Layers of Structure: A comparison of two approaches to developmental assessment

Theo L. Dawson

Graduate School of Education, University of California at Berkeley, theod@uclink4.berkeley.edu

Abstract

Proponents of domain-specific cognitive developmental assessment systems argue that development in particular knowledge domains is characterized by domain-specific structures, concluding from this that developmental assessment should be conducted with domain-specific assessment systems. My colleagues and I have questioned this assertion in a series of studies comparing results obtained with several domain-specific developmental assessment systems and one domain-general developmental assessment system. For the present study, 8 expert stage-scorers were interviewed with a think-aloud procedure as they scored a range of text segments from a diverse sample of interview protocols. Two of these raters employed Kohlberg's Standard Issue Scoring System (Colby & Kohlberg, 1987b) and Armon's Good Life Scoring System (Armon, 1984b) and six employed Commons' Hierarchical Complexity Scoring System (Commons et al., 1995). Despite differences in training, the mean scores of these two groups of raters agreed within one full order of hierarchical complexity 97.6% of the time. In spite of this, the criteria employed to determine the stage of statements varied widely across the two groups of raters. Some of these criteria centered on particular conceptual content, some referred primarily to socio-moral perspective, and others referred to highly formal indicators of hierarchical complexity—hierarchical order of abstraction and logical structure. To aid in explaining the strong correspondence between stage scores awarded with the three scoring systems, the relationships among different types of scoring criteria are modeled as 'layers' of structure representing differing levels of abstraction.

Introduction

As part of a larger project investigating cognitive development across the lifespan, my colleagues and I have conducted numerous comparisons of developmental assessment systems. The present study, a comparison of two domain-based scoring systems with one domain-general scoring system, examines rater behavior. Eight expert stage-scorers were interviewed as they scored a range of text segments from a diverse sample of interviews. Despite their dissimilar training and considerable variation in the criteria employed to determine the stage of text segments, I find relatively little disagreement in stage assignment across raters, even though some of them believe the methods they employ result in stage scores that are significantly different from those generated with alternative systems. In this paper, I explore the possibility that similarities in the scores assigned by raters can be attributed to a common structural core underlying the various scoring schemes employed by these raters.

Though most cognitive developmental scoring systems are intended to tap Piagetian operational structures, such systems actually rely upon the identification of particular conceptual content (Armon, 1984a; Colby & Kohlberg, 1987b; Damon, 1980; Kitchener & King, 1990) rather than the direct identification of these structures. As a consequence, the relationship between stage and particular conceptual content is often confounded, sometimes to the extent that stage is defined in terms of that content, rather than in terms of the operatory structures that form the basis of cognitive developmental theory. One of the several difficulties with this trend is that the more content-bound scoring systems become, the more problematic it is to compare results across content domains or cultural contexts (King, Kitchener, Wood, & Davison, 1989). Other, interrelated, problems with content-dependent scoring systems include (1) the virtual impossibility of providing an exhaustive account of the conceptual content of any domain; (2) the documented presence in individual verbal performances of conceptual content associated with multiple stages (Dawson, 1998); and (3) the potential for cultural or gender bias when scoring systems are based on the responses of non-representative samples.

While stages of development are spoken of as having universal properties—often described in terms of hierarchical complexity (Commons, Trudeau, Stein, Richards, & Krause, 1998; Piaget, 1973)—that hold across domains of reasoning, relatively few efforts have been made to establish a method that can be used to measure these properties across domains. To express the problem another way, we presently have a theoretical construct, developmental stage, which has universal properties in that aspects of the construct are defined similarly across domains, but we have a separate ruler for every domain in which stage is assessed and no satisfactory means for comparing rulers across domains.

Using conventional methods we can neither assess whether criteria for determining stage across domains are equivalent, nor whether some developmental changes in reasoning can be successfully modeled as universal "stages."

Theoretically, if developmental stage could be assessed in terms of its universal features (hierarchical complexity) rather than its contextual features (domain and context specific content), it would be possible to (1) meaningfully compare development across domains and contexts, and (2) examine the relationship between developmental stages and conceptual content. Case (1991), Fischer (1980), and Commons (Commons et al., 1998) have each introduced a generalized developmental theory. Moreover, they all employ the definitions of their developmental levels (or stages) to identify developmental sequences, but only Commons and his colleagues (Commons, Danaher, Miller, & Dawson, 2000) have provided a domain-general developmental scoring system.

Moral and Good Life Stages

Kohlberg's moral stages describe 3 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, though the scoring manual provides scoring criteria for only five of these. The Standard Issue Scoring System is designed around a set of standard moral judgment interviews, each of which presents a moral dilemma dealing with issues of life, law, conscience, contract, authority, and/or punishment. Kohlberg and his colleagues employed what they called a *bootstrapping* process to define moral judgment stages and to construct a scoring manual. They started with a theoretical sequence strongly influenced by both moral philosophy and Piaget's (1965) idea that moral thinking moves from heteronomy to autonomy and from concrete to abstract, and refined their understanding of moral stages as they gathered successive rounds of longitudinal data from a group of New England school boys, the youngest of whom were 10 years old at the outset of the study.

The Standard Issue Scoring System itself was constructed by analyzing seven sets of interviews from Kohlberg's original longitudinal study. Each of these sets of interviews included assessments from all 6 of the test times, which were separated by four-year intervals. Each performance was assigned a global stage score "based on intensive discussion and analysis" (Colby & Kohlberg, 1987a) (p. 40) employing criteria from an earlier version of the scoring system. Then, individual responses to each dilemma provided the basis for examples in the scoring manual. To score, the rater matches the judgments and justifications provided by a given respondent with *criterion judgments* in the scoring manual. Skilled raters become adept at "looking through" the *particular* concepts in a performance, and base their ratings on more structural features such as interpersonal perspective.

Armon's Good Life Stages define 5 stages of evaluative reasoning about the good, covering the period from early childhood to adulthood. The *Good Life Interview* focuses on evaluative reasoning about the good, including the good life, good work, the good person, and good friendship. Armon's interview approach differs from Kohlberg's in that she does not pose dilemmas. Instead, her interview technique employs open-ended questions such as, "What is the good life?" "Why is that good?" Like Kohlberg, Armon employed a bootstrapping process to develop stages of the good life, and was guided in her stage definitions by philosophical categories. Moreover, her stage scoring method relies, as does Kohlberg's, on content descriptions.

