EDUCATIONAL ASSESSMENT, *9*(3&4), 153–191
Copyright © 2003–2004, Lawrence Erlbaum Associates, Inc.

The LAAS: A Computerized Scoring System for Small- and Large-Scale Developmental Assessments

Theo L. Dawson
Cognitive Science
Hampshire College

Mark Wilson
Graduate School of Education
University of California at Berkeley

The evaluation of developmental interventions has been hampered by a lack of practical, reliable, and objective developmental assessment systems. This article describes the construction of a domain-general computerized developmental assessment system for texts: the Lexical Abstraction Assessment System (LAAS). The LAAS provides assessments of the order of hierarchical complexity of oral and written texts, employing scoring rules developed with predictive discriminant analysis. The LAAS is made possible by a feature of conceptual structure we call *hierarchical order of abstraction*, which produces systematic quantifiable changes in lexical composition with development. The LAAS produces scores that agree with human ratings of hierarchical complexity more than 80% of the time within one-third of a complexity order across 6 complexity orders (18 levels), spanning the portion of the lifespan from about 4 years of age through adulthood. This corresponds to a Kendall's tau of .93.

This article describes the development of the Lexical Abstraction Assessment System (LAAS). The LAAS provides assessments of the developmental level of oral and written texts, employing scoring rules developed with predictive discriminant analysis. As we will demonstrate, it produces scores that agree with human ratings more than 80% of the time, within one-third of a level across six levels, spanning

Requests for reprints should be sent to Theo L. Dawson, Cognitive Science, Hampshire College, Amherst, MA 01002. E-mail: tdawson@hampshire.edu

the portion of the lifespan from about 4 years of age through adulthood. The LAAS is made possible by a feature of conceptual structure we call *hierarchical order of abstraction*, which produces systematic quantifiable changes in lexical composition with development.

The evaluation of developmental interventions has been hampered by the lack of practical, reliable, and objective developmental assessment systems. Most cognitive-developmental evaluation methods are expensive and time-consuming to implement. Moreover, they are task- and domain-specific, which means that the range of behaviors that can be evaluated developmentally is restricted. These limitations make developmental assessments impractical for most real-world applications.

To address these problems, we first investigated the possibility of domain-general developmental assessment.

THE HIERARCHICAL COMPLEXITY SCORING SYSTEM

Several attempts have been made to develop a generalized developmental assessment system for human raters. Indeed, Piaget (XXXX) defined each of his developmental stages in generalized terms. Conservation, for example, is a general feature of concrete operations and can be observed on a wide range of tasks. Case (Case, Griffin, McKeough, & Okamoto, 1992) and Fischer (Fischer & Bidell, 1998) and their colleagues have employed generalized definitions extensively to scale performances across domains, but neither has disseminated a generalized scoring system. Based primarily on Commons' General Stage Scoring System (Commons et al., 1995) and Fischer's (1980) skill theory, the Hierarchical Complexity Scoring System (Dawson, 2003), employed here, lays out explicit general criteria for determining the developmental level of performances in any domain of knowledge.

Commons, Trudeau, Stein, Richards, and Krause's (1998) General Stage Model (sometimes referred to as the Model of Hierarchical Complexity), on which Commons based his General Stage Scoring System (Commons et al., 1995), is a model of the hierarchical complexity of tasks. The model specifies 15 stages (subsequently referred to as *complexity orders*). The sequence is 0, computory; 1, sensory and 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. Complexity orders 0 to 12 correspond definitionally to Fischer's (1980) skill levels (Table 1), which include single reflexive actions, reflexive mappings, reflexive systems, single sensorimotor actions, sensorimotor mappings, sensorimotor systems, single representations, representational mappings, representational systems, single abstractions, abstract mappings, abstract systems, and single principles/axioms. Orders of hierarchical complexity also correspond definitionally to the stages

TABLE 1
Correspondences Between Commons's (Commons, et al., 1995) Stages and Fischer's (1980) Skill Levels

Number	*Commons*	*Fischer*
0	Computory	Single reflexive actions
1	Sensory & motor	Reflexive mappings
2	Circular sensory-motor	Reflexive systems
3	Sensory-motor	Single sensorimotor schemes
4	Nominal	Sensorimotor mappings
5	Sentential	Sensorimotor systems
6	Preoperational	Single representations
7	Primary	Representational mappings
8	Concrete	Representational systems
9	Abstract	Single abstractions
10	Formal	Abstract mappings
11	Systematic	Abstract systems
12	Metasystematic	Single principles/axioms

and substages more recently described by Noelting and Rousseau (in press). Sonnert and Commons' (1994) paradigmatic and cross-paradigmatic stages (13 and 14) are largely hypothetical. Because Fischer's skill level labels are not only familiar to more readers, but also more descriptive of the structures observed in text performances than Commons' (XXXX) stage labels, we employ them here as names for complexity orders 0 to 12.[1]

Not only are there definitional correspondences among analogous levels described by Commons, Fischer, and Noelting, (XXXX) there is empirical evidence of correspondences between the levels determined with the Hierarchical Complexity Scoring System and those determined with at least four content-based systems, including Kitchener and King's (1990) stages of reflective judgment (Dawson, 2003; Kitchener, Lynch, Fischer, & Wood, 1993), Armon's Good Life stages (Dawson, 2002), Perry's epistemological positions (Dawson, in press-a), and Kohlberg's moral stages (Dawson, in press-b).

[1]Dr. Fischer (personal communication, September 23, 2002) has agreed to the use of these labels. Skill levels are more descriptive of psychological processes than complexity orders, in that skill theory posits a special relationship between skill levels as observed in behavior and the development of the brain. Hierarchical Complexity Theory, as it is presented here, is focused on behavior and its measurement, which is why we call the levels complexity orders rather than skill levels, even though we are using the same labels. We consider the development of a good developmental ruler to be essential to progress in developmental psychology, including our ability to address the question of the relationship between behavior and changes in brain activity or structure. In a sense, Hierarchical Complexity Theory has been developed to serve skill theory and other cognitive-developmental theories.

When assessing the hierarchical complexity of a text with Dawson's (2003) Hierarchical Complexity Scoring System, the rater attends to two manifestations of hierarchical complexity. The first is its conceptual structure, embodied in the *hierarchical order of abstraction*[2] of the new concepts employed in its arguments, and the second is the most complex *logical structure* of its arguments. Note that conceptual and logical structures are definitionally identical and fundamentally interdependent. We make a distinction between the two types of structure for heuristic and pragmatic reasons. When scoring texts, hierarchical order of abstraction refers primarily to the structure of the elements of arguments, which often must be inferred from their meaning in context, whereas logical structure refers to the explicit way in which these elements are coordinated in a given text.

Each complexity order is associated with a primary hierarchical order of abstraction (reflexive actions, sensorimotor schemes, representations, abstractions, or principles) and a secondary hierarchical order of abstraction (i.e., first-order representations, second-order representations, third-order representations). Each complexity order is also associated with one of three logical forms (elements, mappings or relations, and systems). The hierarchical order of abstraction and logical structure of each complexity order are described in Appendix A, which includes only the eight complexity orders identified in the data employed in the analyses that follow. The examples provided in these descriptions are drawn primarily from Dawson and Gabrielian's (2003) analysis of the conceptions of authority and contract associated with complexity orders in a large sample of moral judgment interviews scored with the Hierarchical Complexity Scoring System. Appendix B describes the scoring process.

Reliability and Validity of the Scoring System

We have undertaken several studies of the reliability and validity of the Hierarchical Complexity Scoring System. First, Dawson and her colleagues (Dawson, Commons, & Wilson, 2003) employed Rasch scaling to investigate patterns of performance in a cross-sectional life-span sample of 602 moral judgment performances scored with the Hierarchical Complexity Scoring System. They found six complexity orders—representational mappings, representational systems, single abstractions, abstract mappings, abstract systems, and single principles/axioms—represented in performance between the ages of 5 and 86 years. The reliability of the scale was assessed through statistical modeling and by examining interrater agreement rates. A Rasch analysis, which provides a reliability estimate that is equivalent to

[2]The word *abstraction* as used in the term *hierarchical order of abstraction* refers to the way in which conceptions increase in generality over the course of development. The concepts that occur for the first time at the single abstractions complexity order are abstract in a more particular sense; the new conceptions of this complexity order are defined in terms of qualities that are increasingly detached from the concrete.

Cronbach's alpha coefficient, provided a reliability of .97. Interrater agreement rates ranged from 80% to 97% within half of a complexity order.

In the data collected for this study, the ages at which the first four complexity orders—representational mappings to abstract mappings—first predominated were 5, 7, 9, and 14 years, respectively. The highest complexity orders—abstract systems and single principles—did not become the plurality until 22 years of age with 3 years of college, and 26 years of age with 3 years of post-graduate work. These age ranges are very similar to those reported by Fischer and Bidell (1998) for the acquisition of skill levels. The authors also showed that the relationship between age and developmental level becomes weaker and less deterministic as age increases, and that, from adolescence through adulthood, educational attainment is a better predictor of developmental level than age.

In the same article, Dawson and her colleagues (Dawson, Commons, & Wilson, 2003) reported that the sequence of acquisition of the complexity orders appears to be invariant and wave-like, with evidence of spurts and plateaus at every transition measured, including the transitions to the two highest complexity orders, which were identified almost exclusively in adulthood. They also demonstrated that patterns of performance in their sample are highly consistent from complexity order to complexity order. They argued that these systematic patterns provide evidence of construct validity in that they are consistent with the postulates of cognitive developmental theory—in particular, the dynamic systems account of development described by Fischer and Bidell (1998).

