JOURNAL OF APPLIED MEASUREMENT, *11*(1), 11-23

Concrete, Abstract, Formal, and Systematic Operations as Observed in a "Piagetian" Balance-Beam Task Series

Theo Linda Dawson
Developmental Testing Services, LLC

Eric Andrew Goodheart
Harvard University

Karen Draney
Mark Wilson
University of California at Berkeley

Michael Lamport Commons
Harvard Medical School

We performed a Rasch analysis of cross-sectional developmental data gathered from children and adults who were presented with a task series derived from Inhelder's and Piaget's balance beam. The partial credit model situates both participants and items along a single hierarchically ordered dimension. As the Model of Hierarchical Complexity predicted, order of hierarchical complexity accurately predicted item difficulty, with notable exceptions at the formal and systematic levels. Gappiness between items was examined using the saltus model. A two level saltus model, which examined the gap between the concrete/abstract and formal/systematic items, was a better predictor of performance than the Rasch analysis ($\chi^2 = 71.91$, $df = 4$, $p < .01$).

Requests for reprints should be sent to Theo L. Dawson, 35 South Park Terrace, Northampton, MA 01060, USA, e-mail: Theo@DevTestService.com.

Overview

The existence of hierarchical stage sequences is a controversial issue (Brainerd, 1973; Broughton, 1984). This controversy cannot be resolved in the absence of task sequences that reliably and accurately reflect specific difficulty levels. In this study, we show that Commons' (Commons, Goodheart, and Bresette, 1995) Balance-Beam instrument, which was designed specifically to assess hierarchical complexity, elicits developmentally sequenced and hierarchically organized behavior. In doing this we demonstrate how Rasch modeling (Rasch, 1980) and saltus analysis (Wilson, 1989) can be applied to the problem of demonstrating that a constructed developmental task series actually measures what it is purported to measure.

Theoretical Framework

The Model of Hierarchical Complexity (MHC) of Commons and his colleagues (Commons, Trudeau, Stein, Richards, and Krause, 1998) is a generalized developmental model that classifies development in terms of the hierarchical organization of ideal actions. The model addresses two conceptually different but related issues. The first issue is the analytic notion of hierarchical complexity. Hierarchical complexity refers to the number of recursions that coordinating actions must perform on a set of primary elements. Actions at a higher order of hierarchical complexity: a) are defined in terms of the actions at the next lower order of hierarchical complexity; b) organize and transform the lower order actions; c) produce organizations of lower order actions that are new, non-arbitrary, and cannot be accomplished by those lower order actions alone.

The second issue is the empirical question of how hierarchical complexity of performance on tasks develops. For the instrument employed in this study—the Balance Beam instrument (Commons et al., 1995)—several tasks have been analytically constructed for each of 6 complexity orders. An individual's hierarchical complexity score on this task series is based on the order of hierarchical complexity of the most complex tasks that are accurately completed. Hierarchical complexity scores are not intended to be measures of competence. They are measures of performance.

Tasks have many features in addition to their hierarchical complexity, many of which affect performance (Fischer and Bidell, 1998). In order to minimize the "noise" from features other than hierarchical complexity, the balance beam problems in our instrument were constructed with as little variation in form and content as possible. In other words, in constructing items for the different complexity orders, we attempted to manipulate only the hierarchical complexity dimension.

If the tasks in an instrument represent a single latent dimension, differing only in their difficulties, then we would expect the data to fit the Rasch model. However, if the tasks in an instrument represent successive orders of hierarchical complexity, and solving tasks of different orders of hierarchical complexity requires qualitatively distinct forms of reasoning, then both item difficulties and person estimates will exhibit relatively rigid, Guttman-like (Guttman, 1950) orderings (Fischer, Knight, and Van Parys, 1993). Moreover, Rasch analysis will reveal clusters of items, representing the different complexity orders, that are separated by statistically significant gaps, and the distribution of person estimates may appear 'toothy' with clusters of person estimates systematically associated with clusters of item difficulties (Dawson, 1998; Dawson-Tunik, Commons, Wilson, and Fisher, 2005). Certain patterns of misfit to the Rasch model are also likely to be observed (Mislevy and Wilson, 1996; Wilson, 1989). The saltus model (Wilson, 1989 #3520), makes it possible to test whether or not these patterns are indicative of discontinuities in development that are consistent with developmental stage theory.