Scoring with the domain-based scoring systems of Armon (1984b) and Kohlberg (Colby & Kohlberg, 1987b), though it is grounded in what both authors argue are structural criteria, involves matching the concepts identified in interviews with similar concepts in a scoring manual. For example, the criterion judgments in Kohlberg's manual 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. This is not to say that these scoring systems entirely fail to distinguish between structure and content. Kohlberg's 6 moral issues, for example, represent different conceptual categories, and the system allows for the separate coding of moral "norms" and "elements". Stage assignment is not dependent upon which issue, element, or norm is employed in an argument. However, the criterion judgments are laden with conceptual content, and one ultimately scores based on how well a respondent's argument matches exemplars on the same issue.

Kohlberg, along with other proponents of domain theories of cognitive development (Cosmides & Tooby, 1994; Demetriou & Efklides, 1994) argues that development in different

knowledge domains involves fundamentally different processes. For Kohlberg, the moral domain constitutes a unique structure parallel to, but progressively differentiated from, both the Piagetian logico-mathematical domain and the perspective-taking domain. From the domain perspective, it is necessary to assess cognitive development in a given knowledge domain with a domain-based scoring system that incorporates the structures specific to that domain.

The Model of Hierarchical Complexity

The Model of Hierarchical Complexity is a model of the hierarchical complexity of tasks. In the Model of Hierarchical Complexity, an action is considered to be at a given order of hierarchical complexity (we will refer to these as *complexity orders*) when it successfully completes a task of that order of hierarchical complexity. Hierarchical complexity refers to the number of non-repeating recursions that coordinating actions must perform on a set of primary elements. Actions at the higher order of hierarchical complexity are: (a) defined in terms of the actions at the next lower order; (b) organize and transform the lower order actions; and (c) produce organizations of lower order actions that are new and not arbitrary and cannot be accomplished by the lower order actions alone. The Model of Hierarchical Complexity does not define stages in terms of specific logical abilities like Piaget's INRC groups or class inclusion. Rather, it proposes that a given class inclusion task, for example, has a particular order of hierarchical complexity, depending on the nature of its elements and the kinds of classes formed.

The Model of Hierarchical Complexity specifies 15 complexity orders. The sequence is—(0) computory, (1) sensory & motor, (2) circular sensory-motor, (3) sensory-motor, (4) nominal, (5) sentential, (6) preoperational, (7) primary, (8) concrete, (9) abstract, (10) formal, (11) systematic, (12) metasystematic, (13) paradigmatic, and (14) cross-paradigmatic. The first three complexity orders correspond to Piaget's sensorimotor stage, 3-5 correspond to his symbolic or preoperational stage, 6-8 correspond to his concrete operational stage, and 9-11 correspond to his formal operational stage. Complexity orders 0 to 12 also correspond definitionally to the levels and tiers originally described by Fischer (1980). Complexity orders 13 and 14 are hypothetical postformal stages (Sonnert & Commons, 1994). For definitions of the complexity orders, see Commons, Trudeau, Stein, Richards, and Krause (Commons et al., 1998) or Dawson, Commons, & Wilson (Dawson, Commons, & Wilson, in review).

Unlike the Standard Issue Scoring System and the Good Life Scoring System, the Hierarchical Complexity Scoring System does not require any concept matching. This is possible because hierarchical complexity is reflected in two aspects of performance that can be abstracted from particular conceptual content. These are (a) hierarchical order of abstraction and (b) the logical organization of arguments. Hierarchical order of abstraction is observable in texts because new concepts are formed at each complexity order as the operations of the previous complexity order are "summarized" into single constructs. Halford (1999) suggests that this summarizing or "chunking" makes advanced forms of thought possible by reducing the number of elements that must be simultaneously coordinated, freeing up processing space and making it possible to produce an argument or conceptualization at a higher complexity order. Interestingly, at the sensory-motor, preoperational, abstract, and metasystematic complexity orders, the new concepts not only coordinate or modify constructions from the previous complexity order, they are qualitatively distinct conceptual forms—symbols, representations, abstractions, and principles, respectively (Fischer, 1980). The appearance of each of these conceptual forms ushers in three repeating logical forms—definitional, linear, and multivariate. Because these three logical forms are repeated several times throughout the course of development, it is only by pairing a logical form with a hierarchical order of abstraction that a rater can make an accurate assessment of the complexity order of a performance. It should be noted that other researchers have observed and described similar conceptual forms and repeating logical structures (Case, Okamoto, Henderson, & McKeough, 1993; Fischer & Bidell, 1998; Piaget, Garcia, & Feider, 1989).

To test the feasibility of domain-general scoring, my colleagues and I have conducted several investigations into the relationship between the Hierarchical Complexity Scoring System and the more domain-specific scoring systems of Armon (1984b) and Kohlberg (Colby & Kohlberg, 1987b). The first of these investigations (Dawson, in press) examines the relationships among three stage scoring systems by comparing performances on three different stage measures employed in scoring three different interviews. These were (a) Kohlberg's moral judgment interview, scored with the

Standard Issue Scoring System; (b) Armon's good education interview, scored with the Hierarchical Complexity Scoring System; and (c) Armon's good life interview, scored with the Good Life Scoring System (Armon, 1984b). Though the internal consistency of the Hierarchical Complexity Scoring System is somewhat greater than that of the Good Life Scoring System or Standard Issue Scoring System (yielding more stage-like patterns of performance), the Hierarchical Complexity Scoring System appears to be somewhat 'easier' than the other two scoring systems (especially at the lower orders of hierarchical complexity). In addition, there is a strong correspondence (disattenuated correlations from .90 to .97) among stage scores awarded with these systems. Dawson argues that the strong correlations among the three systems, combined with patterns in the acquisition of analogous Good Life Stages, Moral Stages, and complexity orders provide evidence that the three scoring systems predominantly assess the same latent dimension: hierarchical complexity.

In a second study, Dawson, Xie, and Wilson (in prep) carried out a multidimensional partial credit Rasch analysis of the relationship between scores obtained with the Standard Issue Scoring System and scores obtained with the Hierarchical Complexity Scoring System on 378 moral judgment interviews from respondents aged 5 to 86. They report a disattenuated correlation of .92 between scores awarded with the two scoring systems, suggesting that to a large extent these two systems assess the same dimension of performance, though the Hierarchical Complexity Scoring System awards somewhat higher scores than the Standard Issue Scoring System, particularly at the lower complexity orders. The Hierarchical Complexity Scoring System also reveals more stage-like patterns of performance than the Standard Issue Scoring System, including evidence of developmental spurts and plateaus.