As previously mentioned, we have conducted numerous comparisons of the Hierarchical Complexity Scoring System with other developmental scoring systems (Dawson, 2002; Dawson, Xie, & Wilson, 2003). These studies provide additional evidence of construct validity, in that they show that the Hierarchical Complexity Scoring System and four other developmental assessment systems—Kohlberg and his colleague's (Colby & Kohlberg, 1987b) Standard Issue Scoring System, Armon's (1984) Good Life Scoring System, Kitchener and King's (1990) Reflective Judgment Scoring System, and a scoring system based on Perry's (1970) epistemological development sequence (Dawson, in press-a)—predominantly assess the same dimension of performance. However, the Hierarchical Complexity Scoring System displays greater internal consistency and provides results that are more consonant with the postulates of cognitive developmental theory than these domain-based scoring systems. Overall, these studies provide evidence that the Hierarchical Complexity Scoring System is a valid and reliable measure of intellectual development from early childhood through adulthood.

Hierarchical Order of Abstraction

Given the evidence from this series of studies, we conclude that the Hierarchical Complexity Scoring System is a reasonably reliable and valid measure of cognitive development. Fortunately, it also provides us with a basis for the development of computerized scoring rules: the construct hierarchical order of abstraction. As already mentioned, it is possible to score the hierarchical complexity of text performances 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) suggested 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 single sensorimotor actions, single representations, single abstractions, and single principles complexity orders, the new concepts not only coordinate or modify constructions from the previous complexity order, but also are qualitatively distinct conceptual forms—sensorimotor actions, representations, abstractions, and principles, respectively (Fischer, 1980). The appearance of each of these conceptual forms ushers in three repeating logical forms—single elements, mappings or relations, and systems (Fischer, 1980). 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. Other researchers have observed and described similar conceptual forms and repeating logical structures (Case, Okamoto, Henderson, & McKeough, 1993; Fischer & Bidell, 1998; Piaget & Garcia, 1989).

Because the three logical forms repeat themselves several times over the course of development, the logical form of arguments does not provide adequate evidence for hierarchical complexity scoring. The logical form of a text is primarily evident in syntax. This means that syntax itself is unlikely to provide a sufficient basis for computerized scoring. For this reason, combined with the difficulties associated with parsing texts, we abandoned the notion of constructing computerized scoring criteria on the basis of syntactic change. However, unique manifestations of hierarchical order of abstraction are evident at every complexity order in the form of new conceptual understandings. Many of these new meanings are embodied in lexical items. For example, the abstract mappings conception of *honor* is constructed from single abstractions conceptions of reputation, honesty, and fairness. The word *honor* rarely appears in texts before the abstract mappings order, and the words *reputation, honesty,* and *fairness* rarely appear in texts before the single abstractions order. On the basis of observations like these, we hypothesized that each complexity order might be associated with a lexicon, and that lexical items from each lexicon might be systematically distributed in text performances in ways that would permit the development of computerized scoring rules.

Another way in which hierarchical order of abstraction should be evident in texts is in the mean word length of vocabulary. Because many morphemes (though not all) stand for more abstract meanings than root words, and most words containing multiple morphemes are more hierarchically complex by definition (they link two concepts to construct a new meaning), an increase in mean word length should reflect increasing hierarchical order of abstraction. For example, the suffix *-ness* raises the hierarchical order of abstraction of a number of concepts such as fair, kind, playful, and neat, and the terms *self-understanding* and *social contract* are constructed on the basis of earlier concepts of self, understanding, society, and contract.

In the following sections we describe our sample, scoring procedures, and the development of a lexical index for each complexity order. This process involved several steps: gathering a large set of appropriate texts, scoring the texts for hierarchical complexity, extracting the vocabulary for each case, sorting the cases by complexity order, combining the vocabulary of texts scored at the same complexity order, and constructing lists (lexical indexes) of the unique vocabulary associated with each complexity order. These steps are described in more detail later. We then show how the lexical indexes and mean word length were employed to calibrate the LAAS.

METHOD

Sample

All of the 1,014 texts used in this project were selected from a larger corpus of texts collected by several researchers between 1955 and the present. There were two main criteria for selection. First, the texts had to be legible enough to submit to analysis. Many of the texts in the larger corpus were scanned from old typewritten manuscripts. Some of the original manuscripts were of poor quality, preventing the efficient use of character recognition software. These were rejected. Second, selected texts had to meet length criteria. These were that the vocabulary employed in each text must exceed 100 words at or below the representational mappings order; 150 words at the single abstractions order; 200 words at the abstract mappings order; 250 words at the abstract systems order; and 350 words at the single principles order. These length criteria were established on the basis of preliminary examinations of the data, which suggested that vocabulary samples shorter than these were inadequate for lexical analysis.

A few of the interviews in the sample were of the same individual, interviewed on more than one occasion. The decision to include these longitudinal cases was based on the need to adequately represent certain age groups and is justified by the long intervals (2 to 4 years) between interviews (Willett, 1989).

In putting together our database, we took care to find an adequate number of cases in each of the represented age groups. Consequently, some texts in the larger corpus were selected because they helped fill in age cells, and some were not selected simply because they represented age groups for which we already had adequate data. In every sense, this is a convenience sample. The subsample sizes and sources are shown in Table 2.

TABLE 2
Subsample Sizes and Sources

Sample size	*Source*	*Description*
116	Armon & Dawson, 1997	A study of moral reasoning and reasoning about the good
119	Dawson, 2001	A study of moral reasoning and reasoning about education
26	Commons, Danaher, Miller, & Dawson, 2000	A study of moral reasoning and reasoning about education among Harvard professors and students
120	Colby, Kohlberg, Gibbs, & Lieberman, 1983	A study of the moral reasoning of boys
24	Ullian, 1977	A study of the moral reasoning of elementary school students
163	Walker, 1989	A longitudinal study of the moral reasoning of children and their parents
32	Drexler, 1998	A study of the moral reasoning of children
45	Berkowitz, Guerra, & Nucci, 1991	A study of the moral reasoning of adolescents
42	Walker, Gustafson, & Hennig, 2001	A study of the moral reasoning of children
71	Miller, Olds, & Pavelich, 2001; Pavelich, Miller, & Olds, 2002	A study of reflective judgment
66	Mensing, 2002	A study of moral and conventional judgments of adolescents and their parents
60	Bates et al., 1988; Carlson-Luden, 1979	A study of early language development
11	Bloom, 1970; Bloom & Lightblown, 1974; Bloom, Lightblown, & Hood, 1975	Studies of early language development
38	Hall, Nagy, & Linn, 1984; Hall, Nagy, & Nottenburg, 1981; Hall & Tirre, 1979	A study of socioeconomic influences on language development
12	Demetras, 1989a; 1989b	Studies of parent-child interaction
37	Henry, 1995; Wilson & Henry, 1998	A study of early language development
7	MacWhinney, 2000	A study of early language development
3	Snow (in MacWhinney, 2000)	A study of early language development
23	Brown, 1973	A study of early language development
5	Higginson, 1985	A study of early language development

Note. Data from the last nine studies were obtained from the CHILDES database (MacWhinney, 2000).

The populations sampled by these various studies were diverse, representing a wide range of socioeconomic and ethnic groups. It is not possible to report a consistent account of these, however, because discrepant reporting methods were employed in the various studies. All of the texts were scored with the Hierarchical Complexity Scoring System. Each of the texts was randomly assigned to either a training condition or a test condition (half to each). The texts assigned to the training condition were employed to develop scoring criteria, whereas the texts assigned to the test condition were employed to evaluate these scoring criteria. Tables 3 and 4 show the breakdown of test and training cases by source and age. Some age and education data are missing. A detailed breakdown of age distributions by subsample is shown in Table 5.

Hierarchical Complexity Scoring

Hierarchical Complexity Scoring involves identifying both the highest hierarchical order of abstraction and the most complex logical form in text performances. A text is considered to be at a given complexity order if its elements embody the hierarchical order of abstraction of that complexity order, and its logical structure meets the formal requirements of that complexity order. Once the unit for analysis

TABLE 3
Distribution of Cases Across Construction and Test Conditions by Study

Study	*Training*	*Test Case*	*Total*
Armon (1984)	58	58	116
Dawson (XXXX)	60	59	119
Commons (XXXX)	13	13	26
Kohlberg (XXXX)	60	60	120
Ullian (1997)	12	12	24
Walker (1989)	81	82	163
Drexler (1998)	16	16	32
Berkowitz, Guerra, & Nucci (1991)	23	22	45
Walker et al. (2001)	21	21	42
Pavelich, Miller, and Olds (2002)	49	49	98
Mensing (2002)	35	35	70
Bates et al. (1988)	11	12	23
Bloom (1970)	6	5	11
Hall (XXXX)	19	19	38
Demetras (XXXX)	6	6	12
Henry (1995)	18	19	37
MacWhinney (2000)	4	3	7
Snow (in MacWhinney, 2000)	1	2	3
Brown (1973)	12	11	23
Higginson (1985)	2	3	5
Total	507	507	1014

TABLE 4
Distribution of Cases Across Construction and Test Conditions by Age

Age	*Training*	*Test Case*	*Total*
2	22	22	44
3	24	28	52
4	29	27	56
5	12	12	24
6	20	18	38
7	15	15	30
8	13	17	30
9	13	8	21
10	22	22	44
11	10	16	26
12	11	12	23
13	14	8	22
14	11	13	24
15	14	11	25
16	9	11	20
17	16	11	27
18	22	14	36
19	10	9	19
20	12	12	24
21–22	21	23	44
23–24	5	8	13
25–26	18	13	31
27–28	6	9	15
29–30	9	7	16
31–32	11	12	23
33–34	14	16	30
35–36	17	19	36
37–38	13	11	24
39–40	11	16	27
41–45	26	25	51
46–50	16	11	27
51–55	9	11	20
56–60	4	8	12
61–65	5	4	9
66–86	6	8	14
Total	490	487	977

has been determined, the hierarchical complexity rater first identifies the most hierarchically abstract elements in a text, and then examines the way in which these elements are employed in arguments. For example, if the highest hierarchical order of abstraction represented in the text is representations, and these are employed in arguments or propositions that take the form of mappings, the text is considered to be at the representational mappings order. Secondary orders of abstraction (i.e., first-order representations, second-order representations, and third-order representations) are rarely employed in human scoring, because it is enough to know the primary order of abstraction when logical structure is also evaluated. Later in this article, when we discuss the development of computerized scoring, the secondary orders of abstraction are of central importance. Appendix B provides examples of hierarchical complexity scoring.