We hypothesize (1) that the ability to solve balance beam problems will emerge in a Rasch analysis as a single latent trait or dimension of performance with item difficulties ordered according to complexity order; (2) that items of different orders of hierarchical complexity will cluster in groups along this dimension and that misfit will be associated with this clustering; and (3) that a saltus analysis will demonstrate that these clusters are indicative of a pattern of performance that supports the notion of developmental stage.

Method

Convenience sampling was used to gather 121 predominantly Caucasian, middle class participants (79 female, 37 male), whose ages ranged from 7 to 56 years ($M = 29.2$, $SD = 12.98$). Participants were solicited from New England public schools and colleges. Participation was voluntary. Participants were given unlimited time to complete a multiple choice pen and paper instrument containing primary, concrete, abstract, formal, systematic, and metasystematic balance beam problems. Only the concrete, abstract, formal, and systematic balance beam problems are analyzed here.

Instrument

The Balance Beam Series is a pen and paper instrument that consists of sets of multiple choice problems of increasing hierarchical complexity. The tasks form a series because every higher order task has the lower order task of the previous complexity order embedded within it. In the present series, there are 5 concrete order items, 5 abstract order items, 4 formal order items, and 4 systematic order items. These items are presented in 4 sections. A set of instructions and an example problem are provided at the beginning of each section. Figure 1 gives examples of items at each level.

Concrete Balance Beam. The concrete Balance Beam task requires the coordination of two operations, as shown in Figure 1: First, the total weight on each side of the beam must be equal. On one side of the fulcrum, there is a single weight. This weight must be balanced by a stated weight plus some unknown additional weight. The participant determines the unknown additional weight by subtracting the stated amount of weight (1) from the total (3).

Abstract Balance Beam. In the abstract Balance Beam tasks (Figure 1, second example), the distances between the weights and the fulcrum are no longer held constant, as they were on the concrete Beam. Therefore, when determining the amount of weight necessary to balance a given weight on the opposite side of the beam, the participant must consider both weight and distance. We created tasks with the following constraint. The distance between the fulcrum and the weight on one side of the beam is equal to the weight on the other side of the beam. Similarly, the weight on one side of the beam is equal to the distance between the fulcrum and the weight on the other side of the beam. Because of this constraint, torque can be calculated either by summing weight and distance or by taking their product:

The task requires the participant to solve for a single unknown variable. But because the participants performing at this complexity order have not acquired the true rule for calculating torque (the product of weight and distance), they will only consistently solve balance beam problems built with the aforementioned constraint. Applying the additive rule to beams that have not been built with this constraint (i.e., formal Balance Beams) produces incorrect solutions.

Formal Balance Beam. In Inhelder and Piaget's formal balance beam (1958), the task was to make the beam balance by moving the weights toward or away from the fulcrum and by increasing or decreasing the size of the weights. The beam is in balance when the torques on each side of the fulcrum are equal in magnitude, where torque is equal to the amount of weight times its distance from the fulcrum. The formal Balance Beam tasks in the present instrument (Figure 1, third example) require the participant to use the multiplicative rule to find an unknown weight or distance that will make the beam balance.

Systematic Balance Beam. A systematic Balance Beam task presents two formal balance beams to the participant. The same two unknown quantities (weights or distances) must be determined for both beams. In order to find the unknowns, the participant is required to solve a set of simultaneous equations. The

participant must relate two formal operational equations to solve these simultaneous equations.

Table 1 shows some anticipated response patterns, according to the orders of hierarchical complexity of the tasks at each complexity order. Not all respondents are expected to conform to one of these response patterns; they can be considered prototypical. Some subjects may be able to correctly answer some, but not all of the problems at the complexity order above which they can answer all problems correctly. Additionally, it is to be expected that some respondents will make isolated errors as a consequence of factors unrelated to their overall stage of performance, such as flaws in the instrument design and variations resulting from the method of administration. Further, because this is a multiple-choice instrument, guessing is possible and should cause a certain amount of random variation. Nevertheless, in a Rasch analysis, if the overall response patterns are similar to the patterns in Table 1, groups of item estimates from each complexity order should cluster together with gaps between these groups.