In a third study, Dawson and Kay (in review) conducted two detailed examinations of the relationship between moral stages and complexity orders. In the first of these, the Hierarchical Complexity Scoring System was used to score the 219 criterion judgments (scoring criteria) from Form A (Heinz and Joe) of the Standard Issue Scoring System. They conclude that the primary, concrete, abstract, formal, systematic, and metasystematic complexity orders correspond predominantly to moral stages 1, 2, 2/3, 3, 4, and 5. Criterion judgments for stages 2/3 through 5 correspond well with analogous complexity orders. However, criterion judgments for Kohlbergian stages do not correspond well to complexity orders below the abstract complexity order, where 23% of stage 1 criterion judgments were scored as concrete instead of primary, and 24% of stage 2 criterion judgments were scored as abstract instead of concrete. Moreover, some of Kohlberg's Stage 1 criterion judgments were scored as preoperational, a complexity order that has no clear parallel in Kohlberg's hierarchy, though other researchers have commented on the acquisition of social and moral concepts in pre-school children (Fischer, Hand, Watson, van Parys, & Tucker, 1984; Killen, 1991).

In a second analysis of the same data set, Dawson and Kay examined the relationship between scores awarded with the Hierarchical Complexity Scoring System and the Standard Issue Scoring System in a sample of 637 moral judgment interviews scored with both systems. Strong correspondences were found between scores awarded by the two systems from the formal to metasystematic complexity orders. In this range, the two systems agree within 1/2 of a complexity order 84% to 89% of the time. At the concrete and abstract complexity orders, agreement rates are considerably lower, with agreement within 1/2 of a complexity order only 50% to 61% of the time. The authors argue that Kohlberg's stages are misspecified below moral stage 3 (the formal complexity order). They provide three sources of evidence. First, they note that several researchers have reported problems with the definition of Kohlberg's lower stages (Damon, 1977; Keller, Eckensberger, & von Rosen, 1989; Kuhn, 1976). Second, Rasch scaling of the Standard Issue Scoring System results consistently reveals more 'noise' at the lower stages than at the higher stages (Dawson, 2002; Xie & Dawson, 2000). And finally, individuals with mean scores at moral stages 1 and 2 tend to receive a statistically significantly wider range of scores (mean range = 1 stage) than those with mean scores at moral stages 3, 4, or 5 (mean ranges = .60 to .85 stage), suggesting that reasoning is either (a) less cohesive at the lower stages than at the higher stages or (b) that scoring criteria are not as well specified at stages 1 and 2 as they are at stages 3 through 5. The misspecification of Kohlberg's lower stages is probably due to the fact that his 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).

Stage and structure

The idea that cognitive development proceeds through a series of stages or levels, each of which has a particular structure, is central to most cognitive-developmental theories. But the definition of structure varies. Kohlberg and Armon (1984) argue that the various definitions of structure can be arranged in a hierarchy, from a content level, through a surface structure level, to a "deep" or "hard" structure level. They maintain that only those systems defined in terms of hard structure can be expected to meet Piagetian criteria for structured wholeness, invariant sequence, hierarchical integration, and qualitative reorganizations of knowledge. Kohlberg claims that his Standard Issue Scoring System meets these four criteria, while Armon sees her good life stages as requiring further validation. Commons"(Commons et al., 1998) claims that his Model of Hierarchical Complexity meets all of these requirements, and adds the requirements that (1) the "stage" properties of performances be abstracted from the particular tasks they address, and (2) that the logic of each stage's tasks must be explicit. Because of these added criteria Commons claims that his stages are even "harder" than those of Kohlberg and Armon.

In the present investigation into scoring behavior, I first compare scores awarded by Standard Issue and Good Life

Table 1: Probe questions for rater interviews

1. What would you say are the most important things you attend to when you score a text for stage?
2. What are first things you notice when you read this statement? Why?
3. What role do the particular words in this statement play in your scoring process? How and why?
4. Are there particular phrases that provide clues to the stage of this statement? What are they? What information do they provide? Is it in the meaning conveyed or in the form of the phrases, or both? How do these relate to one another?
5. Are there clues to the stage of this statement in structural features that transcend its content? What are they? Why do you think they transcend the content?
6. Are there any features of this statement that make it unusually difficult/easy to score? What are they? Why?
7. How do you know this statement is at the given stage rather than the prior or later stage?

Table 2: Sample protocols from rater interview

Good Education 0212

[OK. HOW IMPORTANT IS GOOD EDUCATION IN THE GOOD LIFE? COULD YOU HAVE A GOOD LIFE WITHOUT HAVING HAD A GOOD EDUCATION?]

Yeah, probably so, I would say. I wouldn't, it would be richer with education but it wouldn't…

[WHY DO YOU THINK IT WOULD BE RICHER WITH EDUCATION?]

Well you just; your mind would be open to a lot more things.

Good Life: 008

[WHAT IS THE GOOD OR IDEAL LIFE?]

Did I mention that good work and good life were synonymous?

[YES.]

Right.

[IS THERE ANYTHING OTHER THAN GOOD WORK THAT WOULD GO INTO THE GOOD LIFE?]

Perhaps I should redefine work. First of all they're concurrent; they're working parallel. Good work can be going out and building something and doing a good job. Also work can be your own self-development, if you enhance your awareness of the world through reading or television, or movies, or seminars. There's various communicative means that you have to do. I don't know whether you would say it's work, but it's…

[WOULD YOU SAY IT'S WORK THAT'S THE ONLY THING THAT'S IMPORTANT.]

I don't look at it; it depends on how you define work. Some people would say that work is a drudgery and a burden.

[BUT WHAT DO YOU SAY?]

For me, it is work, but it's a life work for self-education and self-awareness. Actually I feel like there isn't enough time in life to do everything I would like to do. There's too much to know and too much stimulation, and too much intellectual stuff to know, too much physical stuff in the world, to actual do to implement it, it's overwhelming sometimes. But you have to set priorities.

Joe: 008

[SHOULD JOE REFUSE TO GIVE HIS FATHER THE MONEY?]

Yes.

[WHY?]

We're talking about two people having certain expectations, and one of them is working towards them and the other one isn't. Joe has made this contract with himself to save the money to enhance his life and to have potential and growth, and he's gone ahead and worked hard and saved it. He has every right to use this money for his own advancement, his own life experiences, etc. Now his father had an option to have his own money too or cash reserve, and his father doesn't have, as far as I'm concerned, jurisdiction or that type of control over Joe's life. His father's responsibility as a parent is to provide physical support, health, clothing, and also moral guidance showing Joe how to live a good life, a responsible life, and not to abuse people, and he's certainly not doing this by taking his kid's money.