Interrater agreement was high. We found correlations of .95 to .98 among the scores of four independent raters on a subset of 112 randomly selected texts. In this group of raters, agreement rates ranged from 80% to 97% within half a complexity order and from 98% to 100% within a full complexity order. These rates of interrater agreement equal or exceed interrater agreements commonly reported for developmental assessments (Armon, 1984; Colby & Kohlberg, 1987a; Kitchener & King, 1990).

In scoring the texts employed in this study, data from the various subsamples were treated somewhat differently. Our first problem was to determine what we would regard as a scorable segment of text. Because the various data sources provided different types of texts, it was not possible to establish a single criterion for segmenting all of the data. For example, we were unable to establish a consistent method for dividing up the texts from spontaneous speech studies, because the texts were generally short and there were often no shifts of topic or theme to suggest natural divisions. We therefore decided (somewhat reluctantly) to score the texts holistically, assigning each text to a single complexity order and transition—early transitional (some evidence of reasoning at the complexity order following the dominant complexity order), late transitional (reasoning predominantly at the dominant complexity order), and consolidated (reasoning almost exclusively at the dominant complexity order).

In contrast, moral judgment interviews were subdivided into protocols, the preferred method specified in the Hierarchical Complexity Scoring Manual (Dawson, 2003). Each protocol included the complete response to a standard probe question and associated follow-up questions. Each probe question was scored individually, and mean scores were calculated from the protocol scores awarded to each performance.

Evaluative-reasoning-about-good-education texts were also divided into protocols, but it was necessary to employ different criteria than in the case of the moral judgment interviews. Because the good-education interviews were open-ended, there were no standard probes. Consequently, the good-education interviews were divided into protocols representing the individual judgments about education made by each respondent in response to nonstandard probes. Each protocol included the complete justificatory argument for a given judgment. For example, if a respondent stated that good education "is one in which the teacher engages students," the protocol would include this statement along with the argument used to support it. Protocols were scored individually, and a mean score was calculated for

TABLE 5
Age Distribution by Subsample

Age	*Study* Armon (1984)	Dawson (XXXX)	Commons (XXXX)	Kohlberg (XXXX)	Ullian (1999)	Walker (1989)	Drexler (1998)	Berkowitz, Guerra, Nucci (1991)	Walker et al. (2001)	Pavelich, Miller, & Olds (2002)
2										
3										
4										
5		18								
6		23			5	2	7			
7		8				9	13			
8	2	11			9	2	6			
9	2	2				9	6		2	
10		3		9	10	5			17	
11	2	2				2			20	
12	2	3				15			3	
13	2	2		2		5		2		
14	2			15				4		
15	4	2				7		6		
16	1	3		4		5		2		
17	4	1		15		5		2		
18	1	10		6		7		2		10
19	3	6		3				2		5
20	2	3	2						17	
21–22	1		1	15				22	5	
23–24	5	3		3					2	
25–26	2			11				15	3	
27–28	3		1	8				1	2	
29–30	1	1		8				2	4	
31–32	9		1	6		2		2	3	
33–34	5			8		6		5	6	
35–36	10	1		7		11		5	2	
37–38	5	2				11		3	3	
39–40	5		1			18		1	2	
41–45	16	3	3			22		4	3	
46–50	4	1	3			11		2	6	
51–55	3	2	5			6		2	2	
56–60	3	1	5			3				
61–65	5	1	3							
66–86	12	1	1							
Total	116	113	26	120	24	163	32	45	42	71

TABLE 5
Age Distribution by Subsample *(Continued)*

Study										
Mensing (2002)	*Bates et al. (1998)*	*Bloom (1970)*	*Hall (XXXX)*	*Demetras (XXXX)*	*Henry (1995)*	*MacWhinney (2000)*	*Snow (2000)*	*Brown (1973)*	*Higginson (1985)*	*Total*
16	8	4	5	4			5	2	44	
4	3	5	6	15	3	2	11	3	52	
3		25	1	18	2	1	6		56	
			4		1		1		24	
						1				38
										30
										30
										21
										44
										26
										23
9										22
3										24
6										25
5										20
										27
										36
										19
										24
										44
										13
										31
										15
										16
										23
							30			30
										36
										24
										27
										51
										27
										20
										12
										9
										14
66	23	11	38	12	37	7	3	23	5	977

each case. Finally, reflective judgment and moral/conventional interviews, which are composed of a number of subinterviews, were divided into protocols by subinterview. Each protocol was individually scored, and a mean score was calculated for each case.

The rationale for dividing interviews into several protocols includes two components. The first concerns reliability; we can be more confident of the score we award to a performance when we score several protocols than when we award a holistic score. However, this is true only if we are fairly certain that each protocol is a good representation of the individual's reasoning. Consequently, we either look for "complete arguments" or divide a text into large enough segments to be fairly certain that we will capture a representative sample of reasoning in each segment.

Mean complexity order scores (for the cases in which these were calculated) ranged from 0 to 12. Complexity order means were then divided into one of three levels per complexity order: early transitional, late transitional, and consolidated. The use of three levels was justified by the consistent finding that we can have confidence in hierarchical complexity scores within about one third of a complexity order (Dawson & Gabrielian, 2003; Dawson, Xie, & Wilson, 2003). Dividing complexity order scores into three levels was not as simple as determining that three levels are meaningful, however, because no nonarbitrary classification scheme suggested itself. Deciding to err cautiously, we began by considering scores of 0.0, 1.0, 2.0, 3.0, 4.0, and so on, to be the most consolidated cases because over 99% of the cases with these mean scores were consolidated at a single complexity order. (All of their scores were at the same complexity order.)

We then divided each complexity order into three (almost) equal segments. We began by considering a performance consolidated at a complexity order if it had no more than 16% of its performances at a higher or lower complexity order than the modal complexity order (16% + 16% = 32% of a complexity order). This means we assumed that a performance was consolidated at a given complexity order even though one or two of its protocols were assigned to a higher or lower complexity order. Performances with more than 16% and less than 85% of their performances at their higher complexity order were coded into two transitional levels as follows: We considered a performance to be in the late transition to a new complexity order if more than 50% and less than 85% of its responses were at its highest complexity order, and in the early transition to a new complexity order if more than 16% and less than 50% of its protocols were at its higher complexity order.

As previously noted, in developing these criteria we chose to err on the side of caution. This consideration is reflected in the fact that we assigned a case to the consolidated category if up to 16% of its protocols were assigned to a higher or lower complexity order than its modal complexity order. The assignment to a consolidated category actually involves a more stringent criterion than assignment to a transition category, because very few performances spanned more than two complexity orders. This means that, in the vast majority of cases, a consolidated performance, as defined here, is one in which more than 84% of its protocols are scored at the same complexity order. We adopted this cautious approach because, for research purposes that are only tangentially related to this article, we are particularly interested in differentiating between consolidated and transitional cases. In the remainder of this article, the various transitional categories and consolidated categories are referred to as levels.

In the interview data, eight complexity orders were identified: sensorimotor systems, single representations, representational mappings, representational systems, single abstractions, abstract mappings, abstract systems, and single principles/axioms. Each complexity order is represented by three transitional phases: an early transitional phase, a middle transitional phase, and a consolidated phase. In total, there are 22 levels, as shown in Table 6. Table 7 shows the distribution of data types by level.

TABLE 6
Abbreviations for the Names of Levels

Abbreviation	*Level*
SS3	Consolidated sensorimotor systems
SR1	1st transition to single representations
SR2	2nd transition to single representations
SR3	Consolidated single representations
RM1	1st transition to representational mappings
RM2	2nd transition to representational mappings
RM3	Consolidated representational mappings
RS1	1st transition to representational systems
RS2	2nd transition to representational systems
RS3	Consolidated representational systems
SA1	1st transition to single abstractions
SA2	2nd transition to single abstractions
SA3	Consolidated single abstractions
AM1	1st transition to abstract mappings
AM2	2nd transition to abstract mappings
AM3	Consolidated abstract mappings
AS1	1st transition to abstract systems
AS2	2nd transition to abstract systems
AS3	Consolidated abstract systems
SP1	1st transition to single principles
SP2	2nd transition to single principles
SP3	Consolidated single principles

TABLE 7
Distribution of Data Type by Level

Abbreviation	*Moral Judgment*	*Good Education*	*Moral Judgment and Good Education*	*Reflective Judgment*	*Spontaneous*	*Convention*	*Total*
SS3					9		9
SR1					14		14
SR2					21		21
SR3					29		29
RM1					35		35
RM2			1		40		41
RM3	1		24		8		33
RS1	4		7		1		12
RS2	10		9		1		20
RS3	22		9		1		32
SA1	30	2	3				35
SA2	35	3				1	39
SA3	42	8				10	60
AM1	19	2				9	30
AM2	33	3				10	46
AM3	98	2		27		31	158
AS1	39	3		16		6	64
AS2	60	3	1	14		3	81
AS3	109	6		16			131
SP1	22	4		4			30
SP2	15	15		8			38
SP3	26	17		13			56
Total	565	68	54	98	159	70	1014

Note. SS = sensorimotor systems; SR = single representations; RM = representational mappings; RS = representational systems; SA = single abstractions; AM = abstract mappings; AS = abstract systems; SP = single principles.

Abstraction Indexes

Because individual lexical items (individual words as well as hyphenated items, like *social contract* and *self-understanding*) stand for concepts that embody a hierarchical order of abstraction, it is possible to classify many lexical items according to the hierarchical complexity of the underlying concepts they represent. This makes it possible to classify individual lexical items according to their hierarchical order of abstraction. Because each complexity order is associated with a secondary order of abstraction, it is possible to classify lexical items into one list per complexity order. We hypothesized that a set of lists of this kind, which we called abstraction indexes, could be employed to assess the hierarchical complexity of texts, because patterns in the distribution of lexical items would differ from order to order. To test this hypothesis, we employed the training texts and an electronic version of *Webster's New Universal Unabridged Dictionary* (1983) to construct an abstraction index for each complexity order represented in our data.