In the following Rasch analysis we examine the overall pattern of item difficulties and person abilities. Then, saltus analysis (Wilson, 1989) is

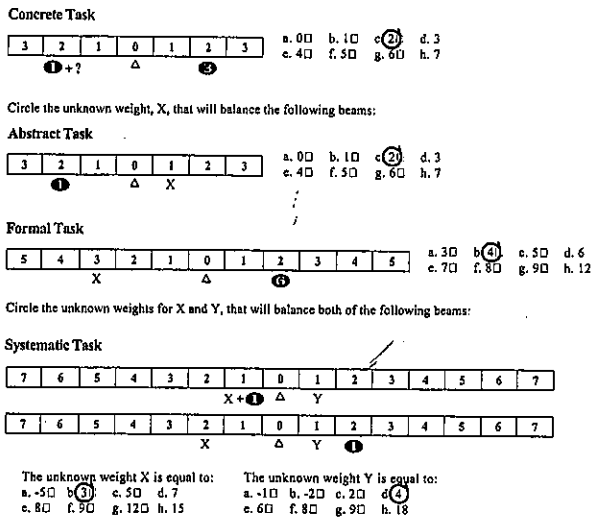

Figure 1. Examples of balance-beam tasks for each complexity order.

Table 1

Expected Performance Patterns

	Items			
Performances	Concrete	Abstract	Formal	Systematic
Preconcrete	00000	00000	0000	0000
Concrete	11111	00000	0000	0000
Abstract	11111	11111	0000	0000
Formal	11111	11111	1111	0000
Systematic	11111	11111	1111	1111

used to investigate whether the expected response patterns occur with enough regularity to suggest a step-like discontinuity between successive orders of hierarchical integration.

A saltus analysis is appropriate here because it determines whether the difficulty of a group of items is significantly different for groups of persons who have different ability estimates. These differences are called *second order discontinuities.* Saltus analysis is related to the family of logistic models (Spada and McGraw, 1985), but is distinguished from members of that family by the inclusion of a latent group membership parameter. For similar examples of the use of saltus, see (Draney and Wilson, 2005; Draney and Wilson, 2007; Draney, 1996; Fieuws, Spiessens, and Draney, 2004; Mislevy and Wilson, 1996; Wilson, 1989).

Analysis

Initially, a Rasch model of the data was estimated with the software, Quest (Adams and Khoo, 1993). Overall, the results of this analysis support the claim that the Balance Beam Task Series measures a single dimension of performance. Reliability of the item estimates is .97,with a mean infit mean square of .94 (*SD* = .13). Standard errors of the item estimates range from .28 to 1.03 (See Table 2). All of the items have infit meansquares that are close enough to the expected value of 1 to fit the model. However, the outfit meansquares are less consistent, with formal item 4 and systematic item 1 exhibiting positive misfit, indicating that performance on these items is more erratic than expected. This problem is discussed further below. Reliability of case estimates is .77, with a mean infit mean square of .95 (SD .64). This low reliability for the person estimates reflects the small number of items in the Balance-beam task series.

The distribution of person and item estimates is shown in Figure 2. The 25 participants who had perfect scores are not included. Performances with positive infit *ts* over 2.00 are indicated with *Y*, while performances with negative infit *ts* below –2.00 are indicated with *Z*. Patterns of fit will be discussed further below. Item difficulties, with two exceptions—systematic 1 and formal 4—conform to the anticipated hierarchy. As expected, difficulty estimates for the abstract and formal order items are clearly differentiated, though, contrary to expectations, there is neither a clear differentiation between the concrete and abstract items (likely due to the small number of individuals performing at the concrete stage), nor the formal and systematic items.

Person fit is not as good as item fit. 21 out of 120, or 17.5% of the performances fail to fit the Rasch model. Thirteen (10.8%) of these exhibit negative misfit. All 13 share the specific pattern of response shown in Table 3, case 132. Interestingly, this type of performance is entirely consistent with Commons' theorized response pattern. The fact that cases consistent with the theory misfit the model is strong evidence that the pattern of performance is not adequately modeled by the Rasch estimates.

Cases with positive misfit comprise 6.7% of the sample. In Table 3, Case 99, for example, is at the same level of performance (.56) as case 132, but in this instance, the respondent has correctly answered some formal and systematic items while missing several abstract items. This pattern of response clearly violates performance expectations.

