EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY
2005, 2 (2), 163–195

Psychology Press
Taylor & Francis Group

The shape of development

Theo L. Dawson-Tunik
Hampshire College, Amherst, MA, USA

Michael Commons
Harvard Medical School, Cambridge, MA, USA

Mark Wilson
University of California, Berkeley, CA, USA

Kurt W. Fischer
Harvard Graduate School, Cambridge, MA, USA

This project examines the *shape* of conceptual development from early childhood through adulthood. To do so we model the attainment of developmental complexity levels in the moral reasoning of a large sample ($n = 747$) of 5- to 86-year-olds. Employing a novel application of the Rasch model to investigate patterns of performance in these data, we show that the acquisition of successive complexity levels proceeds in a pattern suggestive of a series of spurts and plateaus. We also show that there are six complexity levels represented in performance between the ages of 5 and 86; that patterns of performance are consistent with the specified sequence; that these findings apply to both childhood and adulthood levels; that sex is not an important predictor of complexity level once educational attainment has been taken into account; and that both age and educational attainment predict complexity level well during childhood, but educational attainment is a better predictor in late adolescence and adulthood.

Correspondence should be addressed to Theo L. Dawson-Tunik, Cognitive Science, Hampshire College, Amherst, MA 01002, USA. Email: tdawson@hampshire.edu

The research reported in this paper was made possible in part by a grant from the Spencer Foundation.

This work would not have been possible without the participation of Yiyu Xie, Hiro Yamada, Adam Kay, Lainie Kantor, Sonya Gabrielian, and Miles Becker, and the donation of interview data by the Murray Research Center, Larry Walker, Cheryl Armon, Marvin Berkowitz, Michael Commons, and Peggy Drexler. However, the data presented, the statements made, and the views expressed are solely the responsibility of the authors.

http://www.tandf.co.uk/journals/pp/17405629.html DOI: 10.1080/17405620544000011

INTRODUCTION

There has been much debate over the *shape* of cognitive development. Many models have been presented, ranging from those based on the notion that cognitive development is incremental or continuous (Bandura, 1977) to those that consider it to be discontinuous, involving transformations such as hierarchical integration (Case, 1987; Demetriou & Valanides, 1998; Fischer, 1980; Piaget, 1985) or the processes of nonlinear dynamics (Lewis, 2000; Smith & Thelen, 1993; van der Maas & Molenaar, 1995; van Geert, 1998). Interestingly, research results have supported both types of model, and as Fischer and his colleagues (Fischer & Bidell, 1998) have repeatedly demonstrated, development can appear continuous under some conditions and discontinuous under others.

This project investigates the acquisition of levels of hierarchical complexity (complexity levels) by modeling the structure of reasoning represented in a large, lifespan (ages 5–86), cross-sectional sample of moral judgment interviews. Because we required a large sample in order to successfully model developmental patterns across several complexity levels, we collected 747 interviews from a number of Kohlbergian studies of moral judgment development. Combining samples in this way not only produced a large sample, but also increased the extent to which the results could be generalized, because it diluted the effects that the composition of any one sample might exert on the outcome.

Though the data are moral judgment interviews and the results have important implications for moral theory, this paper is not about moral development per se. It is about conceptual development as it plays itself out in the moral domain. Developmental progress is assessed with a domain-general developmental assessment system, the *Lectical*^{TM} *Assessment System* (LAS), which primarily is based on Piaget's conceptualization of reflective abstraction (Piaget, 2000), the *General Stage Model,* also known as the *Model of Hierarchical Complexity* (Commons, Trudeau, Stein, Richards, & Krause, 1998) and *Skill Theory* (Fischer, 1980; Fischer & Bidell, 1998). To investigate patterns of performance in these data, we employ the Rasch model (1980), a well-established psychometric model that is particularly well-suited for examining patterns of performance in developmental data. We address four research questions, including: (1) how many complexity levels are represented in performances between the ages of 5 and 86; (2) are patterns of performance consistent with the specified sequence; (3) does the shape of development as revealed in the rating scale model support the theoretical position that development is discontinuous; and (4) do these findings apply to both childhood and adulthood stages? We also briefly examine relations between complexity level and sex, age, and educational attainment.

REFLECTIVE ABSTRACTION, HIERARCHICAL INTEGRATION, AND HIERARCHICAL COMPLEXITY

Most cognitive-developmental researchers agree that development in different knowledge domains does not necessarily proceed at the same rate (Fischer & Bidell, 1998; Lourenco & Machado, 1996). However, there is still considerable disagreement about whether development in different domains can be characterized in terms of a single, generalized process. Domain theorists argue that different processes apply in different knowledge domains (Kohlberg, 1969; Larivee, Normandeau, & Parent, 2000; Turiel, 1980). Others, though they acknowledge that unique structures and processes are associated with particular domains, also argue that a single general developmental process applies across domains (Case, Okamoto, Henderson, & McKeough, 1993; Fischer & Bidell, 1998). Piaget (2000) called this general process *reflective (or reflecting) abstraction*, through which the actions of one developmental level become the subject of the actions of the subsequent level. The product of reflective abstraction is *hierarchical integration*. In conceptual development, hierarchical integration is observable in the concepts constructed at a new level by coordinating (or integrating) the conceptual elements of the prior level. These new concepts are said to be more hierarchically complex than the concepts of the previous level, in that they integrate earlier knowledge into a new form of knowledge. For example, independent conceptions of *play* and *learning* constructed at one complexity level are integrated into a conception of *learning as play* at the next complexity level (Dawson-Tunik, 2004a). Though it builds on the original *play* and *learning* concepts, the new concept cannot be reduced to the original *play* and *learning* elements. Not only is there a new concept in the recognition that learning can be playful, but the individual meanings of the elements *learning* and *play* have changed, in that each now incorporates some of the meaning embedded in the new construction—the concept of *learning* now includes *play* as a component, and the concept of *play* includes learning as a component.

A number of researchers have described developmental sequences that elaborate the basic notion of hierarchical integration, including Fischer (1980), who has emphasized the development of skill hierarchies in particular contexts; Commons and his colleagues (Commons et al., 1998), who have described a task structure hierarchy; Pascual-Leone and Goodman (1979), who have focused on the growth of mental attention and memory capacity; Case (1991), who described the development of memory capacity and associated processing structures; and Demetriou & Valanides (1998), who have described hierarchical development in terms of processing functions.

Not only are there definitional correspondences among analogous levels described by Commons, Fischer, and Piaget, there is empirical evidence of correspondences between complexity levels, skill levels, and orders of hierarchical complexity and at least three domain-based systems, including Kitchener and King's (Dawson, 2002b; King, Kitchener, Wood, & Davison, 1989; Kitchener & King, 1990; Kitchener, Lynch, Fischer, & Wood, 1993) stages of reflective judgment, Armon's good life stages (Dawson, 2002a), Perry's stages of epistemological development (Dawson, 2004), and Kohlberg's moral stages (Commons et al., 1989; Dawson & Gabrielian, 2003; Dawson, Xie, & Wilson, 2003). These correspondences suggest that, as a community, this group of developmental researchers is moving toward a consensus regarding the detection and aspects of the definition of developmental stages. Table 1 shows the level names, typical ages of appearance, and relations among a number of developmental sequences, including the levels specified in the LAS and those of Piaget (Piaget & Inhelder, 1969), Fischer (Fischer & Bidell, 1998), Commons (Commons, Richards, with Ruf, Armstrong-Roche, & Bretzius, 1984), Armon (1984b), Kitchener and King (1990), and Kohlberg (Colby & Kohlberg, 1987a). The skill level names from Fischer's skill theory are also used to denote LAS complexity levels.[1]

THE LAS

Like Commons' Hierarchical Complexity Scoring System (Commons, Danaher, Miller, & Dawson, 2000), and Rose and Fischer's (1989) system for constructing task sequences, the LAS (Dawson-Tunik, 2004b) is designed to make it possible to assess the complexity level of a performance without reference to *particular* conceptual content. Rather than making the claim that a person's response occupies a complexity level because that person, for example, has elaborated a particular conception of justice, the LAS permits us to identify performances of a particular complexity level and then ask what the range of justice conceptions are at that complexity level. Thus, it avoids much of the circularity of domain-based developmental assessment systems, which define developmental levels in terms of particular conceptual content or content-laden domain-specific structures (Brainerd, 1993).

The LAS focuses on two aspects of texts that can be abstracted from particular conceptual content. The first is conceptual structure, embodied in

[1]Dr Fischer (personal communication, 23 September 2002) has agreed to the use of these labels.