Before beginning the construction of the abstraction indexes, we prepared both training case (the cases used to develop a scoring rule) and test case (the cases used to test the rule) data as follows: First, with text editing software specifically designed for this purpose, we removed all text that was not the actual speech of the intended participant from all of the files. We then created individual files containing the total vocabulary employed by each respondent, and sorted these by complexity order, based on the previously conducted hierarchical complexity scoring.

From the vocabulary files of the training cases only, we then created eight files composed of the total vocabulary found at each complexity order. The consolidated phase and the two preceding transitional phases were considered to comprise the complexity order. From these files we removed all concrete nouns (those referring to particular persons, places, or things or categories of persons places or things), colloquialisms, and nonwords. Concrete nouns were excluded, because the order of their acquisition is clearly linked to context. For example, the name *aardvark* may be a relatively late acquisition in the United States, but may occur much earlier in a country where aardvarks are commonly observed. We then employed our text analysis software to remove all of the words in the sensorimotor systems list from all of the lists representing the seven higher complexity orders. Next, we removed all of the words left in the single representations list from all of the lists representing the six higher complexity orders. Following this, we removed all of the words left in the representational mappings list from all of the lists representing the five higher complexity orders. We continued this process through the single principles order. At this point we had eight unique lists, derived empirically from the training texts we originally scored for hierarchical complexity. The lexical items in each of these lists were included simply because they occurred for the first time at the complexity order represented by the list.

Refining the Abstraction Indexes

We then analyzed these lists to determine whether the lexical items in the lists made sense from a hierarchical order of abstraction perspective. This was essential because of the small size of our sample, relative to the size of the lexicon. We would need a much larger sample if we wished to be certain that the earliest empirical location of a particular lexical item represented its hierarchical order of abstraction.

Beginning with the single principles list, we began the process of scoring the lexical items. First, a group of four researchers examined each lexical item in the single principles list. We asked whether any concepts commonly expressed with the item, or relationships described with the item, could be articulated at a lower secondary hierarchical order of abstraction than that associated with the complexity order at which they were found. For example, the item *social system* occurred for the first time at the single principles order. Given that this concept can be employed to describe a single system, and single systems are third-order abstractions, we concluded that social system could occur before the single-principles order. Similarly, the item *interrelated* occurred for the first time at the single principles order. Although this word can be employed to point to a relationship between systems that reflects reasoning at the single principles order, it can also be employed to describe a relationship between elements (second-order abstractions) of a single system (third-order abstraction) at the abstract systems order. All lexical items like social system and interrelated that we thought could be available earlier than the single principles order were moved to the abstract systems list. Lexical items that we considered unlikely to be employed at an earlier complexity order remained in the single principles list. We then repeated the process with the abstract systems, abstract mappings, single abstractions, representational systems, representational mappings, single representations, and sensorimotor systems lists, respectively. The entire process was repeated three times, to refine our selections.

We then employed the thesaurus in *Webster's New Universal Unabridged Dictionary* (1983) to identify synonyms and antonyms of words already in each of the lists. In this way, we expanded the generality of the lists by incorporating a variety of lexical items for many concepts. Following this process, the lists were once again subjected to the scoring process. The final result was a set of eight abstraction indexes, one for each complexity order from sensorimotor systems to single principles. The indexes were composed of 788; 1,771; 1,213; 651; 2,539; 8,140; 13,754; and 5,107 lexical items, respectively. The numbers of words in these lists suggest that, overall, larger vocabularies are associated with higher complexity orders. However, this trend may also, to some extent, reflect sample distribution. The lower complexity orders were less populated than the higher orders, which may have led to the construction of less complete, and therefore shorter, lists for those orders. Future work will address this issue. Appendix C provides some examples of lexical items from the eight abstraction indexes.

Finally, we calculated the density in vocabulary of lexical items from each of the abstraction indexes, for both training cases and test cases. In other words, we counted the number of items from a given index that were found in a particular performance, divided this by the total number of lexical items in the performance, and multiplied the result by 100. The performance of hypothetical case 01, for example, has a vocabulary of 250 lexical items. Thirty of these are on the abstract mappings list. The density of abstract mappings items in this performance is therefore 12%. The final result of this set of calculations is a set of eight densities for each case, plus the mean length of the vocabulary (mean word length) for each case.

RESULTS

The use of mean word length as a scoring criterion is justified by the strong correlation between mean word length and complexity order (r = .85, n = 1,014). Table 8 shows the relationship between mean word length and complexity order for the entire sample. (There is no difference between the correlations between mean word length and complexity order for the training or test samples and the sample as a whole.) As can be seen from Table 8, the relationship between mean word length and complexity order is nonlinear. In fact, this relationship is best described by a quadratic function. Whereas the r^2 for the linear relationship is .73, the r^2 for the quadratic relationship is .82. This means that mean word length increases somewhat more sharply at the higher complexity orders than at the lower complexity orders.

The densities of lexical items from each of the abstraction indexes were calculated for the vocabulary of each respondent, to examine their distributions by complexity order. Figures 1 through 8 show these distributions for the entire sample, including both test and training cases. (Either the training sample or the test sample could have been used in place of the entire sample; The basic relationships are virtually identical.) These figures suggest that lexical items from abstraction indexes representing a complexity order higher than that of a given performance are unlikely to appear in that performance at all. Moreover, they suggest that the density of lexical items from a given index increases for several complexity orders after items from the index first appear, stabilizes for a time, and—in the case of sensorimotor systems and single representations—finally declines. Overall, these patterns suggest that a performance at a given complexity order should be associated with a particular density pattern. For example, a likely density profile for a

TABLE 8
Mean Word Length by Complexity Order

Abbreviation	*Mean Word Length*	*Standard Error of Mean*	*SD*
SS	4.3999	.1077	.3232
SR	4.3609	.0376	.3005
RM	4.4945	.0253	.2645
RS	4.5084	.0230	.1838
SA	4.7793	.0216	.2497
AM	5.2548	.0214	.3275
AS	5.8198	.0223	.3704
SP	6.4152	.0363	.4040

Note. SS = sensorimotor systems; SR = single representations; RM = representational mappings; RS = representational systems; SA = single abstractions; AM = abstract mappings; AS = abstract systems; SP = single principles.

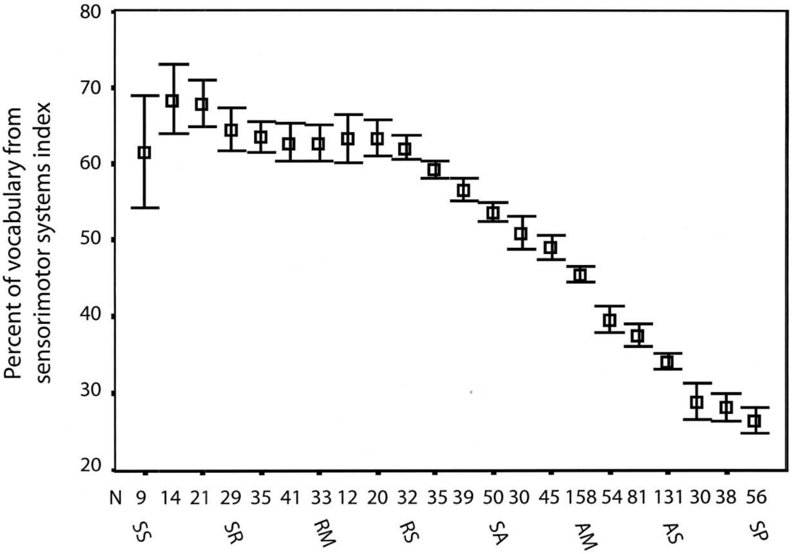

FIGURE 1 Mean percentage vocabulary from the sensorimotor systems index by complexity order.

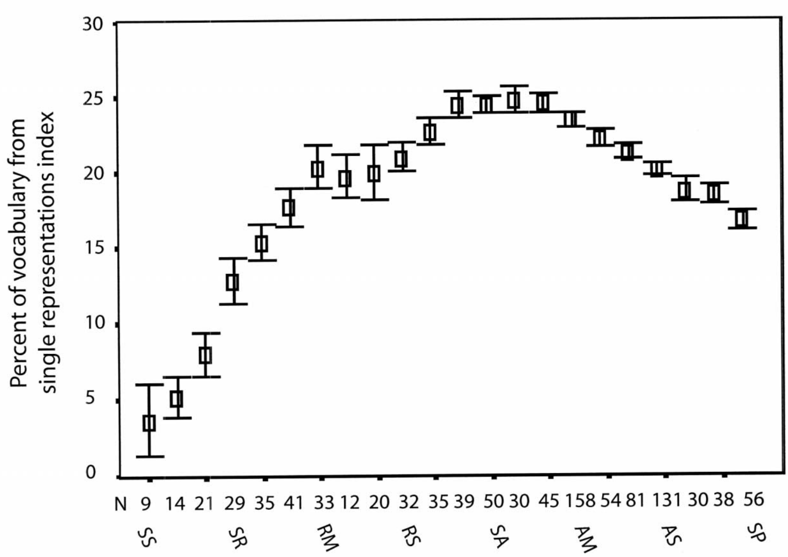

FIGURE 2 Mean percentage vocabulary from the single representations index by complexity order.