Table 3: A Comparison of Three Developmental Numbering Sequences

MHC Stage	Kohlbergian stage	Good Life
Primary	1.0	1.0
Concrete	2.0	2.0
Abstract	2.5	2.5
Formal	3.0	3.0
Systematic	4.0	4.0
Metasystematic	5.0	5.0

raters with those awarded by Hierarchical Complexity raters. I anticipate that these scores will be similar, though earlier research suggests that the Hierarchical Complexity Scoring System will award somewhat higher scores than the Standard Issue Scoring System and Good Life Scoring System and that differences between the two types of scoring systems may be greater at the lower complexity orders because of known problems with the specification of moral stages 1 and 2. I then examine the criteria employed by the two groups of raters in terms of the level of structure they evoke, modeling the results to explain why the Standard Issue Scoring System, Good Life Scoring System, and Hierarchical Complexity Scoring System yield similar stage-scores despite the fact that they employ different scoring criteria. The data are interviews of 8 expert stage-scorers that were interviewed as they rated a common set of text segments from moral judgment, evaluative reasoning, self-understanding, and science reasoning interviews. Two of these raters employed the Standard Issue Scoring System (Colby & Kohlberg, 1987b) and Good Life Scoring System (Armon, 1984b), and six employed the Hierarchical Complexity Scoring System (Commons et al., 2000). I speculate that close examination of the strategies and criteria employed by these raters will reveal a common structural core, consistent with a notion of stages as orders of hierarchical complexity.

Method

Eight expert raters were administered a talk-aloud interview. All of the raters have published results using the assessment procedures about which they were being questioned. The interview consisted of a set of up to 42 text performances (protocols) selected to represent a range of content domains and orders of hierarchical complexity. As each statement was scored, raters were asked a variety of questions about their scoring process (Table 1). Questions were not always posed precisely as presented in Table 1, but were asked as required to stimulate the rater to discuss his or her reasoning process during scoring. The protocols scored by the eight raters were taken from moral judgment and evaluative reasoning interviews collected for earlier studies (Armon & Dawson, 1997; Dawson, 1998). Each protocol consists of the complete argument supporting a claim or set of interrelated claims made by a respondent about a particular issue. Some of these protocols are shown in Table 2.

Hierarchical Complexity raters were asked to score each protocol on a 10-point scale, with 10 as metasystematic, 9 as transitional from systematic to metasystematic, 8 as systematic, 7 as transitional from formal to systematic, 6 as formal, 5 as transitional from abstract to formal, 4 as

Table 4: Correlations of scores among raters

Rater		HCSS1	HCSS2	HCSS3	HCSS4	HCSS5	HCSS6	SISS1
HCSS2	Pearson Corr.	.947						
	N	25						
HCSS3	Pearson Corr.	.925	.888					
	N	19	14					
HCSS4	Pearson Corr.	.895	.966	.896				
	N	10	7	9				
HCSS5	Pearson Corr.	.968	.945	.935	.933			
	N	42	25	19	10			
HCSS6	Pearson Corr.	.940	.917	.875	.846	.960		
	N	43	25	19	10	42		
SISS1	Pearson Corr.	.882	**.677**	**.757**	.928	.869	**.792**	
	N	20	12	8	5	19	20	
SISS2	Pearson Corr.	.906	.921	.845	.864	.923	.900	**.743**
	N	40	22	16	7	39	40	18

abstract, 3 as transitional from concrete to abstract, 2 as concrete, 1 as transitional from primary to concrete, and 0 as primary. Standard Issue and Good Life raters scored their interviews from stage 1 to stage 5 in 1/2 stage increments. The stage labels from the three scoring systems correspond as shown in Table 3 (Dawson & Kay, in review).

The two Standard Issue raters in the present study, who had both studied with Kohlberg, did not employ the scoring manual directly. Instead, they employed what they called the "structural criteria" upon which the scoring manual is based. The same is true for Good Life scoring, which is usually conducted with a manual similar to the Standard Issue Scoring System manual. Standard Issue raters scored protocols by examining the hierarchical order of abstraction and logical structure of performances, and without reference to the scoring manual.

Three raters scored all 42 statements selected for the interview. Five raters were unable to complete the entire interview, scoring from 18 to 40 statements. All interviews were recorded on audiotape. All scoring criteria employed and reported by each respondent were recorded, along with the stage-score awarded for each statement in the interview. Because of the interactive nature of the interview, some of the questions and probes of the interviewer may, on occasion, have influenced scoring. This should be kept in mind when considering the implications of the following analysis of inter-rater agreement.

Results

Inter-rater agreement

Table 4 presents the inter-rater correlation matrix for all 8 raters. Raters are identified with the abbreviation for the scoring system employed, followed by a number. (*SISS* is used to identify the two raters who employed both the Standard Issue and Good Life Scoring Systems.) The inter-rater correlations shown in this matrix are generally between .85 and .97. However, the scores of one rater, SISS 1, correlate less well with the ratings of HCSS 2, HCSS 3, HCSS 6 and SISS 2. These correlations are highlighted in the table in boldface type. With the exception of these four relatively low correlations, the magnitude of the correlations among these raters are similar to inter-rater correlations commonly reported in the literature (Armon, 1984b; Colby & Kohlberg, 1987a; Dawson, 1998; Dawson et al., in review).