TABLE 1
Correspondences among seven developmental sequences

Dawson-Tunik (2004)	*Piaget & Inhelder (1969)*	*Fischer & Bidell (1998)*	*Commons et al. (1998)*	*Colby & Kohlberg (1987)*[a]	*Armon (1984)*[b]	*Kitchener & King (1994)*[c]
Single Principles/ Axioms (SP) 25 years +		P1, A4, Single Principles 23–25 years	Meta-systematic	Stage 5 24 + years	Stage 5 30 + years	Stages 6 & 7 26 + years
Abstract Systems (AS) 18–20 years	Consolidated Formal Operations	Tier 4 A3, Abstract Systems 18–20 years	Systematic	Stage 4 22 years	Stage 4 22 years	Stage 5 21 years
Abstract Mappings (AM) 14–16 years		A2, Abstract mappings 14–16 years	Formal	Stage 3 15 years	Stage 3 15–16 years	Stage 4 16 years
Single Abstractions (SA) 10–11 years		R4, A1, Single Abstractions 10–12 years	Abstract	Transition 2/3 12 years		Stage 3 14 years
Representational ˜ Systems (RS) 6–7 years	Consolidated Concrete Operations	Tier 3, R3, Representational Systems 6–7 years	Concrete	Stage 2 9 years	Stage 2 7–8 years	
Representational Mappings (RM) 3.5–4.5 years		R2, Representational Mappings 3.5–4.5 years	Primary	Stage 1 ?	Stage 1 ?	
Single Representations (SR) 18–24 months		S4, R1, Single Representation 18–24 months	Pre-operational			

Notes: [a]Age estimates based on pooled sample of 996 moral judgment interviews (Dawson, 2002c). Correspondences with complexity levels empirically verified in Dawson, Xie, & Wilson, 2003). [b]Correspondences with complexity levels empirically verified by Dawson (2000). [c]Age estimates represent ages at which a level appears in high support conditions. Correspondences with complexity levels empirically verified in Dawson (2002, January). Kitchener, Lynch, Fischer, & Wood (1993) report somewhat different correspondences.

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. When scoring texts, hierarchical order of abstraction refers primarily to the structure of the elements of arguments, which must be *inferred* from their meaning in context, whereas logical structure refers to the *explicit* way in which these elements are co-ordinated in a given text. Note that conceptual and logical structures are similarly defined and fundamentally interdependent. We make a distinction between the two types of structure for heuristic and pragmatic reasons.

Hierarchical order of abstraction is observable in texts because new concepts are formed at each complexity level as the operations of the previous complexity level are hierarchically integrated into single constructs. Halford (1999) suggests that this integration or "chunking" makes advanced forms of thought possible by reducing the number of elements that must be simultaneously co-ordinated, freeing up processing space and making it possible to produce an argument or conceptualization at a higher complexity level. Interestingly, at the single representations, single abstractions, and single principles complexity levels, the new concepts not only co-ordinate or modify constructions from the previous complexity level, they are qualitatively distinct conceptual forms—representations, abstractions, and principles (or axioms), respectively (Fischer, 1980). The appearance of each of these conceptual forms ushers in three repeating logical forms—elements, mappings or relations, and systems. 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 level of a performance. For example, the statement, "If you hit dogs they might bite you," is structurally identical to the statement, "If you abuse dogs they may become vicious." They are both mappings. It is only by determining the hierarchical order of abstraction of the elements *hit, bite* (representations), *abuse*, and *vicious* (abstractions), that we can accurately place these statements at representational mappings and abstract mappings, respectively.[3] Other researchers have observed and described similar conceptual forms and repeating logical structures (Case, 1998; Fischer & Bidell, 1998; Overton, Ward, Noveck, & Black, 1987; Piaget & Garcia, 1989). Detailed descriptions

[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 level are abstract in a more particular sense; the new conceptions of this complexity level co-ordinate representations.

[3]The determination of the hierarchical order of abstraction of an individual lexical item involves an interpretation of the meaning of the item given the broader context in which it is embedded. Consequently, it is generally inappropriate to score a single sentence.

of complexity levels and the conceptual content associated with these levels (including moral content) can be found at the LAS website (Dawson-Tunik, 2004b).

RELIABILITY AND VALIDITY OF THE LAS

We have undertaken several studies of the reliability and validity of the LAS and its predecessors (Dawson, 2002a, 2003, 2004; Dawson & Gabrielian, 2003; Dawson et al., 2003; Dawson-Tunik, 2004a). We have examined inter-analyst agreement rates, compared scores obtained with the LAS with scores obtained with more conventional scoring systems, and examined scale characteristics with statistical modeling. Inter-analyst agreement rates have been high, 80–97% within half of a complexity level (Dawson, 2004; Dawson & Gabrielian, 2003; Dawson-Tunik, 2004a).[4] Correspondences between other developmental scoring systems and the LAS are also high, consistently revealing agreement rates of 85% or greater within one half of a complexity level (Dawson, 2002a, 2004; Dawson et al., 2003). Employing Rasch scaling, which provides reliability estimates that are equivalent to Cronbach's alpha, we have consistently calculated reliabilities over .95 (Dawson, 2002a; Dawson et al., 2003; Dawson-Tunik, 2004a). Overall, our research shows that the LAS is a valid and reliable general measure of intellectual development from early childhood through adulthood.

PROPERTIES OF STAGES

Because developmental stages represent successive hierarchical integrations—meaning that each new stage is constructed from the actions of the previous stage—the sequence of development, if stages are true, must be from one stage to the next with no skipping.[5] It is conventionally held that the only way to provide evidence in support of invariant sequence is to conduct longitudinal research, showing empirically that, within persons, development proceeds sequentially (Armon, 1984a; Colby, 1981; Walker, 1982). In the moral domain, several longitudinal studies of this kind have been conducted, providing support for invariant sequence (Armon, 1984a; Armon & Dawson, 1997; Colby, 1981; Rest, 1975; Walker, 1982; Walker, Gustafson, & Hennig, 2001). Despite the fact that longitudinal data can

[4]Certified LAS analysts must maintain an agreement rate of 85% within one third of a complexity level with a certified master analyst (Dawson-Tunik, 2004b)

[5]The criterion of no regressions is often added to the invariant sequence requirement. However, some dynamic systems models of developmental processes predict regressions at stage transitions. These have been identified longitudinally (Fischer & Bidell, 1998; van der Maas & Molenaar, 1992; van Geert, 1998).

make a compelling case for invariant sequence, there are patterns of performance in cross-sectional data that provide support for sequentiality (Fischer & Bullock, 1981). For example, sequentiality is supported by evidence that individuals always perform (within measurement error) at complexity levels that are adjacent to one another in the specified sequence.

Each of Piaget's original stages is defined by a set of formal properties that constitute a *structure d'ensemble*, or structure of the whole (Piaget & Inhelder, 1969). This has sometimes been taken to mean that the entire knowledge system forms a single unified global structure (Fischer & Bullock, 1981). In some (not all, see Lourenco & Machado, 1996) interpretations of stage transitions based on Piaget's notion of *structure d'ensemble,* abrupt *global* reorganizations of the entire knowledge system characterize development, which is modeled as a staircase. However, because analogous structures—especially analogous structures in different knowledge domains (Demetriou & Efklides, 1994; Fischer & Bidell, 1998)—often do not develop in parallel, attempts to model development globally (in multiple domains) will almost certainly produce patterns that make development appear more or less continuous. In response to the lack of evidence for global step-like development, some have argued that development is better characterized as smooth and continuous. Flavell (1971), for example, suggests that progress through developmental stages is characterized by a gradual replacement of lower-stage structures over time as shown (in an idealized form) in Figure 1.

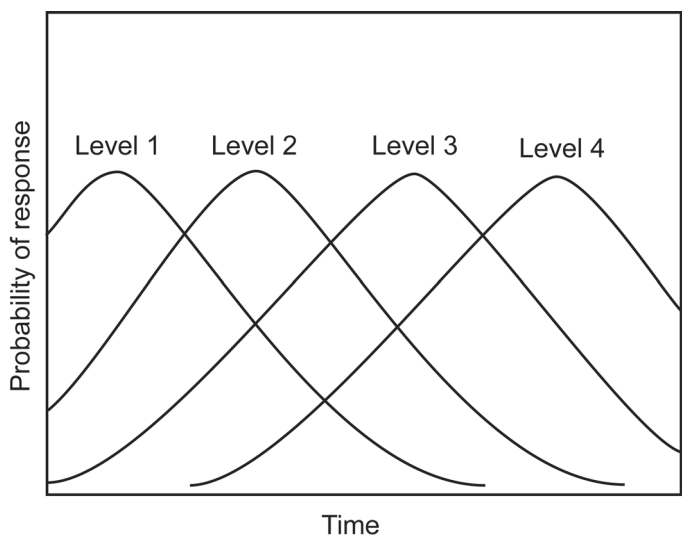

Figure 1. Probability of response, continuous, smooth learning model.