Case 96, also at .56 logits, fits within the parameters of the present model. In this instance, the respondent deviates from the expected pattern of performance by correctly answering formal item 3, while incorrectly answering abstract item 3. Because the Rasch model is probabilistic, a range of variation from the expected pattern can occur without causing misfit.

Performances at the juncture of the formal and systematic levels are more problematic than those at the juncture of the abstract and formal levels. Although three instances of the anticipated formal performance occur, as in case 076 in Table 3, there are also two instances of the pattern found in case 015. (This pattern is problematic, and will be discussed further below.) Most performances around this level of difficulty have a mixture of correct formal and correct systematic answers,

Table 2

Fit statistics for Balance Beam items

ITEM	NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	Concrete 1	94	120	-4.01 1.03	.77	.06	0.0	.6
2	Concrete 2	93	120	-3.29 .75	.81	.15	-.1	.2
3	Concrete 3	93	120	-3.29 .75	.88	.37	0.0	.4
4	Concrete 4	90	120	-2.24 .51	.88	3.29	-.2	1.5
5	Concrete 5	89	120	-2.02 .47	.95	1.02	0.0	.4
6	Abstract 1	89	120	-2.02 .47	.88	.77	-.2	.2
7	Abstract 2	81	120	-.87 .34	1.16	1.63	.8	1.0
8	Abstract 3	85	120	-1.35 .38	1.01	2.75	.1	1.7
9	Abstract 4	82	120	-.98 .35	1.13	2.21	.6	1.5
10	Abstract 5	78	120	-.57 .31	.97	1.67	-.1	1.2
11	Formal 1	33	120	2.38 .26	1.02	1.33	.2	1.0
12	Formal 2	32	120	2.44 .26	.81	.69	-1.5	-.9
13	Formal 3	31	120	2.51 .27	.85	.82	-1.1	-.4
14	Formal 4	22	120	3.16 .29	.82	4.99	-1.1	4.6
15	Systematic 1	44	120	1.70 .25	1.05	1.94	.5	2.9
16	Systematic 2	28	120	2.71 .27	.78	.56	-1.7	-1.2
17	Systematic 3	26	120	2.86 .28	1.09	.79	.7	-.4
18	Systematic 4	26	120	2.86 .28	1.12	1.12	.9	.4
Mean				0.00	.94	1.45	-.1	.8
SD				2.53	.13	1.24	.8	1.4

N = 120, L = 18, Probability Level = .50

as in case 44 (Table 3). In light of the mixture of patterns, it is not surprising that the formal and systematic items are not clearly differentiated in Figure 2.

Saltus Analysis

To study the gappiness in our data, we performed a saltus analysis (Draney and Wilson, 2005; Draney and Wilson, in press; Draney, 1996; Fieuws et al., 2004; Mislevy and Wilson, 1996; Wilson, 1989). A saltus analysis is a mixture model extension of the Rasch model (Mislevy and Verhelst, 1990; Rost, 1990). Whereas a traditional partial credit analysis determines the probability of a given subject performing a given item in terms of item difficulty (delta) and subject ability (beta), a saltus analysis introduces item and subject stage as a third concept. This additional concept (along with the parameter{s} that embody it) can help to determine whether the gappiness and systematic shifts in item misfit present in the earlier two parameter Rasch analysis can be explained as stage change.

Table 3

Sample Performances and Fit Statistics

Case	Fit	Ability	Infit *t*	Concrete	Abstract	Formal	Systematic
132	.17	.56	−2.09	11111	11111	0000	0000
099	2.79	.56	2.27	11111	10010	1000	1001
096	1.16	.56	.45	11111	11011	0010	0000
076	1.07	2.66	1.07	11111	11111	1111	0000
015	0.99	2.66	.04	11111	11111	0000	1111
044	.76	2.66	−1.18	11111	11111	1100	1000

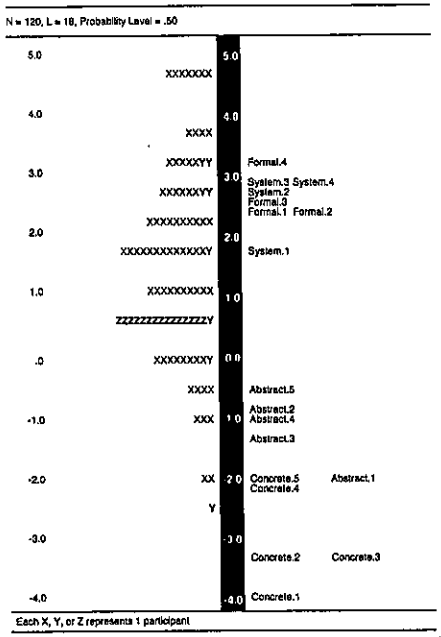

Figure 2. Item and case estimates for balance-beam data.