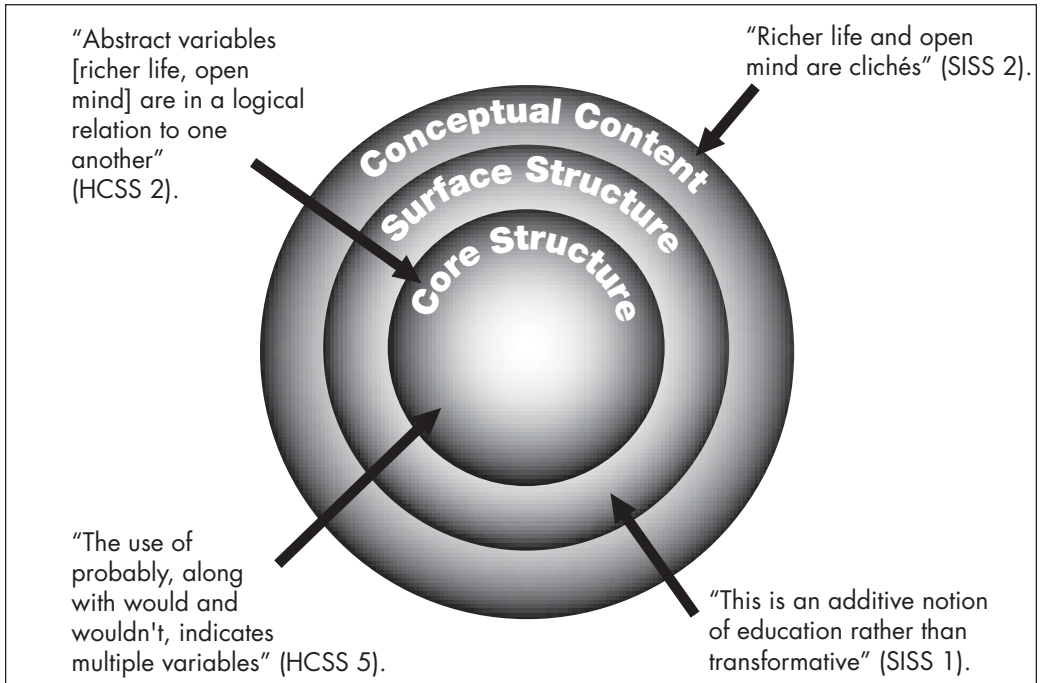
While correlations provide information about the extent to which the stage *sequences* specified by the two types of scoring systems are the same, they do not tell us much about how the stages are related to one another across systems. To understand these relationships, I examined (a) the rates of agreement among the 8 raters and (b) the direction of differences between the two groups of raters. As shown in Table 5, percent agreement rates among Standard Issue raters are within 1 complexity order between 89.5% and 100% (M = 98.0%) of the time and within 1/2 of a complexity order 64.0% to 88.9% (M = 78.5%) of the time. Agreement between the two Standard Issue/Good Life raters is 100% within 1 complexity order and 88.2% within 1/2 of a complexity order. Agreement rates between Hierarchical Complexity Scoring System and Standard Issue/Good Life raters are within 1 complexity order between 85.0% and 100% of the time (M = 97.6%) and within 1/2 of a complexity order 42.0% to 85.0% (M = 71.4%) of the time. These agreement rates, like the correlations reported above, all correspond to inter-rater agreement rates commonly reported in the literature (Armon, 1984b; Colby & Kohlberg,

Table 5: Percent agreement within 1 complexity order

Rater	Rate	HCSS1	HCSS2	HCSS3	HCSS4	HCSS5	HCSS6	SISS1	
HCSS2	Within 1/2		72.0						
	Within 1		**100.0**						
HCSS3	Within 1/2		68.4	64.3					
	Within 1		**94.7**	**92.9**					
HCSS4	Within 1/2		80.0	100.0	88.9				
	Within 1		**100.0**	**100.0**	**100.0**				
HCSS5	Within 1/2		88.1	84.0	78.9	90.0			
	Within 1		**100.0**	**100.0**	**94.7**	**100.0**			
HCSS6	Within 1/2		74.4	64.0	68.4	70.0	85.7		
	Within 1		**97.7**	**96.0**	**89.5**	**100.0**	**100.0**		
SISS1	Within 1/2		60.0	58.3	75.0	80.0	78.9	75.0	
	Within 1		**90.0**	**91.7**	**87.5**	**100.0**	**94.7**	**85.0**	
SISS2	Within 1/2		69.2	72.7	66.7	42.0	86.8	82.1	88.2
	Within 1		**97.4**	**95.5**	**86.7**	**100.0**	**100.0**	**97.4**	**100.0**

1987a; Dawson, 1998; Dawson et al., in review).

Between scoring systems, the overall direction of disagreement reveals that Standard Issue raters award scores that are about 1/2 of a complexity order lower than those awarded by Standard Issue raters. On the 42 protocols scored by members of both groups of raters, only 19.0% of the scores awarded by complexity order raters are lower on average than scores awarded by Standard Issue raters, while 57.1% of the scores awarded by complexity order raters are higher on average than scores awarded by Standard Issue raters. This result corroborates findings from several other comparisons of the Standard Issue Scoring System, Good Life Scoring System, and Hierarchical Complexity Scoring System (Dawson, in press; Dawson & Kay, in review; Dawson et al., in prep), in which Hierarchical Complexity Scoring System scores were found to be somewhat higher than Standard Issue Scoring System scores.



[COULD YOU HAVE A GOOD LIFE WITHOUT HAVING HAD A GOOD EDUCATION?]
Yeah, probably so, I would say. I wouldn't, it would be richer with education but it wouldn't...
[WHY DO YOU THINK IT WOULD BE RICHER WITH EDUCATION?]
Well you just, your mind would be open to a lot more things.

Figure 1: How important is education to the good life?

Layers of structure

To examine the extent to which the "hardness" of the structural criteria employed in stage-scoring impact the scoring process, I have developed a simple model of structural levels onto which the criteria employed by raters in their determination of stage are mapped. Figure 1 shows this three-layer model, in which the outside layer represents conceptual content, the middle layer represents surface structures, and the central layer represents "hard" or "core" structure. This model differs somewhat from the model suggested by Kohlberg and Armon (Kohlberg & Armon, 1984), in that the surface structure level corresponds with their hard structure level and the "core" structure level is defined in terms of Commons' additional hard structure criteria. Recall that these are that (1) the "stage" properties of performances are abstracted from the particular tasks they address, and (2) that the logic of each stage's tasks must be explicit.

The conceptual content layer is the layer referenced when a stage-score is awarded based on the explicit presence or absence of a specific conceptualization without reference to any formal criteria. For example, one rater claims that the central concern raised in a protocol is "care" and awards a score of moral stage 3. This is a content layer criterion. The surface structure layer is referenced when particular conceptual content is said to represent a more general principle or concept that is associated in a given scoring scheme with a particular stage. For example, one rater evokes the interpersonal perspective reflected in the conceptual form of the argument presented in a protocol as grounds for awarding a

stage 4 score. This criterion is at the level of surface structure. The layer representing core structure is referenced when form, organization, or relations are in the foreground and particular conceptual content is seen only as indicative of a particular hierarchical order of abstraction. For example, one rater scores a protocol at the formal order, arguing that the concepts of care, rights, and relationship in the argument are at least at the abstract order, and the way in which they are related to one another is in the form of linear causal arguments. These criteria are at the level of core structure.