Others, like Seigler (1996), have suggested that development is continuous, though not smooth, as represented in Figure 2. Still others argue that development within individual knowledge domains is characterized by periods of consolidation (plateaus) during which performance within a domain tends to be largely homogeneous (is predominantly at a single stage), and transitional periods (spurts) characterized by vacillation between the modal stage and its successor (Fischer & Rose, 1999), as represented in Figure 3. Several researchers have provided evidence compatible with the latter two models, including evidence of spurts, drops, or shifts during developmental transitions in childhood, adolescence, and early adulthood (Andrich & Styles, 1994; Case, 1991; Draney, 1996; Fischer & Rose, 1994, 1999; Fischer & Silvern, 1985; Kitchener et al., 1993; Shultz, 2003; Thomas & Lohaus, 1993; van der Maas & Molenaar, 1992; van Geert, 1998; Walker et al., 2001; Wilson, 1989). These models can be characterized as representing wave-like patterns of development (Siegler, 1996). Figure 3 presents the most step-like model, in that change from one level to the next is more abrupt than in Figures 1 and 2, but it still can be characterized as wave-like, because periods of overlap between adjacent levels are evident. In a cross-sectional sample, with age evenly distributed, the pattern shown in Figure 3 only would be observed if individuals tended to perform primarily at one level or at two adjacent complexity levels.

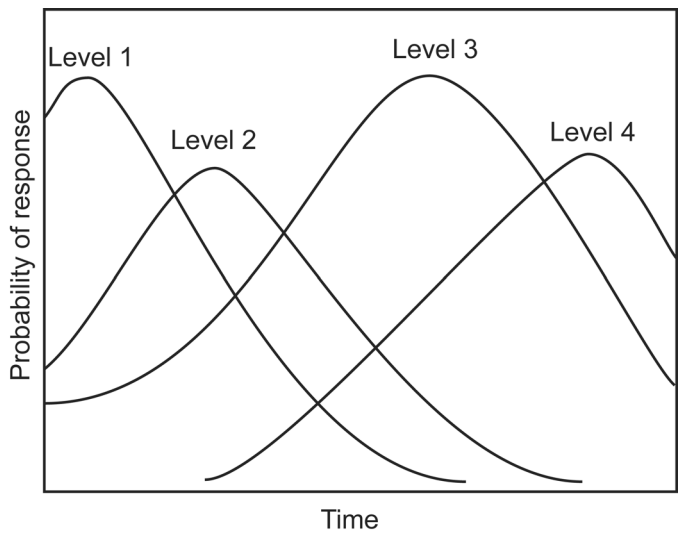

Figure 2. Probability of response, continuous, not smooth development model.

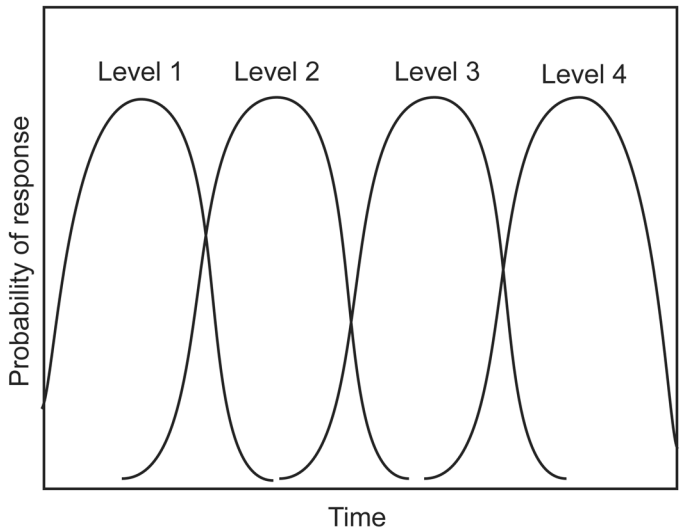

Figure 3. Probability of response, discontinuous model.

THE RASCH MODEL

Whereas they are well-known in psychometric circles, Rasch's (1980) models for measurement have been employed by developmental psychologists only recently (Andrich & Constable, 1984; Bond, 1994; Dawson, 1998, 2000, 2002c; Draney, 1996; Müller, Sokol, & Overton, 1999; Wilson, 1984). These models are designed specifically to examine hierarchies of person and item performance, displaying both person proficiency and item difficulty estimates along a single interval scale (logit scale) under a probabilistic function. In addition, they can be employed to test the extent to which items or scores conform to a theoretically specified hierarchical sequence. A central tenet of stage theory is that cognitive abilities develop in a specified sequence, making the statistical tests implemented in a Rasch analysis especially relevant to understanding stage data. The Rasch model permits researchers to address questions like, ''Are all single abstractions items more difficult than all representational systems level and less difficult than all abstract mappings items?'' Moreover, the detailed information about item functioning and individual performances provided by the software makes it possible to simultaneously examine group and individual effects. These properties make Rasch models uniquely suitable for the investigation of many developmental phenomena.

It is beyond the scope of this paper to provide a comprehensive account of the Rasch model, though we do attempt to provide enough information to allow readers who are unfamiliar with the model to follow the results of the analysis. For an introduction to the Rasch model, see Bond and Fox (2001), Rasch (1980), Smith (2004), or Wilson (2005).

EDUCATION AND AGE

Both age and education are associated with hierarchical development. Approximate ages of acquisition for Fischer's skill levels, LAS complexity levels, Kohlbergian stages, Armon's good life stages, and Kitchener and King's reflective judgment stages are shown in Table 1. As Fischer and his colleagues (Fischer & Silvern, 1985) have shown, the age at which the structures of a particular complexity level typically appear are influenced by the type of task administered, the level of support provided, the domain of the task, and the affective salience of the task. Scoring criteria also influence the age at which complexity levels are identified. For example, some scoring systems, like the those of Kohlberg (Colby & Kohlberg, 1987a) and King and Kitchener (1994), require that particular conceptualizations (or "surface" structures; Dawson, 2001a) are present in a performance before it can be scored at a given level, whereas other systems, like the LAS, require only that the "deep" structures of a given complexity level are present.

Age is most strongly associated with stage in childhood and early adolescence. Dawson (2002c), for example, reports a correlation of .75 between Kohlbergian moral judgment stages and the natural log of age in a lifespan sample of 965 moral judgment interviews, and Armon (1984b) reports an identical correlation of .75 between good life stages and the natural log of age in a lifespan sample of 37 good life interviews. Both authors observe that the relation between age and stage becomes less deterministic as age advances.

Although the relation between age and stage weakens substantially in adulthood, the relation between educational attainment and stage remains fairly strong. Dawson (2002c), for example, reports a strong linear relation ($r = .79$) between educational attainment and Kohlbergian stage in a lifespan sample of 928 moral judgment interviews.

GENDER

One of the disadvantages of domain-specific developmental assessment systems is that they are open to accusations of bias. All domain-specific scoring systems are based on a limited sample of interviews. Kohlberg's scoring system, for example, was based on the performances of seven male respondents who were interviewed on six occasions at four-yearly intervals

(Colby & Kohlberg, 1987a). One well-known criticism of Kohlberg's model, resulting, in part, from the bias introduced by the all-male composition of his construction sample, concerns gender bias (Gilligan, 1977). However, three surveys of research conducted with Kohlberg's instrument provide evidence that there are no systematic differences in the *moral stage* of males and females after differences in education are taken into account (Pratt, Golding, & Hunter, 1984; Walker, 1984, 1994).

Although developmental assessments based on domain-general criteria, such as those based on Fischer's skill theory or Commons' General Stage Model have not been subjected to claims of gender bias, gender differences have been examined. When they are reported, statistically significant gender differences in the rate of development are small, often domain-dependent, and usually age- or cohort-related (Blackburn & Papalia, 1992; Dawson, 1998, 2002c; Kitchener et al., 1993; Overton & Meehan, 1982; Sprinthall & Burke, 1985).

In the following analysis, we employ the Rasch rating scale model, along with more conventional statistical procedures, to examine patterns of performance in a large lifespan sample of moral judgment interviews scored with the LAS.

METHOD

Sample

The interview data for this analysis were collected by several researchers between 1955 and the present. The sample sizes and sources are:

- 167: Kohlberg's original longitudinal study (Colby & Kohlberg, 1987a)—33 from the first test time, 41 from the second test time, 19 from the third test time, 42 from the fourth test time, 9 from the fifth test time, and 23 from the sixth test time;
- 79: Armon's lifespan longitudinal study (Armon & Dawson, 1997)—2 from the first test time, 28 from the second test time, 25 from the third test time, and 24 from the fourth test time;
- 247: Berkowitz study of adolescents and their parents (Berkowitz, Guerra, & Nucci, 1991);
- 2: Walker's longitudinal study of children and their parents (1989)—36 from the first test time and 76 from the second test time;
- 12: Commons' study of Harvard professors and students (Commons, Danaher, Griffin, & Dawson, 2000);
- 23: Ullian's study of elementary school students (1977);
- 31: Drexler's study of young boys (1998); and
- 76: Dawson's study of young children and adolescents (2001b).