The saltus model is based on the assumption that there are H developmental stages in the population of interest. A different set of items represents each one of these stages, such that only persons at or above a stage are fully equipped to answer the items associated with that stage correctly. The saltus model assumes that all persons in stage h answer all items in a manner consistent with membership in that stage. However, persons within a stage may differ by proficiency.

To describe the model, suppose that, as in the partial credit model (Masters, 1982), the random variable $\mathbf{X}_{ni}$ indicates the response to item i. Items have $J_i + 1$ possible response alternatives indexed

$j = 0,1,\ldots,J_i$. The parameter indicating step j for item i will be indicated by β_{ij}; the vector of all β_{ij} by β.

In the saltus model, a person is characterized by a proficiency parameter θ_n and an indicator vector for stage membership ϕ_n. If there are H potential stages, $\phi_n = (\phi_{n1}, \ldots, \phi_{nH})$, where ϕ_{nh} takes the value of 1 if the examinee n is in stage h and 0 if not. Only one of the ϕ_{nh} is theoretically nonzero. As with θ_n, values of ϕ_n are not observable.

Just as persons are associated with one and only one stage, items are associated with one and only one stage. Unlike person stage membership, however, which is unknown and must be estimated, item stage is known a priori, based on the theory that was used to produce the items. It will be useful to denote item stage membership by the indicator vector $\mathbf{b}_i$. As with ϕ_n, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iH})$, where b_{ik} takes the value of 1 if item i belongs to item stage k, and 0 otherwise. The set of all $\mathbf{b}_i$ across all items is denoted by $\mathbf{b}$.

The equation:

$$P(X_{nij} = j \mid \theta_n, \phi_{nh} = 1, \beta_i, \tau_{hk}) = \frac{\exp\sum_{s=0}^{j}(\theta_n - \beta_{is} + \tau_{hk})}{\sum_{t=0}^{J_i}\exp\sum_{s=0}^{t}(\theta_n - \beta_{is} + \tau_{hk})}, \quad (1)$$

indicates the probability of response j to item i. The saltus parameter τ_{hk} describes the additive effect—positive or negative—for people in stage h on the item parameters of all items in stage k In developmental context, this often takes the form of an increase in probability of success as the person achieves the stage at which an item is located, indicated by $\tau_{hk} > 0$ when $h \geq k$ (although this need not be the case). The saltus parameters can be represented together as an H by H matrix $\mathbf{T}$.

The probability that an examinee with stage membership parameter ϕ_n and proficiency θ_n will respond in category j to item i is given by:

$$P(X_{nij}=j|\theta_n, \mathbf{f}_n, \mathbf{b}_i, \mathbf{B}_i, \mathbf{T}) = \prod_h \prod_k P(X_{nij} = j \mid \theta_n, \phi_{nh} = 1, \mathbf{b}_i, \tau_{hk})^{\phi_{nh} b_{ik}}. \quad (2)$$

Assuming conditional independence, the modeled probability of a response vector is:

$$P(\mathbf{X}_n = \mathbf{X}_n \mid \theta_n, \mathbf{f}_n, \mathbf{b}_i, \mathbf{B}_i, \mathbf{T}) = \prod_h \prod_k \prod_i P(X_{nij} = x_{ij} \mid \theta_n, \phi_{nh} = 1, \mathbf{b}_i, \tau_{hk})^{\phi_{nh} b_{ik}}. \quad (3)$$

The model requires a number of constraints on the parameters. For item step parameters, we use two traditional constraints: first, $\beta_{i0} = 0$ for every item, and second, the sum of all the β_{ij} is set equal to zero. Some constraints are also necessary on the saltus parameters. The set of constraints we have chosen is the same as that used by Mislevy and Wilson (1996), and will allow us to interpret the saltus parameters as changes relative to the first (lowest) developmental stage. Two sets of constraints are used. First $\tau_{h1} = 0$; thus, the difficulty of the first stage of items is held constant for all person groups; changes in the difficulty of items representing higher stages are interpreted with respect to this first stage of items for all person stages. Also $\tau_{1k} = 0$; thus, items as seen by person stages higher than 1 will be interpreted relative to the difficulty of the items as seen by persons in the lowest developmental stage.