To illustrate the diversity of criteria employed by raters in scoring for stage, two examples have been selected from the scorer interviews. Figure 1 shows the first of these. Individual responses of four different raters, appealing to all three layers of structure, are represented in this figure. All 8 raters scored this performance at the formal complexity order (Moral and Good Life stage 3). Rater SISS 2 refers to the conceptual content layer by focusing on the meaning of richer life and open mind. Though he points out that these concepts, if fully formed, could represent a higher level of structure than formal operations, he does not think it possible, given this fragment of the performance, to determine whether the full meanings are present. Rater SISS 2 looks deeper than the particular conceptions, noting that the performance reflects an additive, rather than transformative, notion of the place of education in the good life. While this assessment is certainly less dependent on particular conceptual content than that of rater A, it is still a reference to the particular domain represented in the performance. Initially, it was difficult to assign this criterion to one of the three levels of structure presented here, in part, because additivity and transformation could be references to core structural features of certain complexity orders. I chose to place it here, however, because the rater appeared to be referring specifically to (1) how a good education can make life better by providing more content, versus (2) how it can make life better by changing its quality, without a general notion of the deeper structural source of this distinction. The criteria employed by raters HCSS 2 and HCSS 5, on the other hand, do not refer at all to the particular meanings in this performance. Instead, they focus on the form of the argument, citing particular words or phrases as exemplars of abstract forms. In rater HCSS 2's analysis, for instance, richer life and open mind are abstract variables employed in a logical argument at the formal level.

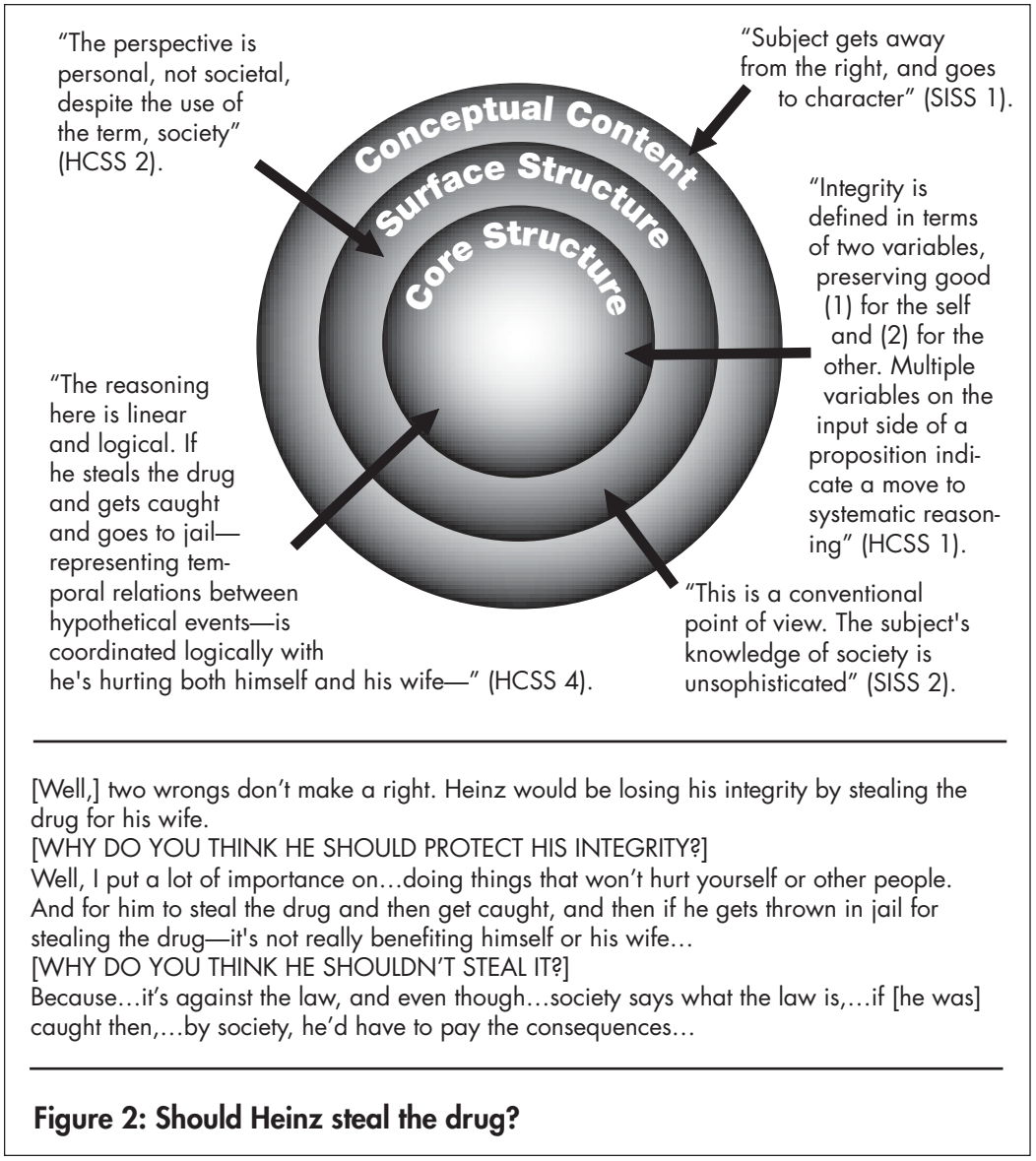


Figure 2: Should Heinz steal the drug?

Figure 2 shows a second example of scoring behavior. In this case, raters were asked to score a segment from a response to Kohlberg's Heinz dilemma. Some of the criteria employed by 5 different raters are illustrated here. Stage scores for this performance were either formal operational or, in one case (rater HCSS 1), transitional to systematic. Rater SISS 1 appeals to the conceptual content layer by focusing on which aspects of the moral are highlighted in the performance. This is seen as evidence that the performance is not systematic, at which level the subject, according to this rater, would have focused on the right instead of character. Rater SISS 2 and HCSS 2 refer to the surface structure level when they focus, respectively, on the conventionality (versus pre- or post-conventionality) of the performance, and its social perspective. Raters HCSS 1 and HCSS 4, who focus explicitly on the organizational and logical structure of the performance, present the core structure level.

One of the questions raised by this look at rater behavior concerns the extent to which the three different approaches to scoring reflect, at some level, a common core. The central diagram shown in Figures 1 and 2 suggests such a core, but the examples in these figures, while they demonstrate that raters employ different levels of structure in their analyses, do not explicitly tie together the three levels of structure. To do this, it must be demonstrated that the strong relationships (found in earlier research) among stage performance estimates derived from the different stage scoring systems (Dawson, 1998b) can be explained in terms of a shared structural core.