The total number of interviews is 747. All of the interviews are Kohlbergian moral judgment interviews. Other interview material collected by the original researchers is not included. Ages range from 5 to 86 years ($M = 25.38$, $SD = 15.93$), and educational attainment ranges from 0 to 22 years ($M = 0.38$, $SD = 5.74$). Males outnumbered females (459 male, 288 female) largely because Kohlberg's and Drexler's studies included only males. The population sampled is 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. The data-collection methods were also somewhat different from study to study, as described further below. We anticipated that the differences in data completeness and quality from study to study would add noise to the results of this analysis rather than resulting in systematic patterns that supported our theoretical position.

We did not include all of the interviews collected for the above studies in this data set. Interviews were excluded for one of two reasons. First, most of the original data were in hard-copy form. To serve our purposes, the data had to be translated into electronic form.[6] We scanned only the interviews that could be successfully translated with our software. Second, some of the respondents in the original studies did not receive the interviews we decided to include in our sample, as explained further below.

Some of the interviews in our sample, which we treat as a cross-sectional sample, are actually interviews of the same respondent, conducted at different times. Most of these, including those in the Kohlberg and Armon sample, were conducted at four-yearly intervals. A small number, those in Walker's study, were conducted at two-yearly intervals. We treat all of these as independent observations in the following analyses. Treating the interviews collected at more than one test-time as independent observations is justified when test-times are separated by a sufficient interval (Willett, 1989), and is a common practice in this type of research (Armon & Dawson, 1997; Bond & Fox, 2001; Dawson, 2000, 2002c). Given that (1) the intervals between test-times are 2 to 4 years, and (2) patterns in the data remain the same when the longitudinal cases are eliminated from the data, we elected to increase the power of the analysis by pooling the longitudinal data into the cross-sectional sample.

Older participants were interviewed with Form A of the MJI (the Heinz and Joe dilemmas) as described in the Standard Issue Scoring Manual (Colby & Kohlberg, 1987b). Because the Heinz dilemma is difficult for young children to follow, the youngest participants were administered the Joe dilemma and/or the Picnic dilemma, the latter of which was developed

[6]Electronic data were required because these interviews were also employed in the development of an electronic version of the LAS (Dawson & Wilson, 2004).

specifically for young children by Damon (1980). All of the interviews were similarly designed to elicit moral judgments and justifications.

Because participants responded to different numbers and sets of dilemmas, it was important to determine whether the dilemmas functioned similarly as assessments of conceptual development in the moral domain. To make this determination, we examined the correlations between complexity level scores on pairs of dilemmas. The correlation between mean scores on 615 Heinz and Joe dilemmas was .93. Mean scores agreed within one half of a complexity level 90% of the time and differences were evenly distributed around zero. The correlation between mean scores on 63 Joe and Picnic dilemmas was .85. Mean scores agreed within one half of a complexity level 92% of the time and differences were equally distributed around zero.

The dilemmas were read to respondents. Pictures were used to assist the understanding of 4 to 6 year olds. A series of standard follow-up questions were employed to probe for respondents' understanding of the dilemma and their reasoning on each of the probe questions. For example, respondents who received the Joe dilemma were asked how important it is for a father to keep a promise to his son, and their responses were further probed to reveal the reasoning behind them. Interviews were audio recorded and transcribed by the original investigators.

The range of standard probes that were actually administered from study to study varied greatly. Though each dilemma included nine standard probes (total = 27), each respondent received only 5 to 16 standard-probe questions. This means, for example, that an adolescent or adult might have received probes 1–9 from the Joe dilemma and only probes 1–7 from the Heinz dilemma or that a young child might have received probes 1–3 from the Joe dilemma and 1–8 from the Picnic dilemma. The mean number of probes received was 9.46 ($SD = 2.22$). There were three main reasons for the missing data. First, for a variety of reasons, including time constraints and inconsistent interviewing, some researchers only used a subset of probes in their studies. Second, respondents sometimes tired before the interviews were complete. Finally, some probes were considered unscorable because they did not include both judgments and justifications (one of the requirements of the LAS).

The scoring unit for this study was the complete response associated with each probe question. We called each of these units a protocol. This meant that 5 to 16 protocols were scored for each case, depending on the number of probe questions administered. These were individually scored by a single trained analyst, the first author, following LAS criteria. Each protocol was treated as an individual item in the following analyses.

Missing data pose problems for statistical analyses. For example, when data are missing, mean scores cannot be assumed to be reliable, because they are calculated on the basis of different items from case to case. Though a

complete data set is always desirable, an important feature of the Rasch model is that it is robust to missing data. In the rating scale analysis that follows, missing data are treated as missing at random. The main consequence of treating missing data in this way is that estimates for cases with a great deal of missing data will be (1) associated with larger error estimates, and (2) biased toward the mean. Both can obscure evidence of developmental patterns.

Scoring

The LAS (Dawson-Tunik, 2004b) requires the analyst to identify both the highest level of abstraction and most complex form of organization in text performances. A protocol is considered to be at a given complexity level if its elements embody the hierarchical order of abstraction of that complexity level and the complexity of its logical structure meets the formal requirements of that complexity level. In these data, seven consecutive complexity levels were identified:

- (0): single representations (SR, in tables and figures);
- (1): representational mappings (RM, in tables and figures);
- (2): representational systems (RS, in tables and figures);
- (3): single abstractions (SA, in tables and figures);
- (4): abstract mappings (AM, in tables and figures);
- (5): abstract systems (AS, in tables and figures); and
- (6): single principles (SP, in tables and figures).

One score was awarded to each protocol. Ideally, a protocol should represent a complete argument on a given topic, including at least one judgment and a justification for that judgment. The score awarded is always for the most hierarchically complex argument in the protocol. Fragmentary arguments are usually treated as unscorable. However, in cases where this would have reduced the number of protocols to fewer than five (primarily interviews of young children), we chose to score some fragmentary protocols if adjacent protocols in a given text provided enough information to aid in their interpretation. This meant that the rater sometimes accessed the entire interview when scoring. This is the standard practice in this type of research (Armon, 1984b; Colby & Kohlberg, 1987a; King & Kitchener, 1994). However, we would have preferred if it had been possible to score each protocol blind to its origins as was done in Dawson's (1998) study of evaluative reasoning about education. Although the sample in this earlier study was considerably smaller than the sample in the present study, restricting the generalizability of the results, patterns of performance were substantively the same as those reported here.

Correlations between mean scores of four independent raters on a subset of 2 randomly selected cases ranged from .95 to .98. Agreement rates ranged from 80% to 97% within half a complexity level and from 98% to 100% within a full complexity level. This equals or exceeds inter-rater agreements commonly reported in this field (Armon, 1984a; Colby & Kohlberg, 1987a; King & Kitchener, 1994).

Rasch analysis

When complexity level scores are in their raw ordinal form, it is possible to calculate a mean score for each case, to examine the proportions of respondents assigned to each complexity level or transitional phase, and to account for the range of complexity levels represented in a given performance. On the other hand, little can be said about the confidence we can place in these mean scores, the amount of difficulty associated with moving from complexity level to complexity level, or whether the difficulty of making transitions changes depending on where you are in the developmental sequence. Rasch analysis software makes it possible to address these questions by using a log-odds transformation (Wright & Masters, 1982) to convert ordinal data into distinct quantitative estimates of (1) item difficulty and (2) person performance, both of which are expressed in the same equal-interval units.

The formulation of the original dichotomous Rasch model can be expressed as:

$$\log_e(P_{ni1}P_{ni0}) - B_n - D_i$$

Where P_{ni1} is the probability that person n encountering item i is observed in category 1, B_n is the "ability" measure of person n, and D_i is the "difficulty" measure of item i—the point where the highest and lowest categories of the item are equally probable. This model has been extended to specify the polytomous rating scale and partial credit models, which can be expressed as:

$$\text{Rating scale}: \log(P_{nij}/P_{ni(j-1)} = B_n - D_i - F_j$$

$$\text{Partial credit}: \log(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_{ij} = B_n - D_{ij}$$

Where P_{nij} is the probability that person n encountering item i is observed in category j, B_n is the "ability" measure of person n, D_i is the "difficulty" measure of item i—the point where the highest and lowest categories of the item are equally probable, D_{ij} is the difficulty measure of item i, category j, and F_j is the "calibration" measure of category j relative to category j (1, the point where categories j (1 and j are equally probable relative to the measure of the item.