As in Mislevy and Wilson (1996), the EM algorithm (Dempster, Laird, and Rubin, 1977) is used to estimate the structural parameters for the model. Empirical Bayes estimation is then used to obtain estimates of the probabilities of stage membership for each subject, as well as proficiency estimates given membership in each stage. A person is said to be classified into the stage for which that person's probability of membership is highest.

The Model of Hierarchical Complexity makes no predictions about gappiness. It does predict that when items are of identical hierarchical complexity, differences in performance will be explained by the presence of differences among items in horizontal complexity: prior exposure, training, etc. It also predicts that when horizontal complexity is held constant among items of identical hierarchical complexity, differences in performances will be explained by differences in hierarchical complexity. In such controlled circumstances, hierarchical complexity should predict the order in which proficiency is demonstrated.

The Commons stage-transition model does predict that toward the end of a transition from one complexity order to the next, developmental change occurs more rapidly. Changes in performance that occur over time are by definition dynamic. However, in a cross-sectional study, they may appear as gaps (i.e., Second order discontinuities) in a static representation of item difficulty. The saltus model provides a tool for studying these discontinuities.

In order to perform a saltus analysis, groups or *classes* of performance are specified. Participants are grouped by their performances on the items of each class. For example in a two class Saltus analysis, participants who perform successfully on items of the first class but fail items of the second class are grouped in the first class. Participants who perform items of both classes are grouped in the class associated with the greater difficulty. Like the Model of Hierarchical Complexity, the saltus model demands that those who perform items of the more difficult class also master items of the less difficult class. However, assignment of participants to a class is probabilistic. Performances that violate the assumptions of the Saltus model will have a significant probability of belonging to either class.

In the present case, we conducted a series of pairwise analyses in which we examined performance on concrete versus abstract, abstract versus formal, and formal versus systematic classes. These were followed by an analysis in which we specified a concrete/abstract class, contrasting performance in this class with performance in the formal class, and an analysis in which we specified the same concrete/abstract class, contrasting performance in this class with a formal/systematic class.

The results of the concrete versus abstract analysis did not support the notion of a stage shift or second order discontinuity between these two complexity orders. Because only three participants performed at the concrete level, it is unlikely that this result says very much about the balance beam instrument. Note the large error ranges for items 1-5, shown in Table 2.

The pair-wise analysis of abstract versus formal classes is more revealing. Figure 3 shows the item difficulties for these classes. Note the large gap between the difficulty estimates for formal and abstract items in Class 1, and its complete disappearance in Class 2. Clearly, the formal items are much harder, relative to abstract items, for

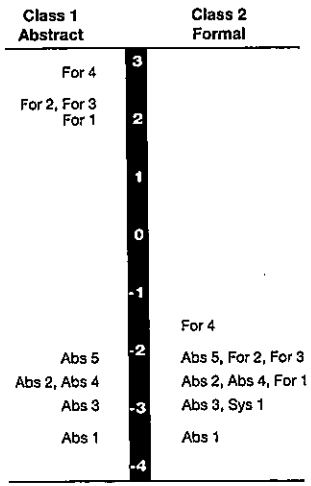

Figure 3. Difficulty estimates by class for abstract vs. formal classes.

members of Class 1 than for members of Class 2. This result supports the hypothesis that a second-order discontinuity occurs at the juncture of the abstract and formal complexity orders.

Figure 4 shows the pair-wise analysis of formal versus systematic classes. There is no gap between formal and systematic items in Class 1. In fact, item difficulties for Class 1 and Class 2 are almost identical. Further, the formal items are more difficult for both groups than the systematic items. Possible reasons for this reversal are discussed further below.