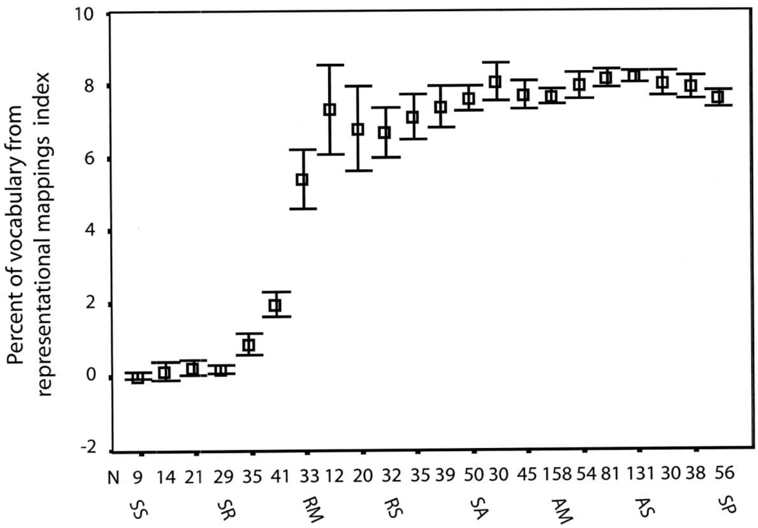

FIGURE 3 Mean percentage vocabulary from the representational mappings index by complexity order.

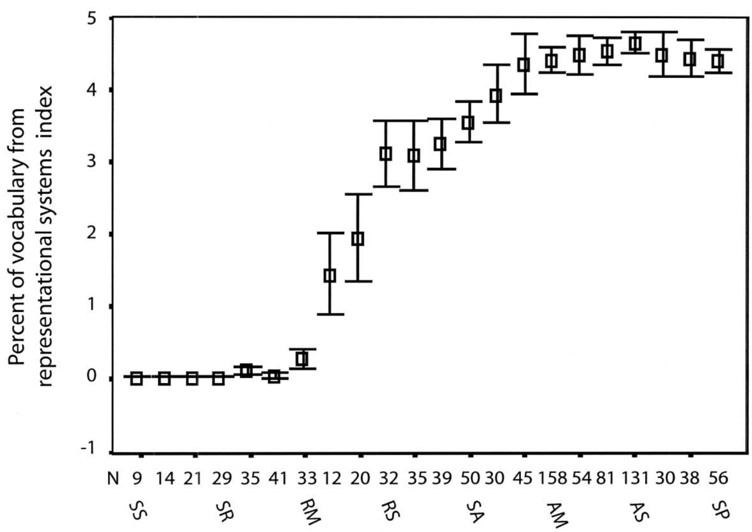

FIGURE 4 Mean percentage vocabulary from the representational systems index by complexity order.

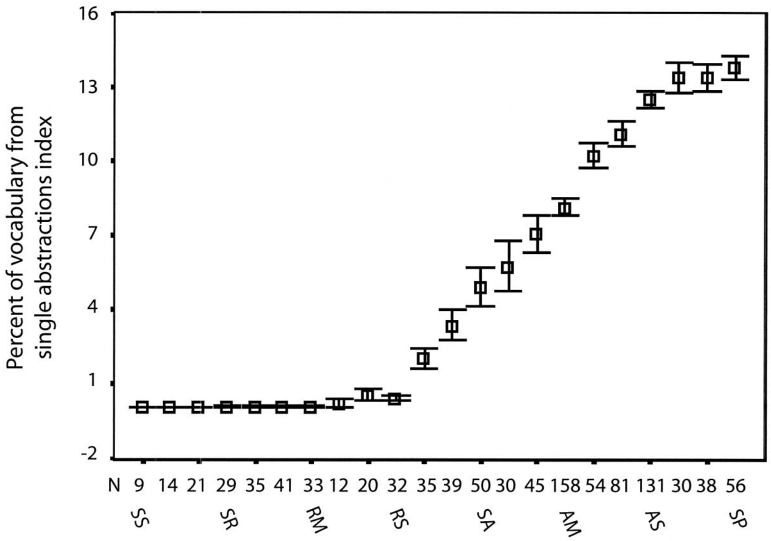

FIGURE 5 Mean percentage vocabulary from the single abstractions index by complexity order.

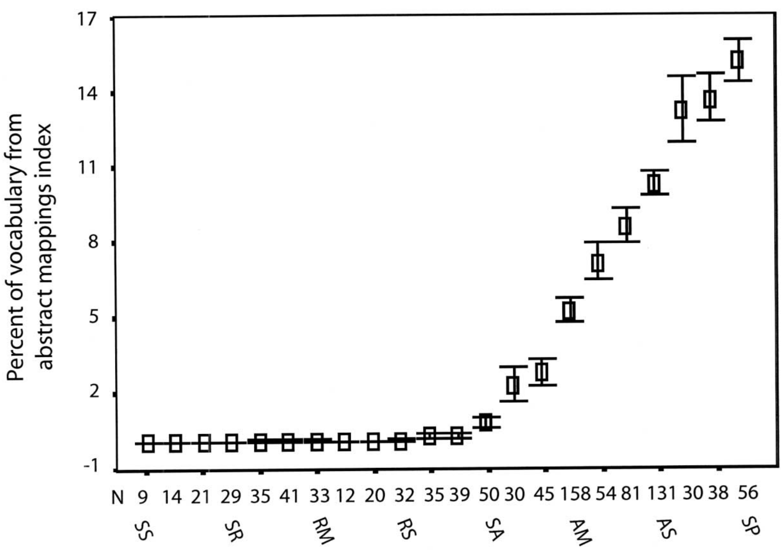

FIGURE 6 Mean percentage vocabulary from the abstract mappings index by complexity order.

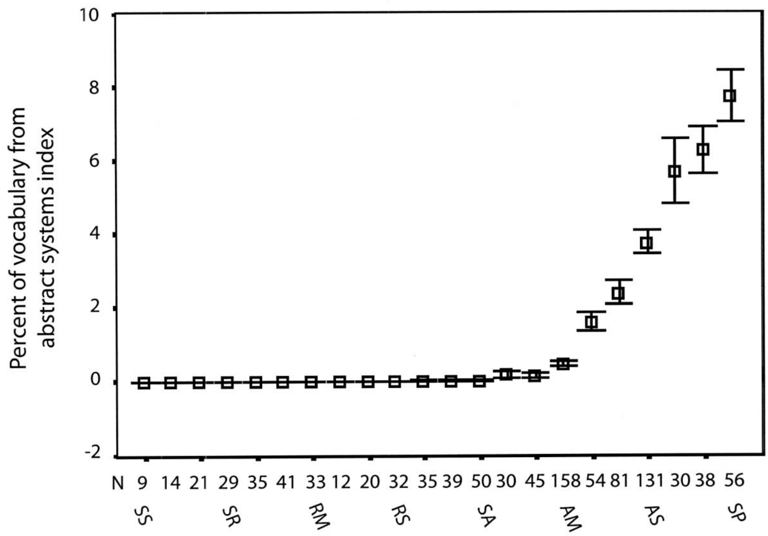

FIGURE 7 Mean percentage vocabulary from the abstract systems index by complexity order.

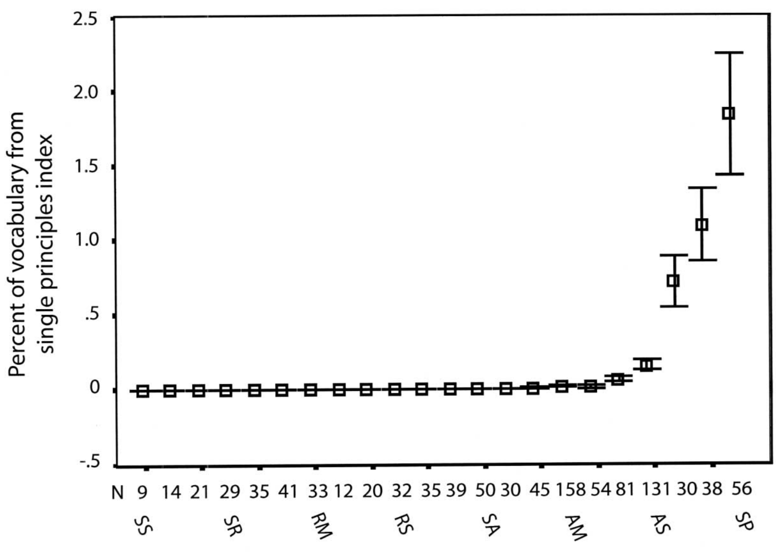

FIGURE 8 Mean percentage vocabulary from the single principles index by complexity order.

performance at the abstract mappings order would be 40%, 23%, 8%, 4.5%, 8%, 6%, 0%, and 0% for the sensorimotor systems, single representations, representational mappings, representational systems, single abstractions, abstract mappings, abstract systems, and single principles indexes, respectively.

To test the systematicity of these patterns, a discriminant analysis was run in Sxxxxxx Pxxxxxx Sxxxxxx Sxxxxxx on the 507 training cases only. Mean word length and the eight abstraction index densities were entered as variables, and 22 complexity order levels were entered as groups. Priors were set to SIZE (compute from group sizes) rather than EQUAL (equal group sizes), because earlier research suggested that the distribution of complexity order transitional phases is unequal in the population, with more individuals in consolidated phases than in transitional phases, and this is the distribution represented in our sample (Dawson, 1998; Dawson, Commons, & Wilson, 2003). Moreover, setting priors to SIZE produces predictions that more closely match human ratings. Predicted complexity order values were calculated, and the classification rule from the discriminant analysis was saved. The Pearson correlation coefficient between the predicted values and human-awarded complexity orders was .99, similar to the correlation between scores awarded by different human raters. Kendall's tau was .93, indicating that the predicted scores reliably reproduced the human ratings. Overall, discriminant function predictions agreed with human ratings within one-third of a complexity order 87% of the time. This agreement rate is solidly within the range of human interrater agreement.

Testing the Classification Rule

At this point, we tested the saved classification rule on the 507 test cases by importing the saved discriminant matrix from the initial analysis. In this case, the correlation between the predicted values and human-awarded complexity orders was .98, similar to the correlation between scores awarded by different human raters. Kendall's tau was .90, indicating that the classification rule from the earlier discriminant analysis reliably reproduced the human ratings in the test condition. Overall, the discriminant function predictions agreed with human ratings within one-third of a complexity order 81% of the time. This agreement rate is solidly within the range of human interrater agreement. When the two lowest complexity orders (sensorimotor systems and single representations), for which there were few data for estimation, were eliminated from the analysis, the discriminant function provided accurate predictions within one-third of a complexity order 83% of the time. These agreement rates meet or exceed human interrater reliabilities and agreement rates generally reported in the cognitive-developmental literature (Armon, 1984; Colby & Kohlberg, 1987a; Kitchener & King, 1990).

