CA.

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

Dawson, T. L. (2002a, January). *The Hierarchical Complexity Scoring System.* Retrieved November 2002, from http://gseacademic.harvard.edu/~hcs/base/index.shtml.

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

Dawson, T. L. (2002c). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement, 3*, 146–189.

Dawson, T. L. (in press). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology.*

Dawson, T. L., Commons, M. L., & Wilson, M. (submitted for publication). The shape of development.

Dawson, T. L., Commons, M. L., Wilson, M., & Xie, Y. (1999, June). *The general model of Hierarchical Complexity Scoring System: Refinements and additions.* Paper presented at the annual symposium of the Jean Piaget Society, Mexico City, Mexico.

Dawson, T. L., & Gabrielian, S. (in press). Developing conceptions of authority and contract across the life-span: Two perspectives. *Developmental Review.*

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

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

Fischer, K. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477–531.

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

Fischer, K. W., & Bullock, D. H. (1981). Patterns of data: Sequence, synchrony, and constraint in cognitive development. In K. W. Fischer (Ed.), *Cognitive development* (pp. 1–22). San Francisco, CA: Jossey-Bass.

Keller, M., Eckensberger, L. H., & von Rosen, 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.

Kitchener, K. S., & King, P. M. (1990). The reflective judgment model: Ten years of research. In M. L. Commons, C. Armon, L. Kohlberg, F. A. Richards, T. A. Grotzer, & J. D. Sinnott (Eds.), *Adult development* (Vol. 2, pp. 62–78). New York: Praeger.

Kitchener, K. S., King, P. M., Wood, P. K., & Davison, M. L. (1989). Sequentiality and consistency in the development of reflective judgment: A six-year longitudinal study. *Journal of Applied Developmetal Psychology, 10*, 73–95.

Kitchener, K. S., Lynch, C. L., Fischer, K. W., & Wood, P. K. (1993). Developmental range of reflective judgment: The effect of contextual support and practice on developmental stage. *Developmental Psychology, 29*, 893–906.

Kohlberg, L., & Candee, D. (1984). Stage and sequence: The cognitive developmental approach to socialization. In *The psychology of moral development: The nature and validity of moral stages* (pp. 7–169). San Francisco, CA: Jossey Bass.

Kuhn, D. (1976). Short-term longitudinal evidence for the sequentiality of Kohlberg's early stages of moral judgment. *Developmental Psychology, 12*, 162–166.

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

Piaget, J. (1985). *The equilibration of cognitive structures: The central problem of intellectual development* (T. Brown & K. J. Thampy, Trans.). Chicago: The University of Chicago Press.

Rest, J. R. (1975). Longitudinal study of the Defining Issues Test of moral judgment: A strategy for analyzing developmental change. *Developmental Psychology, 11*, 738–748.

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

Stevens, D. A. (2000, June). *Dynamic systems: Continuities and discontinuities with Piaget's theory of equilibration*. Paper presented at the annual meeting of the Jean Piaget Society.

Ullian, D. Z. (1977). The development of conceptions of masculinity and femininity. *Dissertation Abstracts International, 37*(7-B).

van Geert, P. (2000). The dynamics of general developmental mechanisms: From Piaget and Vygotsky to dynamic systems models. *Current Directions in Psychological Science, 9*, 64–68.

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

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

Wu, M., Adams, R., & Wilson, M. (1998). *ConQuest user guide*. Hawthorn, Australia, ACER.