Because we did not find evidence of a second order discontinuity between the concrete and abstract orders, we combined the concrete and abstract classes into a single concrete/abstract class for the remaining analyses. Figure 5 shows a pair-wise analysis of concrete/abstract versus formal classes. As in the second analysis, a large gap is apparent between concrete/abstract and formal items in Class 1. Once again, this gap disappears in Class 2. The formal items are much harder, relative to concrete and abstract items, for members of Class 1 than for members of Class 2. This result lends additional support to the hypothesis that a second-order discontinuity occurs at the juncture of the abstract and formal orders.

Because we did not find evidence of a second order discontinuity between the formal and systematic complexity orders, we combined the formal and systematic classes into a single formal/systematic class for the final analysis. Figure 6

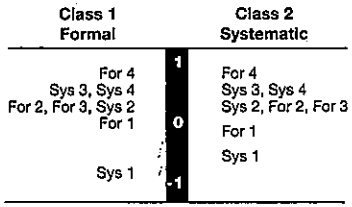

Figure 4. Difficulty estimates by class for formal vs. systematic classes.

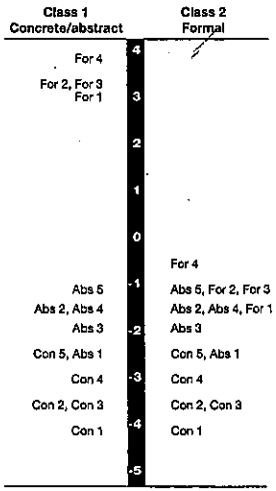

Figure 5. Difficulty estimates by class for concrete/abstract vs. formal classes.

shows the results of this pair-wise analysis of concrete/abstract versus formal/systematic classes. A large gap is apparent between concrete/abstract and formal/systematic items in Class 1, though the gap in this analysis is somewhat smaller than in the second and fourth analyses, due to the low difficulty of systematic item 1. Once again, this gap disappears in Class 2. The formal/systematic items are much harder, relative to concrete and abstract items, for members of Class 1 than for members of Class 2. This result lends additional support to the hypothesis that a second-order discontinuity occurs at the juncture of the abstract complexity order and the complexity order that follows. However, it appears from this result that the systematic items are easier than the formal items, rendering the precise nature of the transition unclear. The likelihood ratio for this saltus analysis was compared with the likelihood ratio for the Rasch analysis. The saltus model predicts performance on the balance beam task with greater accuracy than the Rasch model (χ^2= 71.91, $df = 4$, $p < .01$).

Discussion

The Model of Hierarchical Complexity predicts that performance on items of a given hierarchical complexity will be consistent within individuals. In other words, proficiency at performing on an item of a given order of hierarchical complexity increases the probability of performing other items of the same order of complexity.

Further, the model predicts that individuals who are proficient at solving problems of a particular order of hierarchical complexity will also be proficient at solving problems of lower orders of hierarchical complexity. They should also have a lower probability of performing accurately on items whose order of hierarchical complexity is beyond the complexity order at which they generally perform. In the Rasch analysis, these probabilities were expressed in terms of the hierarchy of items and persons on the items by persons map. With the exception of two items, systematic 1 and formal 4, item difficulties were ordered as expected, though there was not as much differentiation between the concrete and abstract items and the formal and systematic items as we expected, based on the Model of Hierarchical Complexity. The lack of differentiation between the concrete and abstract problems could be associated with the small number of participants who performed at the concrete level. Note the large error ranges for items 1-5.

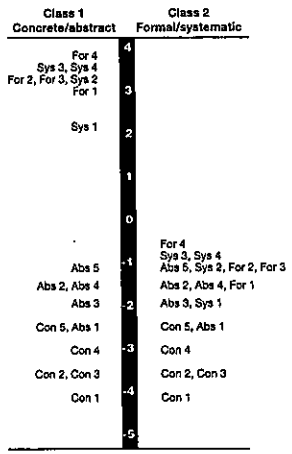

Figure 6. Difficulty estimates by class for concrete/abstract vs. formal/systematic classes.