The product of a Rasch analysis is an equal-interval scale, along which both item difficulty and respondent performance estimates are arranged. Each unit on the scale is referred to as a logit, each of which represents an identical increase in difficulty. The range of difficulty represented in the items of a scale determines the number of logits. The probabilistic interpretation of logits is explained further below.

All Rasch analysis software packages provide error terms for all item and person estimates, establishing the confidence one can place in them. Performances and items with more missing data are associated with larger error terms than those with less missing data, and performances that are predominantly at a single level are associated with larger error terms than performances that include a mixture of levels.[7] Estimates for cases with missing data are biased toward the mean.

The fit statistics included in the following analysis are called infits. Infit statistics are used to assess whether a given performance (or item) is consistent with other performances (or items). They are based on the difference between observed and expected performances. Infits near 1 are desirable. *Z*-scores are calculated to assess the significance of both positive and negative divergences of infit statistics from 1. *Z*-scores between (2.0 and + 2.0 are considered acceptable. Interpretation of infit statistics and *Z*-scores are demonstrated below, in the results section.

All analyses were conducted with the computer program, *Winsteps* (Linacre & Wright, 2000).[8] The *rating scale model* (Wright & Masters, 1982), a special case of the Rasch partial credit model (Masters, 1982), is employed here. Whereas the partial credit Rasch model estimates values for each level of each item independently, the rating scale model assumes that steps between adjacent item levels can be considered to be equivalent across items (see model specifications, above). This means, for example, that the difficulty of moving from the RS level to the SA level should be the same (taking into

[7]In a sense, performances that are scored predominantly at a single level provide less information than performances scored at a mixture of levels. Consequently, consolidated performances are associated with larger error terms in the same way that performances with fewer data points are associated with larger error terms.

[8]In keeping with the original formulation of the Rasch model, *Winsteps* treats person parameters as fixed effects. It has been argued that this limitation of the model restricts the generalizability of the results of Rasch analyses (Bartholomew & Knott, 1999; Mislevy & Wilson, 1996), though the specific implications for research of the present kind are not entirely clear. Moreover, several researchers employ *Winsteps* and other software that treats person parameters as fixed effects to explore developmental constructs similar to those examined here (Bond & Fox, 2001; Dawson, 2002c; Müller et al., 1999). In any case, concerns about generalizability are minimized in the present project by the large size and heterogeneity of the sample (Canadian Christian families, boys from New England private schools, working class mid-western families, California elementary school students, and a convenience sample from all over the United States).

account measurement error) across all interview questions. Likewise, the difficulty of moving from the SA to AM level should be the same (taking into account measurement error) across all interview questions. From a theoretical perspective, the difficulty of moving from any given complexity level to the next should be the same across protocols since the criteria for determining the complexity level of responses do not vary from protocol to protocol, and we expect the reasoning of any one individual to be relatively consistent across protocols. In other words, our measure and theory are compatible with this requirement of the rating scale model. A second requirement of the rating scale model is that the number of hierarchically ordered response possibilities is the same for each item. The present data meet this requirement, in the sense that it is possible to perform at any complexity level on any item.

To test whether patterns of performance provide an adequate empirical justification for using the rating scale model, we also ran a partial credit analysis (Masters, 1982), and compared the results with those from the rating scale analysis. The correlation between these item estimates was .99. With the exception of six outlying estimates, the item level estimates for the two models lay on the identity line, indicating that the more parsimonious rating scale model provides item level estimates that correspond well to those of the more saturated partial credit model. Moreover, the results of the partial credit analysis were identical to those from the rating scale analysis with respect to the research questions addressed here. For these two reasons, we concluded that the more parsimonious rating scale model provided an adequate account of patterns in these data (for more on model selection, see Wright, 1999).

RESULTS

Rating scale analysis

The key results of the rating scale analysis are shown in the form of a variable map in Figure 4, where both person performance and item level difficulties are located on the same interval scale. At the left of the figure is the logit scale, which spans 46 logits. In the middle are the person performance estimates. Here, each vertical line represents one person. To the right are *item level* estimates. The complexity level represented by each cluster of items is indicated on the far right of the figure.

The wide logit range shown in Figure 4 reflects both the wide range of complexity levels represented in the sample and the tendency for individuals to perform predominantly at a single complexity level or at two adjacent complexity levels rather than three or more complexity levels. When performances are less Guttman-like (1944)—the number of logits repre-

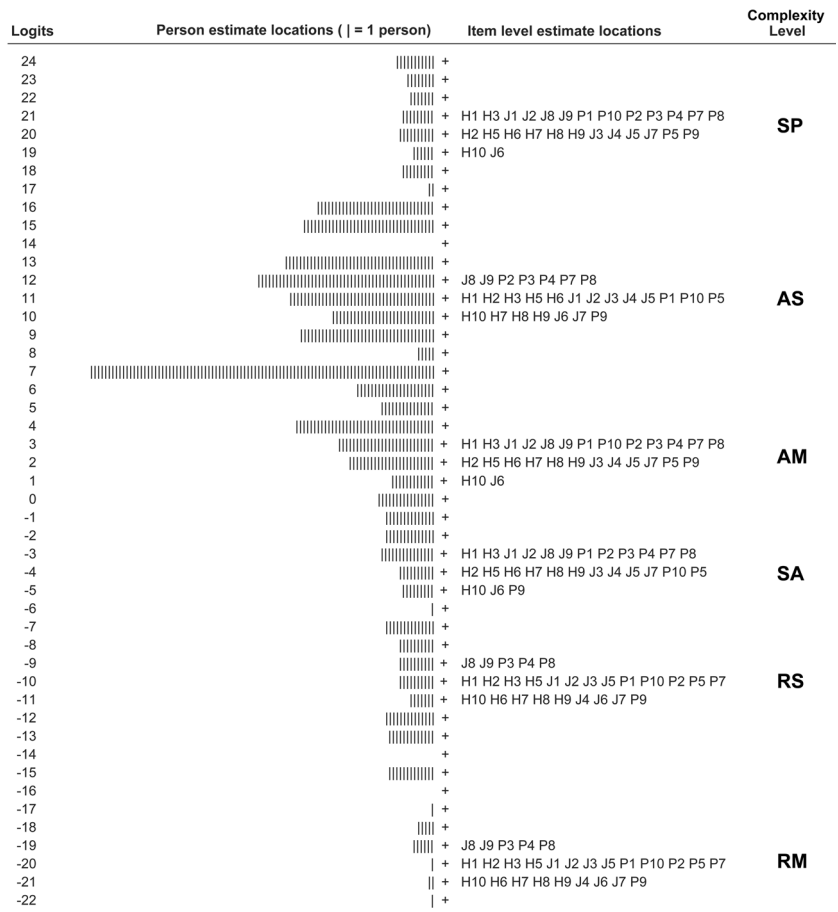

Figure 4. Map of person and item level estimates.

sented in the resulting Rasch model are far fewer. In Figure 4, the distance between adjacent groups of item level estimates spans 6 to 11 logits. Gaps spanning several logits between groups of item level estimates, are a strong indication of a stage-like, discontinuous growth pattern (Wilson, 1985). This is because the distance between logits has a particular probabilistic meaning. In the present case, an ability estimate for a given individual means that the probability of that individual performing accurately on an item at the same level is 50%.[9] There is a 73% probability that the same individual will perform accurately on an item whose difficulty estimate is one logit easier,

[9]This is the default setting in most Rasch analysis software.

an 88% probability that he or she will perform accurately on an item whose difficulty estimate is two logits easier, and a 95% probability that he or she will perform accurately on an item whose difficulty estimate is three logits easier. The same relationships apply, only in reverse, for items that are one, two, and three logits harder. In the present case, this means, for example, that an individual whose estimate is between 13 and 18 logits has a high probability of performing at the AS level on all protocols, whereas an individual whose estimate is in the range of 10 to 12 logits has a high probability of providing some arguments scored at the AM level and some scored at the AS level.

Person performance analysis

The overall separation reliability for the person performance estimates is .97. The separation reliability statistic is equivalent to Cronbach's alpha, and is based on the ratio of the variation in the mean squares (the standard deviation) to the error of measurement (Wright & Masters, 1982). In keeping with established fit interpretation criteria (Wright & Masters, 1982), the infit statistics for all person performance estimates were considered to fit the model if z-scores were greater than -2.0 and less than $+2.0$. Fit statistics higher than 2.0 indicate less consistency within performances than expected. Only four performances out of 747 ($< 1\%$) are too erratic/unpredictable to fit the measurement requirements of the Rasch model. Each of these subjects provided one or more protocols that were scored higher or lower than expected for a person located at that overall performance level. Fit statistics lower than ± 2.0 indicate less performance variability, or more consistency within performances than expected by the Rasch model. Thirteen performances (2%) show these very consistent patterns of protocol rating. These low rates of misfit to the Rasch model are well within the routinely accepted 95% confidence interval for the normal distribution.