Paradoxically, a comment from one of the rater respondents, which was intended to point out the differences between scoring systems, may provide the key. In the midst of scoring the statement represented in Figure 1, this respondent explained, "I could score this statement on the basis of its cognitive structure—the form of the argument—but I'm scoring it in the good life domain, and cognitive and good life scoring are not the same thing" (rater SISS 2). This rater, however, was within one stage of agreement with the 6 raters employing the Hierarchical Complexity Scoring System over 90% of the time, indicating that, at least from the standpoint of awarding stage scores, the Good Life Scoring System, Standard Issue Scoring System and the more formal Hierarchical Complexity Scoring System accomplish the same thing.

The position taken by rater SISS 2 is similar to that of Kohlberg (1984), with respect to moral stages. Kohlberg set out to demonstrate that there were specifically moral stages of cognition, distinct from (though not independent of) logico-mathematical structures. For this reason, his scoring system was based on moral content rather than directly upon cognitive organization or processes. Consequently, though he was strongly influenced by Piagetian notions of development, his scoring system was, arguably, more dependent on an empirical analysis of the conceptualizations of his longitudinal sample than on fundamental Piagetian principles. Dawson (2002) has argued elsewhere that Kohlberg's scoring system increasingly reflected a Piagetian structural analysis as it was refined over the decades. Earlier versions, which were more dependent on particular philosophical conceptualizations, failed to yield results that were consistent with actual developmental trends, necessitating re-analysis of the scoring criteria. The final scoring system (Colby & Kohlberg, 1987b) incorporates notions of conceptual complexity and perspective (which I have placed at the surface structure level) into the scoring criteria to a much greater extent than earlier versions. However, as noted above, scoring with the Standard Issue Scoring System involves matching arguments from performances with similar arguments in a scoring manual, which does not require the rater to make a structural analysis of performances.

Interestingly, the two raters trained in the Standard Issue Scoring System who participated in this study did not refer directly to either Kohlberg's or Armon's scoring manuals as they scored. Instead, they predominantly specified principles that underlie these scoring systems, such as social perspective, levels of conventionality, and transformations. Both of these raters participated in the development of these scoring systems, which may make their ratings more formal than those of typical Standard Issue raters, influencing the present results. For the purposes of this analysis, I assume that the performances of the present raters are similar to those of Standard Issue or Good Life raters as a whole.

So, how are the criteria assigned to the three levels of structure proposed here related conceptually, such that they can be considered as layers of structure? First, the developmental history of the Standard Issue Scoring System (Colby & Kohlberg, 1987a) provides evidence that the conceptual and surface structure levels are conceptually related, in that the

stage assignment of exemplars in the scoring manual was, at least to some extent, dependent upon the identification of surface structure features like social perspective and conventionality. Second, the behavior of the two Standard Issue raters who participated in this study indicates that these surface structure features can be employed as scoring criteria. Third, the 6 raters in this study who reported using the Hierarchical Complexity Scoring System, employed both surface structure and core structure criteria (with a few references to conceptual content) indicating a relationship between these levels of structure. The first and second points are discussed above, but the third point requires further explication. An example of the scoring process of one of the Hierarchical Complexity Scoring System raters will help. Here, rater HCSS 2 responds to the protocol, Good Education 0212:

> Well, this isn't a very sophisticated notion of the role of education in the good life. Especially because, at first, I thought that he was saying that you'd be richer, money-wise (laughter), with an education. That would make richer a concrete notion, but I see that it's actually at least abstract, because it's related to this idea of open-mindedness. It seems there are two variables [richer life, open mind] that are in a logical relation to one another…as in'"If you get a good education, your mind will be more open and therefore you will have a richer life." This would make it at least formal, but could it be higher than that. Well,"*richer life* could be higher than abstract, and so could *open mind,* so I'm looking for evidence that they are…but the perspective here is of the individual person and his life, without bringing in anyone else's perspective, or a social perspective, so you can't say, really. Formal, I'll stick with formal.

In this example, the rater appeals to all three levels of structure. The content level is employed in her initial attempt to understand the argument, and again when she double-checks her understanding at the end. The surface structure level is briefly included when she examines the social perspective of the respondent to see if there are grounds for considering the possibility that the statement is systematic. The core structure is reflected in her analysis of the logical structure of the argument.

This rater's performance is a good illustration of the approach most often taken by the Hierarchical Complexity raters in this study. The analysis of the core level of structure appears to be at least somewhat dependent on interpretations of meaning at the content level, as well as features of surface structure that are associated with particular stages. Interestingly, some of the most experienced scorers seemed to get a bit "lazy" about their analyses of core structure, relying, instead, on past experience with strong associations between features of core and surface structures.

Discussion

This paper attempts to explain the high level of agreement among raters employing three different scoring systems, by positing that core structural features influence stage ratings in all three systems, whether or not these principles are made explicit in a given scoring system. Several pieces of evidence are brought together in this effort: (1) strong correlations between stage scores assigned with two different types of scoring systems; (2) the high rate of agreement among the raters employing these scoring systems; (3) the identification of three distinct levels of structural criteria across rating performances; and (4) the identification of multiple levels of structural criteria within rating performances.

The accumulating evidence that the Hierarchical Complexity Scoring System, Standard Issue Scoring System, and Good Life Scoring System assess the same dimension of ability is increasingly compelling. However, in the present instance, it is important to keep in mind that the interview technique was open-ended and conversational. Consequently, it is likely that the stage scores awarded by respondents were, at times, influenced by the interviewer, which may have generated more agreement among raters than would ordinarily be found. Moreover, though the rate of agreement within a stage was high across all 8 raters, there was still substantial disagreement. Dawson (1998) has speculated that this disagreement may be due to differential reliance upon the presence or absence of specific conceptual content in performances. For example, it is possible that a core structure might be present in a performance while a particular conceptualization, necessary for stage scoring with a concept-matching system, is not. A Hierarchical Complexity rater might score such a performance at a higher stage than a Standard Issue rater. Alternatively, a concept might appear to be present in a performance when it is actually being parroted at a lower level of hierarchical complexity. A Hierarchical Complexity rater might score such a perfor-

mance at a lower level than a Standard Issue rater.

The fact that hierarchical complexity ratings tend to be a bit lower than Standard Issue and Good Life ratings suggests that the more domain-specific systems, by requiring evidence related to surface structure and content, assess an additional dimension related to their content domains. However, the cumulative evidence indicates that over 80% of the variance between scores awarded with the hierarchical complexity scoring system and these other scoring systems is explained by a shared dimension. Taking into account the random noise introduced by measurement error, this leaves little variance for the content dimension to explain—arguably, too little to support the claim that good life and moral reasoning represent unique structures.