The lack of differentiation between the formal and systematic items is more problematic for the Balance-Beam instrument and Commons' theory. Our first concern was the low difficulty estimate for systematic item 1 (shown in Figure 1). In order to understand why this item might be easier than all of the formal and the remaining systematic items, we compared its form and content with the other systematic items. We discovered that this item could be solved by coordinating an abstract, addition based equation with a formal, multiplication-based equation, while the other systematic items required the coordination of two formal, multiplication based, equations. This would explain why the item is easier than the other systematic items, and it may also explain why it is easier than most of the formal items. If a formal operation is the coordination of abstract operations, then systematic item 1 could well be an easier formal item than the intentionally formal items on this instrument, because it provides more support. In effect, the need for coordination is more explicit when there are two balance beam problems to solve than when there is only one. Whereas the need for coordination can be overlooked in the case of the formal items, it is made explicit in the systematic items. This problem is compounded in the Balance-Beam instrument, because the rule for balancing the beam[1] is explicitly provided prior to the presentation of the systematic items. This rule is not provided prior to the formal items. It seems likely that some able individuals carelessly applied the abstract order additive rule to the formal balance beams, but switched to the multiplicative rule once it was described.

The results of the saltus analysis, though they support one of the hypothesized second-order discontinuities between stages, also magnify the failure of the formal and systematic items to accurately measure what they are intended to measure. When Class is included as a dimension, difficulties of the formal and systematic items are even less differentiated than in the Rasch analysis. If this pattern is the result of characteristics of the instrument rather than fundamental flaws in the theory, then it is clear that some instrument redesign is necessary.

We suggest the following. (1) Increase the number of items within each complexity order in order to increase the reliability of person estimates and saltus parameters. When the number of items within a saltus class is low (as it is in the present study), the reliability of the saltus parameters is also diminished. (2) Consider eliminating the multiple-choice format. Problems at the highest levels can be solved using operations of a lower order of hierarchical complexity by substituting the various multiple-choice responses until the correct response is found. This makes it very difficult to know what abilities the items are actually measuring. Further, the multiple-choice format permits guessing, adding "noise" that is difficult to interpret. (3) Redesign systematic item 1 to make it impossible to solve it with an addition strategy.

The construction of items that meet strict criteria for hierarchical complexity is a challenging task. Commons' effort to create items with task demands appropriate to specific complexity orders appear to have been more successful at the lower levels of hierarchical complexity than at the higher levels. The present analysis has provided useful information that can guide ongoing efforts in this area.

1 The diagrams in Figure 1 represent balance beams. The torque(s) on the left side of the balance beam balance the torque(s) on the right side, where torque is equal to weight times distance from the fulcrum (i.e., distance from 0).

References

Adams, R. J., and Khoo, S.-T. (1993). *Quest: The interactive test analysis system*. Camberwell, VIC, Australia: Australian Council for Educational Research.

Brainerd, C. J. (1973). Neo-Piagetian training experiments revisited: Is there any support for the cognitive-developmental stage hypothesis? *Cognition, 2*, 349-370.

Broughton, J. (1984). Not beyond formal operations but beyond Piaget. In M. L. Commons, F. A. Richards and C. A. Armon (Eds.), *Beyond formal operations: Late adolescent and adult development* (Vol. 1, pp. 395-411). New York: Praeger.

Commons, M. L., Goodheart, E. A., and Bresette, L. M. (1995). Formal, systematic, and metasystematic operations with a balance-beam task series: A reply to Kallio's claim of no distinct systematic stage. *Journal of Adult Development, 2*, 193-199.

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

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

Dawson-Tunik, T. L., Commons, M. L., Wilson, M., and Fisher, K. W. (2005). The shape of development. *The European Journal of Developmental Psychology, 2*(2), 163-196.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Draney, K., and Wilson, M. (2005). Application of the polytomous saltus model to stage-like proportional reasoning. In L. A. v. d. Ark, M. A. Croon, and K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 207-226). Mahwah, NJ: Lawrence Erlbaum.

Draney, K., and Wilson, M. (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. v. Davier and C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 119-130). New York: Springer.

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.

Fieuws, S., Spiessens, B., and Draney, K. (2004). Mixture models. In P. De Boeck and M. Wilson (Eds.). *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 317-340). New York: Springer.

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

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer and e. al. (Eds.), *Studies in social psychology in World War II* (Vol. 4, pp. 60-90). New York: Wiley.

Inhelder, B., and Piaget, J. (1958). *The growth of logical thinking forom childhood to adolescence*. New York: Basic Books.

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

Mislevy, R. J., and Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rost, J. (1990). Rasch models in latent class analysis: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282

Spada, H., and McGraw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and paychometrics* (pp. 169-194). Orlando, FL: Academic Press.

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