The mean standard error for person performance estimates is 1.22 logits ($SD = 0.62$), indicating that, on average, the confidence interval around person performance estimates is -2.44 to $+2.44$ logits. While a confidence interval around person performance estimates of -2.44 to $+2.44$ logits seems imprecise, it is relatively small in the context of a 46-logit scale. In fact, locations are precise (on average) within approximately one third of a complexity level above or below the estimated complexity level.

Item analysis—restructuring and invariant sequence

The infit statistics for all of the item level difficulty estimates were considered to fit the model if z-scores were greater than -2.0 and smaller than $+2.0$.

All of the infit *z*s are well below 2.0. However, the infit *z*s for five items are less than −2.0. There is less random variation in performances on these items than expected by the model. Take, for example, the case of question H3, *What should Heinz do if he does not love his wife?* It has an infit *z*-score of ±2.3, suggesting that individuals who have a high probability of performance at a given complexity level, say the representational systems complexity level, are almost always awarded a representational systems score on Heinz 3, whereas individuals with estimates that reflect a high probability of performance at the representational mappings, single abstractions, abstract mappings, abstract systems, or single principles complexity levels are almost never awarded a representational systems score on Heinz 3. In a sense, from the perspective of the Rasch model, the pattern of performance on this item is "too good to be true." However, this is an acceptable pattern from our developmental perspective (Wilson, 1985), reflecting the fact that a large percentage of individual performances are predominantly at a single complexity level, a pattern that is expected when qualitative rather than additive change takes place. See Wilson (1989) for a discussion of the psychometric modeling perspective on this phenomenon in the context of development.

The mean standard error for item level estimates is .22 logits ($SD = 0.07$), indicating that, on average, a difference between item level estimates of over .88 logits would be statistically significant at the $p < .05$ level. The distance between estimates for adjacent complexity levels is always greater than four logits—thus there is no overlap in 95% confidence intervals around groups of item level estimates. In other words, the gaps between groups of item-level estimates are statistically significant, supporting the notion that complexity levels represent distinct forms of reasoning marked by periods of remarkable coherence (Fischer & Bidell, 1998; Wilson, 1985).

The graph of response probabilities in Figure 5 illustrates how distinct the complexity levels are in a qualitative sense. The pattern is strikingly similar to the pattern shown in Figure 3. Each curve is labeled with its corresponding complexity level abbreviation. Here, for example, an individual with a performance estimate of +13 logits has about a 10% probability of performing at the abstract mappings complexity level, and a 90% probability of performing at the abstract systems complexity level on any protocol. In this figure the only overlap of item level estimates occurs at adjacent complexity levels. Note how the curves for non-adjacent complexity levels converge near the 0% probability level. So, for example, an individual with a score of −7 has a near 0% probability of scoring at the representational mappings or single abstractions complexity levels and an over 95% probability of scoring at the representational systems complexity level. Further, this pattern is virtually identical from complexity level to complexity level, though there is a slightly lower probability for

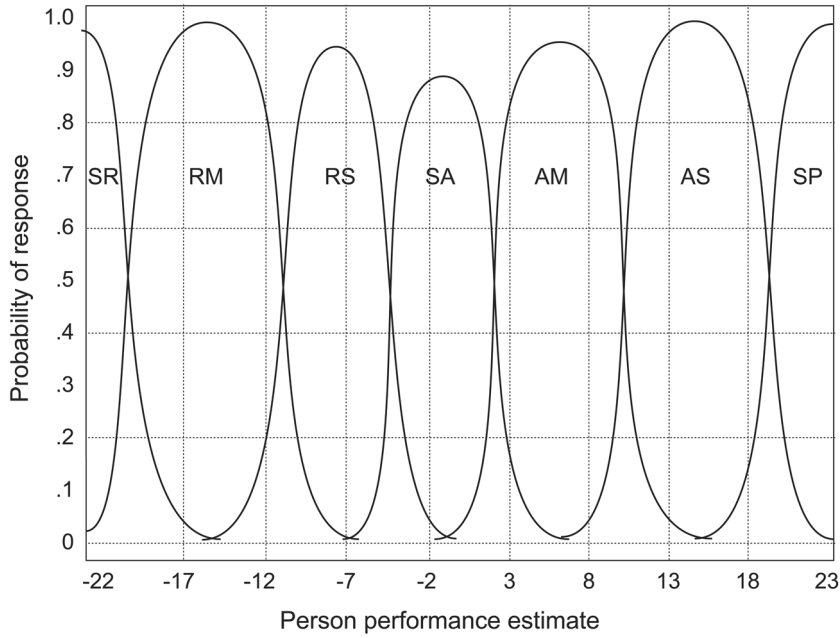

Figure 5. Probability of complexity order response by person performance estimate.

consolidated performances at the representational systems and single abstractions complexity levels than at the representational mappings, abstract mappings, and abstract systems complexity levels.

Figure 6 shows the relation between expected scores and person estimates. Plateaus are clearly visible at the representational mappings, representational systems, single abstractions, abstract mappings, and abstract systems complexity orders, in the sense that there are relatively wide logit ranges in which consolidated performances are expected, followed by periods characterized by a mixture of reasoning at the current complexity level and an increasing amount of reasoning at the subsequent complexity level. The wide logit ranges in which consolidated performances are expected reflect the fact that in this cross-sectional sample there are more consolidated performances than would be expected to occur in the population at any one point in time if development was smoothly continuous. If development was smoothly continuous, there would be no more consolidated performances than performances representing any other mixture of complexity levels. The Rasch analysis reveals a pattern that is consistent with Fischer & Rose's (1999) characterization of development as

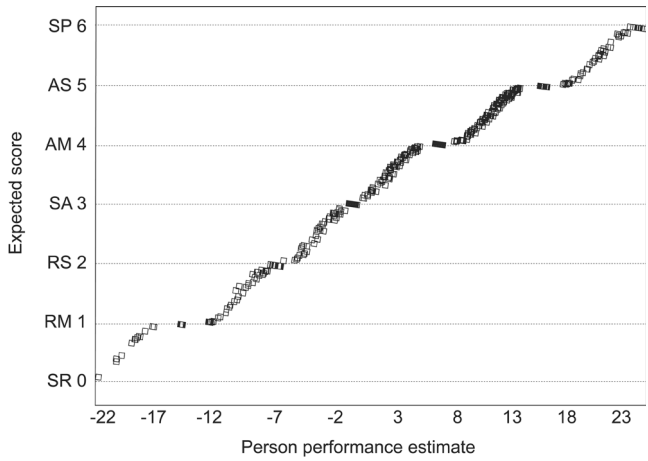

Figure 6. Expected scores by person performance estimates.

proceeding through successive periods of consolidation (plateaus) during which performance within a domain is predominantly at a single level followed by transitional periods (spurts) characterized by vacillation between the modal stage and its successor.

Age, educational attainment, and sex

The sample collected for this project differs from those employed by others to examine age and education effects in one important way. Most research on the relations between cognitive development and age or education were conducted with persons who were still in the process of receiving their formal education (continuously educated persons). Many of the adults in our sample either had completed their education long before being interviewed or had resumed their education after several years outside of the educational system. Consequently, to allow comparison of the ages at which the higher complexity levels emerged in this sample with ages of emergence reported elsewhere, the following analysis is restricted to the continuously educated cases in the larger sample. There are 406 cases in this subsample, with an age range of 5 to 23 and educational attainment from pre-kindergarten to four years of college. In this subsample, representational mappings, representational systems, single abstractions, abstract mappings, and abstract systems performances do not dominate until the ages of 5, 7, 10, 14, and 22, respectively, corresponding to educational attainment (completion) levels of Pre-K, 1, 4, 8, and 16 years. The relation between age

TABLE 2
Percentage of cases in each age group assigned to each complexity order

		Complexity order					
Age	*N*	*RM*	*RS*	*SA*	*AM*	*AS*	*SP*
5	28	**89.3**	03.6				
6	35	**60.0**	34.3				
7	29	20.6	**75.8**	03.4			
8	21	04.8	47.6	47.6			
9	16		50.0	50.0			
10	28		03.6	**92.8**	03.6		
11	04		25.0	**75.0**			
12	10			**60.0**	40.0		
13	22		04.5	50.0	45.5		
14	37			29.7	**70.2**		
15	24			33.4	**66.6**		
16	38			07.8	**92.1**		
17	46			06.5	**84.8**	08.6	
18	33			06.1	**75.8**	18.2	
19	03				**66.6**	33.3	
20	12				**58.3**	41.6	
21	10				50.0	50.0	
22	04					**75.0**	25.0
23	09					**77.8**	22.2

and complexity level in this subsample is further illustrated in Table 2, which shows the percentage of performances assigned to the representational mappings to abstract systems complexity levels by age. Percentages over 50% have been highlighted to show the age-ranges within which each complexity level dominates. Complexity level assignments represent the modal complexity level at which each respondent performed; individuals with 50% of their protocols at one complexity level and 50% at an adjacent complexity level were assigned to the lower complexity level. In this sample, the ages at which each complexity level first becomes dominant are similar to those reported by Fischer for analogous skill levels, with the exception of the abstract systems order, which Fischer and his colleagues (Fischer & Bidell, 1998) found from 18 to 20 years of age, and which does not dominate in this subsample until 22 years of age.