Table 9 provides a more detailed breakdown of the relationship between complexity order and the values predicted by the discriminant function. The rates of agreement (hit rates) within one-third of a complexity order are highlighted in bold

TABLE 9
Difference Between Hierarchical Complexity Scores and Lexical Abstraction Assessment System Scores by Complexity Order for 501 Test Cases

	SR		*RM*		*RS*		*SA*		*AM*		*AS*		*SP*		*Total*	
	Count	*Col %*	*Count*	*Col %*	*Count*	*Col %*	*Count*	*Col %*	*Count*	*Col %*	*Count*	*Col %*	*Count*	*Col %*	*Count*	*Col %*
0 OHC	7	22.6%	31	56.4%	11	34.4%	32	47.8%	66	55.9%	52	38.0%	23	37.7%	222	43.8%
1/3 OHC	11	35.5%	19	34.5%	15	46.9%	27	40.3%	33	28.0%	58	42.3%	22	36.1%	185	37.1%
0-1/3 OHC		**58.1%**		**90.9%**		**81.3%**		**88.1%**		**83.9%**		**80.3%**		**73.8%**		**80.9%**
2/3 OHC	11	35.5%	3	5.5%	4	12.5%	8	11.9%	11	9.3%	23	16.8%	10	16.4%	70	13.8%
1 OHC	2	6.5%			1	3.1%			8	6.8%	4	2.9%	6	9.8%	21	4.7%
1 1/3 OHC			1	1.8%	1	3.1%									2	0.4%
1 2/3 OHC			1	1.8%											1	0.2%
Total	31	100.0%	55	100.0%	32	100.0%	67	100.0%	118	100.0%	137	100.0%	61	100.0%	501	100.0%

Note. SS = sensorimotor systems; SR = single representations; RM = representational mappings; RS = representational systems; SA = single abstractions; AM = abstract mappings; AS = abstract systems; SP = single principles; Col % = column percent; OHC = complexity order.

type. From the single representations complexity order to the abstract systems complexity order, these hit rates were greater than 80%. The lower rate of agreement at the single representations complexity order can be explained by the lack of data for estimation at that level. The reason for the lower rate of agreement at the single principles complexity order is less clear.

DISCUSSION

We have shown that the construct of hierarchical order of abstraction can provide the basis for accurate and reliable computerized developmental assessments. The classification rule, generated from a discriminant analysis of a set of training cases, reproduced human ratings on the test cases with a high degree of accuracy—within the range of human interrater agreement. These results suggest that, for the first time, large-scale, objective developmental assessments of text performances are feasible. This is a major advance for developmental psychology. Moreover, the construct of hierarchical order of abstraction suggests a new, theory-based approach to the study of lexical acquisition.

As far as we know, this is the first time that principles from a strong developmental theory have been employed to inform the construction of a reliable and accurate computerized scoring system. (A fully automated system is currently being tested.) Other approaches to the assessment of texts rest largely on empirical evidence without clear grounding in a psychological theory of development. One such approach is based on quantifiable differences between texts assigned to different levels by human raters employing multidimensional scoring rubrics (Page, 1994). Another is based on word frequency and sentence length (Stenner, 1997). A third is based on a computationally sophisticated assessment of empirical relationships between word pairs (Landauer & Dumais, 1997). We argue that, as sophisticated as these scoring systems may be, they do not provide satisfying accounts of the developmental processes that underlie the sequences they specify. Based on our knowledge of these systems, we strongly suspect that hierarchical complexity is the primary latent dimension underlying all of them. We are eager to test this hypothesis.

Although the principle on which the LAAS is based is general and should apply to a wide range of texts, broader application of the LAAS will require further research. The abstraction indexes will, no doubt, require further adjustments. They still do an imperfect job of accurately reflecting hierarchical order of abstraction. There is some indication, for example, that the lists can be improved by eliminating some types of lexical items. We have already eliminated concrete nouns, and are considering the possibility of excluding some types of words that can be employed as nouns, because their appearance is less predictive of complexity order than other word forms. For example, the word *responsibility* can be employed at the representational mappings order to mean "job," whereas its root word, *responsible*, refers to notions of causality, personal conduct, or obligation that are not available before the single abstractions order. Ultimately, the lists not only should accurately reflect hierarchical order of abstraction, but also should include only those lexical items that are of value in assessing the hierarchical complexity of texts.

In addition to conducting further work on the abstraction indexes, we are testing the classification rule produced from the discriminant analysis on additional samples of interviews. We have employed this classification rule to score (a) a set of 146 adolescent interviews about peer relations, (b) a set of 40 essays on the nature of truth and knowledge written by returning adult college freshmen, (c) a set of 101 literary essays written by traditional-aged college freshmen, (d) a small selection of clinical psychology students' case evaluations, (e) 149 epistemology interviews (Dawson, in press-a), and (f) a sample of 60 interviews investigating the development of conceptions of *flow* (Jackson, 2003). Results suggest that the classifications work well across all of these data sources, approximating human ratings well (within one-third of a complexity order).

Another objective is to demonstrate the utility of the LAAS as an assessment system. Several applications suggest themselves. First, the LAAS may be useful as an indicator of readiness for some kinds of learning. For example, it is well documented that vocabulary and developmental achievement are predictors of the ease with which children learn to read (Cannella, 1980; Shinn-Strieker, House, & Klink, 1989; Spencer, 1986; Sterner & McCallum, 1988). It is quite possible that an evaluation of the hierarchical complexity of children's verbal reasoning will prove to be a good predictor of reading readiness. If this is the case, the LAAS could possibly be employed as part of an evaluation of reading readiness. It could also be used to estimate the cognitive demands of the texts employed in reading programs. Employing the same assessments to evaluate students and texts should simplify the task of selecting appropriate texts for individual readers.

A second potential application of the LAAS is in program assessment. Many primary, secondary, and college programs have set developmental objectives for their students. Developmental measures like Kohlberg's (Colby & Kohlberg, 1987b) Standard Issue Scoring System, and Kitchener and King's (1990) Reflective Judgment instrument have frequently been employed to assess developmental interventions (Higgins, 1991; Sakalys, 1984) but are impractical for large-scale assessments. The LAAS can reasonably be employed to assess the ongoing intellectual development of large numbers of students, providing evidence of the efficacy of program objectives. We are currently piloting this type of application for the LAAS.

A third potential application of the LAAS is in the college entrance process. Given the high correspondence between educational attainment and the hierarchical complexity of reasoning (Dawson, Commons, & Wilson, 2003), as well as the documented relationship between academic success and cognitive development (Campbell & Ramey, 1990; Endler & Bond, 2001; Jones, 1995; Shayer, 1996; Wood, Sher, & Bartholow, 2002), we anticipate finding an association between academic success and performance on the LAAS. The written work of high school seniors can be evaluated for its hierarchical complexity, and these evaluations can be employed as additional information about students' academic qualifications. To explore this application of the LAAS, we are presently examining the relationship between LAAS classifications of the essays of a group of high school students and their later performance at university. There are numerous other potential applications for the LAAS.

By examining interview data from the perspectives of both hierarchical complexity and linguistic content as suggested in Overton's (1998) framework for a relational methodology, we discovered an important systematic relation between lexical acquisition and cognitive development. This has provided a new insight into their interrelationship, as well as a novel instrument for the measurement of developmental change. In several ways, the LAAS is a first. It is the first developmentally based computerized assessment system for texts; it is the first computerized developmental assessment system that measures development along a single trait from early childhood through adulthood; and it represents the first application of the hierarchical order of abstraction construct to real-world assessment. The potential applications of a practical, generalized developmental assessment system are numerous. The LAAS promises not only to make it possible to conduct large-scale developmental assessments for practical applications, but also to allow researchers to address fundamental questions about developmental processes that were heretofore impractical to consider.

ACKNOWLEDGMENTS

Special thanks to the Murray Research Center, Brian MacWhinney and the CHILDES project, Cheryl Armon, Marvin Berkowitz, Peggy Drexler, Jim Mensing, Mike Pavelich, and Larry Walker for the use of their interview data. This project is partially funded through a grant from the Spencer Foundation (XXXXXXXXXX). The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

REFERENCES

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

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

Bates, E., Bretherton, I., Snyder, L., Beeghly, M., Shore, C., McNew, S., et al. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* New York: Cambridge University Press.

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: Lawrence Erlbaum Associates, Inc.

Bloom, L. (1970). *Language and development: Form and function in emerging grammars.* Cambridge, MA: MIT Press.

Bloom, L., & Lightblown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology, 6,* 380–420.

Bloom, L., Lightblown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40*(2, Serial No. 160).

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Campbell, F. A., & Ramey, C. T. (1990). The relationship between Piagetian cognitive development, mental test performance, and academic achievement in high-risk students with and without early educational experience. *Intelligence, 14,* 293–308.

Cannella, G. S. (1980). Beginning reading: Cognitive developmental research. *Journal of Instructional Psychology, 7,* 137–141.

Carlson-Luden, V. (1979). *Causal understanding in the 10-month-old.* Unpublished doctoral dissertation, University of Colorado at Boulder.

Case, R., Griffin, S., McKeough, A., & Okamoto, Y. (1992). Parallels in the development of children's social, numerical, and spatial thought. In R. Case (Ed.), *The mind staircase* (pp. 269–284). 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* (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.

Colby, A., Kohlberg, L., Gibbs, J., & Lieberman, M. (1983). A longitudinal study of moral judgment. *Monographs of the Society for Research in Child Development, 48*(1–2, Serial No. 200).

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., Straughn, J., Meaney, M., Johnstone, J., Weaver, J. H., Lichtenbaum, E., et al. (1995, November). *The general stage scoring system: How to score anything.* Paper presented at the annual meeting of the Association for Moral Education, New York.

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.

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.