One way to think about these results is to consider that the Good Life Scoring System and Standard Issue Scoring System, rather than measuring unique domain-specific stage structures, actually assess *hierarchical complexity* within the moral and good life domains, while adding surface structure and content requirements for stage assignment. To investigate this possibility, I am currently piloting a scoring system that first determines the hierarchical complexity of texts, then separately examines conceptual differentiation within the assessment domain. One can think of the hierarchical complexity dimension as a vertical dimension representing hierarchical integration, and the content dimension as a horizontal dimension representing differentiation. The vertical dimension tells us what kind of "hard" structures a respondent has applied to the problems posed by an interviewer. The horizontal dimension shows the extent to which these structures are reflected in conceptual differentiation specific to the assessment domain. For example, an individual may apply formal structures to reasoning about a science problem but lack the necessary domain-specific conceptual knowledge to produce an adequate solution. Developmentally, this individual has very different needs than another individual who demonstrates neither the conceptual differentiation nor the formal structures required to reach a satisfactory solution.

While additional research is required to conclude with confidence that the Hierarchical Complexity Scoring System produces stage assessments that are empirically and interpretively consistent with those of more content-bound systems, the evidence we have gathered is increasingly compelling. As this evidence mounts, I am increasingly enthusiastic about the implications. A domain-general scoring system confers several advantages. First, a generalized scoring system like the Hierarchical Complexity Scoring System is easier to learn than content-bound systems like the Standard Issue Scoring System and Good Life Scoring System.

Second, domain-general assessment makes it possible to score a wide variety of texts, employing identical criteria across knowledge domains and contexts. This permits meaningful cross-domain and cross-context comparisons of developmental progress. Third, domain-general assessment permits the independent analysis of hierarchical structure and conceptual content. Overton (1998) has called for research methodologies that intentionally explore developmental processes from multiple perspectives. Domain-general assessment helps make this possible by "isolating" the hierarchical dimension and permitting us to treat hierarchical structure and conceptual content as different perspectives on development. By analyzing these dimensions independently and then exploring their interrelationship, we can substantially improve our understanding of developmental processes. My colleagues and I have already employed this approach to construct descriptions of conceptual development in the evaluative reasoning and moral domains (Dawson, 1998; Dawson & Gabrielian, in review). We have also used a similar approach to study the relationship between language development and hierarchical development. One result of this project is the first reliable and accurate computerized developmental assessment system (Dawson, 2002, January; Dawson & Wilson, in prep).

Acknowledgements

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

References

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

Armon, C. (1984b). *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.

Case, R. (1991). *The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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* (Vol. 23, pp. 71-100). Basel, Switzerland: S. 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.

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 City.

Commons, M. L., Straughn, J., Meaney, M., Johnstone, J., Weaver, J. H., Lichtenbaum, E., Sonnert, G., & Rodriquez, J. (1995, November). *Measuring stage of development across domains: A universal scoring scheme.* Paper presented at the Annual meeting of the Association for Moral Education.

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.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* (pp. 85-116). New York: Cambridge University Press.

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

Damon, W. (1980). Patterns of change in children's social reasoning: A two-year longitudinal study. *Child Development, 51,* 1010-1017.

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. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development.*

Dawson, T. L. (2002, January). *Measuring intellectual development across the lifespan.* Paper presented at Powerful learning & the Perry scheme: Exploring intellectual development's role in knowing, learning, and reasoning. California State University, Fullerton, CA.

Dawson, T. L. (in press). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement.*

Dawson, T. L., Commons, M. L., & Wilson, M. (in review). The shape of development. *Developmental Psychology.*

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

Dawson, T. L., & Kay, A. (in review). A stage is a stage is a stage: A direct comparison of two scoring systems.

Dawson, T. L. & Wilson, M. (in prep). The LAAS: A computerized developmental scoring system for small- and large-scale assessments.

Dawson, T. L., Xie, Y., & Wilson, J. (in prep). A multidimensional item response model comparing two developmental assessment systems.

Demetriou, A., & Efklides, A. (1994). Structure, development, and dynamics of mind: A meta-Piagetian theory. In A. Demetriou & A. Efklides (Eds.), *Intelligence, mind, and reasoning: Structure and development. Advances in psychology* (pp. 75-109). Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers.

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* (5 ed., pp. 467-561). New York: John Wiley & Sons.

Fischer, K. W., Hand, H. H., Watson, M. W., van Parys, M., & Tucker, J. (1984). Putting the child into socialization: The development of social categories in preschool children. In L. Katz (Ed.), *Current topics in early childhood education* (Vol. 5, pp. 27-72). Norwood, NJ: Ablex.

Halford, G. S. (1999). The properties of representations used in higher cognitive processes: Developmental implications, *Development of mental representation: Theories and applications.* (pp. 147-168). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., Publishers.

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.

Killen, M. (1991). Social and moral development in early childhood. In J. L. G. William M. Kurtines (Ed.), *Handbook of moral behavior and development, Vol. 1: Theory; Vol. 2: Research; Vol. 3: Application.* (pp. 115-138): Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

King, P. M., Kitchener, K. S., Wood, P. K., & Davison, M. L. (1989). Relationships across developmental domains: A longitudinal study of intellectual, moral, and ego development. In M. L. Commons & J. D. Sinnot & F. A. Richards & C. Armon (Eds.), *Adult development. Volume 1: Comparisons and applications of developmental models* (pp. 57-71). New York: Praeger.

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 NY: Praeger.

Kohlberg, L., & Armon, C. (1984). Three types of stage models in the study of adult development. In M. L. Commons & F. A. Richards & T. A. Grotzer & J. D. Sinnot (Eds.), *Beyond formal operations: Vol 1. Late adolescent and adult cognitive development* (pp. 357-381). New York: Praeger.

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

Overton, W. (1998). Developmental psychology: Philosophy, concepts, and methodology. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development.* New York: John Wiley & Sons.

Piaget, J. (1965). *The moral judgment of the child* (M. Gabian, Trans.). New York: The Free Press.

Piaget, J. (1973). The theory of stages in cognitive development, *Measurement and Piaget.* U. Geneva, Switzerland.

Piaget, J., Garcia, R., & Feider, H. t. (1989). *Psychogenesis and the history of science.* Columbia University Press; New York.

Sonnert, G., & Commons, M. L. (1994). Society and the highest stages of moral development. *Politics & the Individual, 4,* 31-55.

Xie, Y., & Dawson, T. L. (2000, April). *Multidimensional models in a developmental context.* Paper presented at the International Objective Measurement Workshop, New Orleans, LA.