It is not possible to compare the age of emergence of the single principles complexity level with those reported by other researchers due to the constricted age range in this subsample. In fact, single principles performances never constitute the largest percentage of performances by age group, even when the entire sample of 747 interviews is included in the

analysis. However, single principles performances are the plurality after the attainment of a doctoral degree or its equivalent, which, in continuously educated groups, would make the modal age for the attainment of single principles reasoning somewhere in the range of 26 to 30 years, somewhat older than the 23 to 25 age range for principled reasoning found by Fischer and his colleagues. Despite the need for additional research into the specific ages of their emergence, one thing is clear: the abstract systems and single principles complexity levels can legitimately be thought of as adult levels.

In the complete data-set, we find the relation between age and complexity level to be strong and linear up to age 18 ($r = .92$, $p < .01$, $n = 371$), but it is weak in adulthood ($r = .16$, $p < .01$, $n = 376$). The relation between educational attainment and complexity level tells a somewhat different story. Up to age 18, the correlation between complexity level and education is the same as the correlation between complexity level and age ($r = .92$). Above age 18, the correlation between educational attainment and complexity level is moderate, at .46. In the entire sample the relation between education and complexity level is strong. The linear equation is:

$$\text{Stage estimate} = -10.34 + 1.45_{ed}$$

$$F(1, 745) = 2894.06, R^2 = .80, p < .01$$

In terms of complexity levels, this means that, on average, every additional year of education results in a little more than one sixth of a complexity level of development. A quadratic equation on the residuals from the linear equation explains an additional 2% of the variance in complexity level estimates, with a decrease in the effect of education on complexity level at the higher complexity levels.

Once the linear and quadratic relations between education and complexity level have been accounted for, there is a weak but statistically significant effect of sex on complexity level, accounting for about 1% of the variance in complexity level (with $F(1, 745) = 9.19$, $R^2 = .01$, $p < .01$). Females score an average of .06 logits, or about one hundredth of a complexity level, lower than males. Though statistically significant, this difference is not meaningful.

DISCUSSION

The preceding analyses suggest that the attainment of successive complexity orders within the moral domain can be characterized as discontinuous. They provide evidence of patterns in performance that are consistent with the notion that development proceeds in a series of spurts and plateaus across six complexity levels, covering a large portion of the lifespan. Although the

two highest complexity levels are found predominantly in adulthood, the developmental pattern for these complexity levels is virtually identical to the pattern for those found in childhood. Moreover, examinations of sex, age, and education effects reveal that sex is not an important correlate of development once educational attainment has been taken into account, and education and age are both good predictors of complexity level in childhood and early adolescence, whereas education is a better predictor in late adolescence and adulthood.

As shown in the analyses presented here, patterns of performance are highly consistent from complexity level to complexity level. If any complexity level was incorrectly specified, or if complexity levels were left out, we would expect a much less systematic pattern of performance. From this we conclude that there are indeed six complexity levels in the portion of the developmental continuum modeled here.

Strong evidence for the specified sequence is found in the Rasch analysis of these data, which shows that the probability that non-adjacent complexity levels will co-occur is near zero in every instance.

Evidence that development is discontinuous comes from two sources. First, groups of item level estimates cluster in relatively narrow ranges that are separated by several logits. This shows that complexity levels are distinct from one another in a qualitative sense. Second, consolidated performances are expected more frequently than they would be if development was smoothly continuous.

Abstract systems and single principles performances are found predominantly in adulthood. Abstract systems performances do not dominate until 16 years of education have been completed, and single principles performances do not dominate until the PhD has been completed.

Although cross-sectional evidence of developmental patterns is generally not as compelling as direct longitudinal evidence, the strength and persistence of patterns in these data are striking. The Rasch model generated from these data reveals robust patterns (despite differences in data collection, missing data, and sampling variations) that are consistent with the claim that development can be characterized as proceeding in a series of spurts and plateaus.

Patterns of performance on ''adult'' stages are virtually identical to patterns of performance on childhood stages, indicating that moral cognitive development continues well into adulthood. Also, these patterns provide some evidence to support the position that the mechanisms of development may be similar across childhood and adulthood, in the sense that development appears to proceed from a period of consolidation at a given complexity level, through a period of transition, in which an individual employs the structures of adjacent complexity levels, to another period of consolidation at the subsequent complexity level.

Patterns like those described here have often eluded detection. We think this is at least partly due to the lack of reliable and accurate developmental assessment instruments. For example, domain-based instruments like Kohlberg's standard issue scoring system (Colby & Kohlberg, 1987b) have been developed by analyzing the performances of small "construction" samples and then producing complex scoring systems that incorporate examples of reasoning produced by these samples. This results in an overdependence upon concept-matching as a scoring strategy. The first author has argued elsewhere that reliance upon concept matching introduces measurement error, obscuring developmental patterns (Dawson et al., 2003). In contrast, the LAS, as reported above, provides highly reliable and accurate assessments of developmental level.

This analysis has focused on describing patterns in performances scored with the LAS. By modeling these patterns we have gained important insight into developmental processes. However, although the hierarchical complexity of a performance tells us a great deal about its form and order of abstraction, it makes no direct reference to its *specific* conceptual content. When complexity level is assessed with the LAS, specific conceptual content must be assessed independently, then reintegrated with hierarchical complexity information. While this may initially look like a weakness in the approach, it is actually a strength. For example, because complexity levels are assessed independently of *particular* conceptual content, it is possible to address questions about the relation between complexity level and meaning (Dawson, 2004; Dawson & Gabrielian, 2003; Dawson & Stein, 2004; Dawson-Tunik, 2004a; Dawson-Tunik & Stein, 2004a, 2004b; Dawson-Tunik & Stein, 2005a, 2005b; Drexler, 1998).

The independent assessment of complexity level and conceptual content also makes it possible to address questions about individual (or cultural) differences in same level behavior (Fischer & Ayoub, 1994; Fischer, Knight, & Van Parys, 1993). Though, in this sample, complexity level does not appear to be importantly correlated with sex, we can independently ask whether sex appears to relate to the conceptual content of performances. For example, are girls performing at a given complexity level more likely to refer to *care* and boys at the same complexity level more likely to refer to *justice* as predicted by Gilligan (1982)? We believe this question has never been satisfactorily addressed, because there has not previously been an objective, content-independent method for determining developmental level. The way is now open to address questions of this kind.

The version of the LAS employed here required the analyst to award a "whole" complexity level score to each protocol. When these data were scored we had no basis for awarding graded scores. As a result of lexical analyses conducted on these (and other) data (Dawson-Tunik, 2004b), we have since learned that levels of within-complexity-order

elaboration can be employed to provide more precise estimates of the level of a given performance. We can now determine whether a protocol is transitional, unelaborated, elaborated, or highly elaborated at a given complexity level. By providing these more precise assessments of the developmental level of individual protocols, we hope to eliminate concerns that the developmental pattern exposed in the current analysis is an artifact of the scoring system.

Manuscript received 7 September 2004
Revised manuscript accepted 1 March 2005

REFERENCES

Andrich, D., & Constable, E. (1984, March). *Studying unfolding developmental stage data based on Piagetian tasks with a formal probabilistic model*. Paper presented at the Annual meeting of the AERA, New Orleans.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *Journal of Early Adolescence*, *14*, 328–344.

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* (pp. 357–381). New York, NY: Praeger.

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

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

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Oxford 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.

Blackburn, J. A., & Papalia, D. E. (1992). The study of adult cognition from a Piagetian perspective. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 141–196). New York, NY: Cambridge University Press.

Bond, T. G. (1994). Piaget and measurement II: Empirical validation of the Piagetian model. *Archives de Psychologie*, *63*, 155–185.