Dawson, T. L. (2001, June). *Meaning language and stage: A new look at moral thought.* Paper presented at the annual meeting of the Jean Piaget Society, Berkeley, CA.

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

Dawson, T. L. (2003). *The Hierarchical Complexity Scoring System*. Retrieved January 31, 2003, from http://gseacademic.harvard.edu/~hcs/

Dawson, T. L. (in press-a). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development*.

Dawson, T. L. (in press-b). 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. (2003). *The shape of development.* Manuscript submitted for publication.

Dawson, T. L., & Gabrielian, S. (2003). Developing conceptions of authority and contract across the life-span: Two perspectives. *Developmental Review, 23,* 162–218.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development, 18,* 61–78.

Demetras, M. (1989a, XXMONTH). *Changes in parents conversational responses: A function of grammatical development.* Paper presented at the annual convention of the American Speech and Hearing Association, St. Louis, MO.

Demetras, M. (1989b). *Working parents conversational responses to their two-year-old sons.* Unpublished manuscript, University of Arizona, Tucson.

Drexler, P. (1998). *Moral reasoning in sons of lesbian and heterosexual parent families: The oedipal period of development.* Unpublished doctoral dissertation, California School of Professional Psychology, Berkeley/Alameda.

Endler, L. C., & Bond, T. G. (2001). Cognitive development in a secondary science setting. *Research in Science Education, 30,* 403–416.

Fischer, K. W. (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: John Wiley & Sons.

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

Hall, W. S., Nagy, W. E., & Linn, R. (1984*). Spoken words: Effects of situation and social group on oral word usage and frequency.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hall, W. S., Nagy, W. E., & Nottenburg, G. (1981). *Situational variation in the use of internal state words.* Urbana–Champaign: University of Illinois, Center for the Study of Reading.

Hall, W. S., & Tirre, W. C. (1979). *The communicative environment of young children: Social class, ethnic and situational differences.* Urbana–Champaign: University of Illinois, Center for the Study of Reading.

Henry, A. (1995). *Belfast English and standard English: Dialect variation and parameter setting.* New York: Oxford University Press.

Higgins, A. (1991). The Just Community approach to moral education: Evolution of the idea and recent findings. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development, Vol. 1* (pp. 111–141). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Higginson, R. P. (1985). *Fixing: Assimilation in language acquisition.* Unpublished doctoral dissertation, Washington State University.

Jackson, B. (2003). *In search of peak experiences through life. Seeking the arenas and strategies for replicating the flow experience: A developmental perspective.* Unpublished doctoral dissertation, Fielding Graduate Institute, Santa Barbara, CA.

Jones, D. E. (1995). *Predicting college grade point average: The contribution of four developmental domains.* Unpublished manuscript, University of Kansas, Lawrence.

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., 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.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mensing, J. (2002). *Collectivism, individualism, and interpersonal responsibilities in families: Differences and similarities in social reasoning between individuals in poor urban families in Colombia and the United States.* Unpublished doctoral dissertation, University of California at Berkeley.

Miller, R. L., Olds, B. M., & Pavelich, M. J. (2001). *A computer-based expert system to measure intellectual development in college students* (Report to the Fund for the Improvement of Post-Secondary Education [FIPSE], Grant No. P116B70050). Golden: Colorado School of Mines, Chemical Engineering Department.

Noelting, G. N., & Rousseau, J. P. (Eds.). (in press). *Genetic epistemology and the construction of knowledge.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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* (5th ed., pp. 107–188). New York: John Wiley & Sons.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62,* 127–142.

Pavelich, M. J., Miller, R. L., & Olds, B. M. (2002, June). *Software for measuring the intellectual development of students: Advantages and limitations.* Paper presented at the annual conference of the American Society for Engineering Education, Washington, DC.

Perry, W. G. (1970). *Forms of intellectual and ethical development in the college years*. New York: Holt, Rinehart, & Winston.

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

Sakalys, J. A. (1984). Effects of an undergraduate research course on cognitive development. *Nursing Research, 33,* 290–295.

Shayer, M. (1996). *The long-term effects of Cognitive Acceleration on pupils' school achievement.* London, England: The Centre for the Advancement of Thinking.

Shinn-Strieker, T., House, G., & Klink, B. (1989). The roles of cognitive processing and language development in emergent literacy. *RASE: Remedial & Special Education, 10*(5), 43–50.

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

Spencer, B. R. (1986). The concept of word in young children: Tacit and explicit awareness of children at different operational levels. *National Reading Conference Yearbook, 35,* 271–280.

Stenner, A. J. (1997). *The objective measurement of reading comprehension* (Technical Report). Durham, NC: Duke University, Institute of Statistics and Decision Sciences.

Sterner, A. G., & McCallum, R. S. (1988). Relationship of the Gesell Developmental Exam and the Bracken Basic Concept Scale to academic achievement. *Journal of School Psychology, 26,* 297–300.

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

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

Walker, L. J., Gustafson, P., & Hennig, K. H. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology, 37,* 187–197.

Webster's New Universal Unabridged Dictionary (2nd ed., J. L. NcKechnie et al., eds.). (1983). New York: Simon & Schuster.

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

Wilson, J., & Henry, A. (1998). Parameter setting within a socially realistic linguistics. *Language in Society, 27,* 1–21.

Wood, P., Sher, K., & Bartholow, B. (2002). Alcohol use disorders and cognitive abilities in young adulthood: A prospective study. *Clinical and Consulting Psychology, 70,* 897–907.

APPENDIX A
The Sensorimotor Systems to Single Principles Complexity Orders

The complexity order of a verbal performance is reflected in two interrelated aspects of its structure. First, there is a hierarchical complexity of performance associated with the hierarchical order of abstraction of concepts at that complexity order. Second, there is a hierarchical complexity of performance associated with the most complex logical structure exhibited at the complexity order, which is observable in the organization of its conceptual content. Here, the conceptual and logical organization of complexity orders 5 through 12 are described (the complexity orders identified in our sample). In these descriptions we primarily employ examples from responses to the Joe dilemma (from Form A of Kohlberg's Standard Issue Scoring System, Colby & Kohlberg, 1987b):

> Joe is a 14-year-old boy who wanted to go to camp very much. His father promised him he could go if he saved the money for himself. So Joe worked hard at his paper route and saved up the $100 it cost to go to camp and a little more besides. But just before camp was going to start, his father changed his mind. Some of his friends decided to go on a special fishing trip, and Joe's father was short of the money it would cost. So he told Joe to give him the money he had saved from the paper route. Joe didn't want to give up going to camp, so he thinks he might not give his father the money.

At the sensorimotor systems complexity order, the new concepts are referred to as symbolic systems. These coordinate second-order symbolic sets (the concepts of the previous complexity order). In responses to the Joe dilemma, for example, the concept of *swimming* coordinates the idea of being in the water with making certain kinds of movements, and the concept of *painting* coordinates the idea of making marks on paper with the particular tools and products involved in painting (vs. drawing, for example). The most complex logical structure of this complexity order is multivariate, identifying multiple aspects of symbolic sets, as in "Mommy, I painting. I put red paint on wall with brush," which describes the painting system.

At the single representations complexity order, the new concepts are referred to as first-order representational sets. These coordinate sensorimotor systems. In responses to the Joe dilemma, for example, the concept of *camping* coordinates activities like swimming, sleeping in a tent, and painting, and the concept of a *paper route* coordinates activities like riding a bike, delivering papers, and receiving money. The most complex logical structure of this complexity order is definitional, identifying one aspect of a single representation, as in "Camping is fun," in which fun is an aspect of camp.

At the representational mappings complexity 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 very popular representational mappings order concept of *having favorites*, for example, can be employed to rank camping and fishing. "Camping is my favorite, and fishing is my next favorite." Concepts like *being mean, keeping a promise, changing one's mind,* and *sharing* also become common at this complexity order. "[Joe's father] is just being mean, he is taking the money away from his kids." The most complex logical structure of this complexity order is linear, coordinating one aspect of two or more representations, as in "If you do not do what your father tells you to do, he will get really mad at you," in which doing what your father says and not doing what your father says are coordinated by his anticipated reaction.

At the representational systems complexity order, the new concepts are third-order representational sets. These coordinate elements of representational systems. For example, the concept of *trust,* articulated for the first time at the representational systems order, can be used to describe the system of interactions between Joe and his father. "Joe trusted [his Dad] that he could go to the camp if he saved enough money, and then his father just breaks it and the promise is very important." Concepts like *to turn against, to blame, to believe,* and *being fair* are also infrequently observed before this complexity order. "[If you break a promise] they will not like you anymore, and your friends will turn against you." The most complex logical structure of this complexity order is multivariate, coordinating multiple aspects of two or more representations, as in, "If Joe's Dad says Joe can go to camp, then he says he can't go to camp, that's not fair because Joe worked hard and then his Dad changed his mind," in which two conflicting representations of Dad's authority are evaluated in terms of his changed mind and Joe's hard work.

At the single abstractions complexity order, the new concepts are referred to as first-order abstractions. These coordinate representational systems. For example, the concept of *trustworthiness*, articulated for the first time at this complexity order, defines those qualities that make a person trustworthy rather than describing a particular situation in which trust is felt or not felt. It is composed of qualities that produce trust, such as telling the truth, keeping secrets, and keeping promises. "It's always nice … to be trustworthy. Because then, if [someone has] a secret, they can come and talk to you." Concepts like *kindness, keeping your word, respect,* and *guilt* are also rare before the single abstractions order. “If you don’t do something you promise, you’ll feel really guilty.” The most complex logical structure of this complexity order identifies one aspect of a single abstraction, as in “Making a promise is giving your word,” in which giving one’s word is an aspect of a promise.