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

Brainerd, C. J. (1993). Cognitive development is abrupt (but not stage-like). *Monographs of the Society for Research in Child Development*, *58*, 170–190.

Case, R. (1987). The structure and process of intellectual development. *International Journal of Psychology*, *22*, 571–607.

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. (1998). The development of conceptual structures. In W. Damon, D. Kuhn, & R. S. Siegler (Eds.), *Handbook of child psychology. Volume 2: Cognition, perception, and language* (5th ed., pp. 745–800). New York, NY: McGraw-Hill.

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

Colby, A. (1981). *Invariant sequence and internal consistency in moral judgment*. Cambridge, MA: Harvard University, Graduate School of Education.

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

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

Commons, M. L., Armon, C., Richards, F. A., Schrader, D. E., Farrell, E. W., Tappan, M. B., & Bauer, N. F. (1989). A multidomain study of adult development. In D. Sinnott, F. A. Richards, & C. Armon (Eds.), *Adult development, Vol. 1: Comparisons and applications of developmental models*. (pp. 33–56). New York, NY: Praeger.

Commons, M. L., Danaher, D., Griffin, M. M., & Dawson, T. L. (2000, June). *Transition to the metasystematic stage in Harvard faculty, staff, and administration*. Paper presented at the meeting of the Jean Piaget Society, Montreal, Quebec.

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

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

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

Damon, W. (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 lifespan investigation of developmental and conceptual features of evaluative reasoning about education*. Unpublished doctoral dissertation, University of California at Berkeley, Berkeley, CA.

Dawson, T. L. (2000). Moral reasoning and evaluative reasoning about the good life. *Journal of Applied Measurement*, *1*(4), 372–397.

Dawson, T. L. (2001a). Layers of structure: A comparison of two approaches to developmental assessment. *Genetic Epistemologist*, *29*, 1–10.

Dawson, T. L. (2001b, 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. (2002a). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement*, *3*, 146–189.

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

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

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology*, *164*, 335–364.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development*, *11*, 71–85.

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

Dawson, T. L., & Stein, Z. (2004). *Decision-making curricular development database: Skill map, skill definitions, & activities*. Hatfield, MA: Developmental Testing Service, LLC.

Dawson, T. L., & Wilson, M. (2004). The LAAS: A computerized developmental scoring system for small- and large-scale assessments. *Educational Assessment*, *9*, 153–191.

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.

Dawson-Tunik, T. L. (2004a). "A good education is ..." The development of evaluative thought across the lifespan. *Genetic, Social, and General Psychology Monographs*, *130*(1), 4–112.

Dawson-Tunik, T. L. (2004b, November). The Lectical™ Assessment System. 1. Retrieved December, 2004, from http://www.lectica.info.

Dawson-Tunik, T. L., & Stein, Z. (2004a). *Critical thinking seminar pre and post assessment results*. Hatfield, MA: Developmental Testing Service, LLC.

Dawson-Tunik, T. L., & Stein, Z. (2004b). *National Leadership Study results*. Hatfield, MA: Developmental Testing Service, LLC.

Dawson-Tunik, T. L., & Stein, Z. (2005a). *"It has bounciness inside!" Developing conceptions of energy*. Manuscript submitted for publication.

Dawson-Tunik, T. L., & Stein, Z. (2005b). *It's all good: Moral relativism and the millennium*. Manuscript submitted for publication.

Demetriou, A., & Efklides, A. (Eds.) (1994). *Intelligence, mind, and reasoning: Structure and development*. Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers.

Demetriou, A., & Valanides, N. (1998). A three-level theory of the developing mind: Basic principles and implications for instruction and assessment. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment: Theory into practice* (pp. 211–246). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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.

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., & Ayoub, C. (1994). Affective splitting and dissociation in normal and maltreated children: Developmental pathways for self in relationships. In D. Cicchetti & S. L. Toth (Eds.), *Disorders and dysfunctions of the self: Rochester symposium on developmental psychopathology* (Vol. 5, pp. 149–222). Rochester, NY: University of Rochester Press.

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, NY: Wiley.

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

Fischer, K. W., Knight, C. C., & Van Parys, M. (1993). Analyzing diversity in developmental pathways: Methods and concepts. In R. Case & W. Edelstein (Eds.), *The new structuralism in cognitive development: Theory and research on individual pathways* (Vol. 23, pp. 33–56). Basel: Karger.

Fischer, K. W., & Rose, S. P. (1994). Dynamic development of co-ordination of components in brain and behavior: A framework for theory and research. In G. Dawson & K. W. Fischer (Eds.), *Human behavior and the developing brain* (pp. 3–66). New York, NY: Guilford Press.

Fischer, K. W., & Rose, S. P. (1999). Rulers, clocks, and non-linear dynamics: Measurement and method in developmental research. In G. Savelsbergh, H. van der Maas, & P. van Geert (Eds.), *Nonlinear developmental processes* (pp. 197–212). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Fischer, K. W., & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology*, *36*, 613–648.

Flavell, J. H. (1971). Stage-related properties of cognitive development. *Cognitive Psychology*, *2*, 421–453.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard Educational Review*, *47*, 481–517.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

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.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco, CA: Jossey Bass.

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, NY: 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.

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

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago, IL: Rand McNally.

Larivee, S., Normandeau, S., & Parent, S. (2000). The French connection: Contributions of French-language research in the post-Piagetian era. *Child Development*, *71*(4), 823–839.

Lewis, M. D. (2000). The promise of dynamic systems approaches for an integrated account of human development. *Child Development*, *71*, 36–43.

Linacre, J. M., & Wright, B. D. (2000). *Winsteps*. Chicago, IL: MESA Press.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrica*, *61*, 41–47.

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

Overton, W. F., & Meehan, A. M. (1982). Individual differences in formal operational thought: Sex role and learned helplessness. *Child Development*, *53*(6), 1536–1543.

Overton, W. F., Ward, S. L., Noveck, I. A., & Black, J. (1987). Form and content in the development of deductive reasoning. *Developmental Psychology*, *23*, 22–30.

Pascual-Leone, J., & Goodman, D. (1979). Intelligence and experience: A neoPiagetian approach. *Instructional Science*, *8*, 301–367.

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

Piaget, J. (2000). *Studies in reflecting abstraction* (R. L. Campbell, Trans.). Hove, UK: Psychology Press.

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

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York, NY: Basic Books.

Pratt, M. W., Golding, G., & Hunter, W. J. (1984). Does morality have a gender? Sex, sex role, and moral judgment relationships across the adult lifespan. *Merrill-Palmer Quarterly*, *30*, 321–340.

Rasch, G. (1980). *Probabilistic model for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Rose, S. P., & Fischer, K. W. (1989). *Constructing task sequences: A structured approach to skill theory* (Instructional Manual). Cambridge, MA: Harvard University Press.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York, NY: Oxford University Press.

Smith, L. B., & Thelen, E. (1993). *A dynamic systems approach to development: Applications*. Cambridge, MA: M1T Press.

Smith, R. M. (Ed.) (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Sprinthall, N. A., & Burke, S. M. (1985). Intellectual, interpersonal, and emotional development during childhood. *Journal of Humanistic Education & Development*, *24*, 50–58.

Thomas, H., & Lohaus, A. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, *58*(9), i–v, 1–169.

Turiel, E. (1980). The development of social-conventional and moral concepts. In M. Windmiller, N. Lambert, & E. Turiel (Eds.), *Moral development and socialization* (pp. 69–106). Boston, MA: Allyn & Bacon.

Ullian, D. Z. (1977). The development of conceptions of masculinity and femininity. In B. Lind & J. Archer (Eds.), *Exploring sex differences* (pp. 25–47) New York: Academic Press.

van der Maas, H. L., & Molenaar, P. C. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*, 395–417.

van der Maas, H. L., & Molenaar, P. C. (1995). Catastrophe analysis of discontinuous development. In A. A. van Eye & C. C. Clogg (Eds.), *Categorical variables in developmental research: Methods of analysis* (pp. 77–105). New York, NY: Academic Press.

van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review*, *105*, 634–677.

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

Walker, L. J. (1984). Sex differences in the development of moral reasoning: A critical review. *Child Development*, *55*, 677–691.

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.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational & Psychological Measurement*, *49*(3), 587–602.

Wilson, M. (1984). *A psychometric model of hierarchical development*. Unpublished doctoral dissertation, University of Chicago, Chicago, IL, USA.

Wilson, M. (1985). *Measuring stages of growth: A psychometric model of hierarchical development*. *Occasional paper No. 29*. Hawthorn, Victoria, Australia: Australian Council for Educational Research.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*, 276–289.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wright, B. D. (1999). Model selection: Rating scale or partial credit. *Rasch Measurement Transactions*, *12*(3), 641–642.

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