At the abstract mappings complexity order, the new concepts are referred to as second-order abstractions. These coordinate or modify single abstractions. For example, the abstract mappings order concept *basis* can be employed to coordinate the elements essential to a good relationship. “To me, [trust and respect are] the basis of a relationship, and without them you really don’t have one.” Concepts like *coming to an agreement, making a commitment, building trust,* and *compromise* are also rare before the abstract mappings order. “I think [Joe and his father] could come to an agreement or compromise that they are both comfortable with.” The most complex logical structure of this complexity order coordinates one aspect of two or more abstractions, as in “Joe has a right to go to camp because his father said he could go if he saved up the money, and Joe lived up to his commitment.” Here, Joe’s fulfillment of his father’s conditions determines whether Joe has a right or does not have a right to go to camp.

At the abstract systems complexity order, the new concepts are referred to as third-order abstractions. These coordinate elements of abstract systems. For example, the concept of *personal integrity*, which is rare before the abstract systems order, refers to the coordination of and adherence to notions of fairness, trustworthiness, honesty, preservation of the golden rule, etc. in one’s actions. “[You should keep your word] for your own integrity. For your own self-worth, really. Just to always be the kind of person that you would want to be dealing with.” Concepts like *verbal contract, moral commitment, functional, development, social structure,* and *foundation* are also uncommon before the abstract systems order. “A promise is the verbal contract, the moral commitment that the father made to his son. It is the only way for the child to … develop his moral thinking—from watching his parent’s moral attitude.” The most complex logical structure of this complexity order coordinates multiple aspects of two or more abstractions. “Following through with his commitment and actually experiencing camp combine to promote Joe’s growth and development, not just physically, but psychologically, emotionally, and spiritually.” Here multiple facets of Joe’s personal development are promoted when he both keeps his commitment and accomplishes his goal.

At the single principles/axioms complexity order, the new concepts are referred to as first-order principles. These coordinate abstract systems. The notion of the *social contract* (as it is constructed at this order), for example, results from the coordination of human interests (where individual human beings are treated as systems). “Everybody wants to be treated equally and have a sense of fair play. Because this is so, we have an obligation to one another to enter into a social contract that optimizes equality and fairness.” Concepts like *autonomy, fair play, heteronomy, higher order principle,* and *philosophical principle* are rare before the single principles order.

"The only time we're justified in breaking the social contract is when a higher principle, such as the right to life, intervenes." The most complex logical structure of this complexity order identifies one aspect of a principle or axiom coordinating systems, as in "Contracts are articulations of a unique human quality, mutual trust, which coordinates human relations." Here, contracts are seen as the instantiation of a broader principle coordinating human interactions.

APPENDIX B
Scoring With the Hierarchical Complexity Scoring System

Scoring with the Hierarchical Complexity Scoring System involves identifying the hierarchical order of abstraction and logical structure of a text. The simplest way to demonstrate how these two constructs work together is by representing texts as concept maps, in which the elements of arguments can be seen in relation to one another. Here, we provide maps of four different conceptions of *contract/promise,* representing single abstractions, abstract mappings, abstract systems, and single principles constructions.

Single Abstractions

Figure B1 portrays a visual representation, in the form of a concept map, of a 54-year-old[3] respondent's argument about why promises should be kept. The respondent argues that a person should keep a promise because keeping promises is "the right thing to do." When probed, the respondent comes up with three separate (uncoordinated) reasons for keeping promises: because people expect promises to be kept, because "people will trust you" if you keep a promise, and because "you might feel guilty if you break a promise." All three of these reasons for keeping promises are considered to be first-order abstractions, because they extract general abstract notions by coordinating concepts that appear for the first time at the representational systems order (Dawson & Gabrielian, 2003). Keeping promises will create trust, in general; people, in general, have expectations when promises are made; and breaking promises can produce negative emotional consequences, in general, for the promise breaker. It is important to keep in mind that the particular concepts expressed by a respondent are important only to the extent that they embody a particular hierarchical order of abstraction. A rater must "look through" the meaning of a particular conceptual element to abstract its hierarchical order of abstraction.

[3]Single abstractions dominate by age 10 or 11 in most populations that have been sampled by developmental researchers. However, a small percentage of adults do not move beyond this complexity order in their moral reasoning.

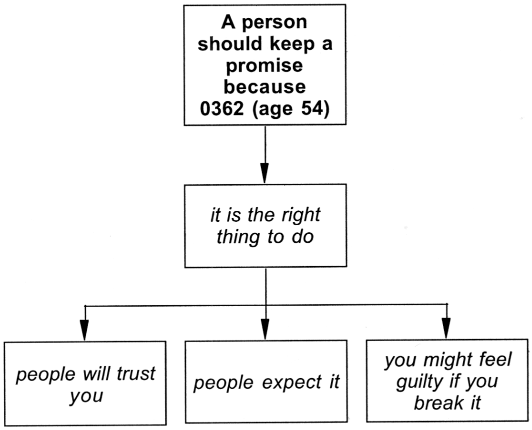

FIGURE B1 Single abstractions conception of *promise*.

Abstract Mappings

Figure B2 provides a map of the performance of a 58-year-old male, who provides three reasons for keeping promises. There are two mappings in this performance. The first is the assertion that "broken promises can harm relationships, because they cause pain and reduce trust." This mapping coordinates two abstract conse-

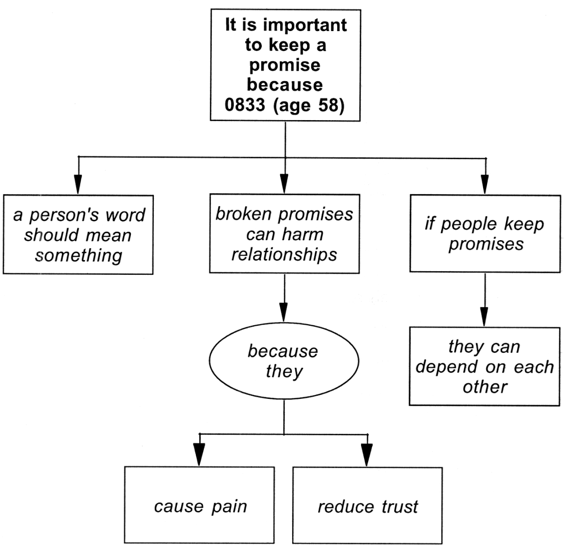

FIGURE B2 Abstract mappings conception of *promise*.

quences of promise-breaking into the general notion that broken promises do harm to relationships. The second is the assertion that keeping promises makes it possible for people to "depend on one another." This mapping coordinates the perspectives of at least two individuals to form the notion that keeping promises produces mutual benefits. Note how this idea builds on the single abstractions notion that people will trust you if you keep promises.

Abstract Systems

Figure B3 provides a map of the performance of a 51-year-old female. The respondent describes a system in which promise-keeping is both obligatory and sometimes impossible, "due to unforeseen circumstances." The reason for keeping promises is that one must stand by one's commitments. Doing so not only preserves one's personal integrity, but also builds a sense of trust, "which keeps society functioning." The notion of standing by one's commitments, the idea that doing so preserves one's integrity, the argument that the sense of trust built through promise-keeping keeps society functioning, and the notion of unforeseen circumstances are all examples of second-order abstractions. Note how the notion that the trust built from promise-keeping keeps society functioning (even in the presence of

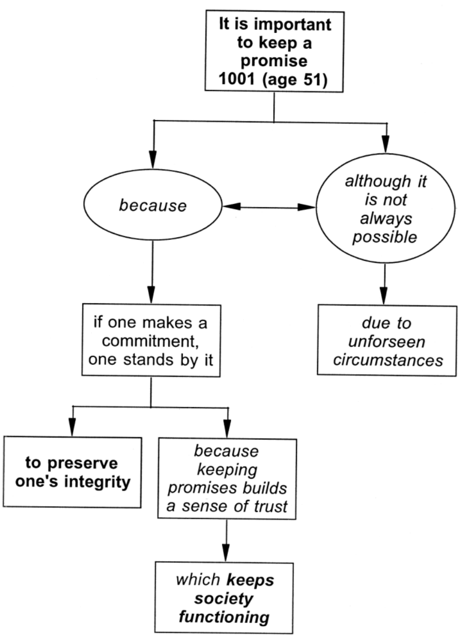

FIGURE B3 Abstract systems conception of *promise*.

the effects of unforeseen circumstances) builds on the abstract mappings idea that keeping promises makes it possible for people to depend on one another.

Single Principles

Figure B4 presents a map of the performance of a 57-year-old male. Here, "mutual trust" is employed as a single principle supporting an argument for keeping promises. The rationale for employing this principle is that "most social conventions" and "all moral principles" are based on trust. Both "all moral principles" and "most social conventions" are third-order abstractions. Note how this single principles argument builds on the abstract systems notion that trust keeps society functioning.

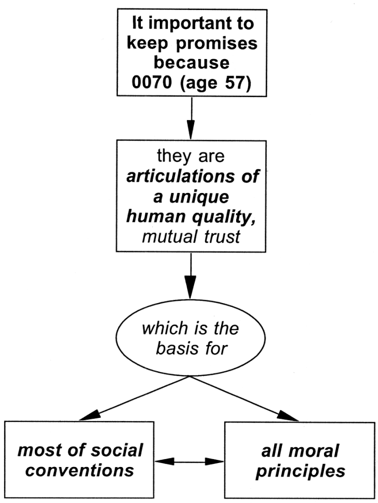

FIGURE B4 Single principles conception of *promise*.

APPENDIX C
Lexical Items by Complexity Order (From the Epistemology Lexicon)

SS	*SR*	*RM*	*RS*	*SA*	*AM*	*AS*	*SP*
know	guess	believe	certain	actual	accurate	actuality	counterfactual
think	learn	correct	doubt	belief	certainty	cognizant	determinants
wrong	question	lie	explain	cause	determine	fallacious	exactitude
right	teach	probably	fact	deceive	deceptive	presumably	materiality
		real	honest	honestly	plausible	plausibility	specious
		reason	imagine	imaginary	rational	rationality	spurious
		true	possible	knowledge	skeptical	skepticism	truisms
		understand	trust	memorize	justify	self-knowledge	

Note. SS = sensorimotor systems; SR = single representations; RM = representational mappings; RS = representational systems; SA = single abstractions; AM = abstract mappings; AS = abstract systems; SP = single principles.