CHAPTER 8

MIXTURE MODELS IN A DEVELOPMENTAL CONTEXT

Karen Draney and Mark Wilson
University of California, Berkeley

Judith Glück and Christiane Spiel
Institut für Psychologie der Universität Wien

Mixture item response theory (IRT) models are based on the assumption that the population being measured is composed of two or more latent subpopulations, each of which responds to a set of tasks in predictably different ways. Within each subpopulation, a latent trait model holds for the entire set of tasks; however, between the subpopulations, there are differences that cannot be described within the constraints of the latent trait model used for a given subpopulation.

One of the most general mixture IRT models is the mixed Rasch model (Rost, 1990). This model assumes that the population in question is made up of H subpopulations, and that a Rasch model holds within each subpopulation. There is no necessary relation among the various Rasch models; the ordering of the items in terms of their difficulty can be entirely different for each subpopulation. This model is exploratory in the sense that it simply divides the population into the "best" (i.e., most different) set of subpopulations. The user must then determine what is interesting about the differences among subpopulations.

Advances in Latent Variable Mixture Models, pages 199–216

Other mixture item response models are more confirmatory in nature. One example is given by Mislevy and Verhelst (1990). This model is an extension of the Linear Logistic Test Model (LLTM; Fischer, 1983). It posits a particular structure for item difficulty parameters within each subpopulation, based on characteristics of the tasks. A different LLTM may hold for each subpopulation, if each set of persons responds differently to the task characteristics, or to a different set of task characteristics. Such a model may even include a subpopulation of random guessers, whose probability of correctly answering multiple choice items is a simple function of the number of choices. This model is a special case of Rost's mixed Rasch model described above.

The *saltus* model (*saltus* is Latin for "leap") (Draney, 1996; Wilson, 1989) is another type of confirmatory mixture IRT model. It was originally designed for the investigation of developmental stages. This model is a special case of both of the preceding models, with linear restrictions on the relations among sets of item difficulties for the different subpopulations. This model was developed as a method for detecting and analyzing discontinuities in performance that are hypothesized to occur as a result of rapidly occurring person growth (e.g., Fischer, Pipp, & Bullock, 1984). Such discontinuities are often theorized to occur as the result of progression through developmental stages or levels. Thus, the model is built upon the assumption that the subpopulations are ordered in some way (as are developmental stages in children), and that groups of items become predictably easier (or perhaps more difficult) for subpopulations further along the developmental continuum.

One of the most influential developmental theories was posited by Jean Piaget (e.g., Piaget, 1950; Inhelder & Piaget, 1958). The work of Piaget describes the cognitive developmental stages through which children progress as they grow. In particular, school-age children progress from the *preoperational* stage, through the *concrete operational* stage, to the *formal operational* stage. In the preoperational stage, children are able for the first time to produce mental representations of objects and events, but unable to consistently perform logical mental operations with these representations. In the concrete operational stage, children are able to perform logical operations, but only on representations of concrete objects. In the formal operational stage, which starts to occur around the beginning of adolescence, children are able to perform abstract operations on abstractions as well as concrete objects. According to Piaget, progress from stage to stage is characterized by more than simple linear growth in reasoning ability. The transition from one stage to another involves a major reorganization of the thinking processes used by children to solve various sorts of problems.

Theories with similar structure, but perhaps different substantive focus, are described by the many neo-Piagetian researchers, and by other edu-

cational and psychological researchers who use stage-based theories. For example, Siegler (1981) used the work of Piaget to develop sets of items regarding which side of a balance scale would go down, when one placed different combinations of weights and distances from the fulcrum on the two sides of the balance scale. These sets of items changed predictably in difficulty for different age groups of children, as the children progressed through Piagetian-based stages. Some groups of items became easier and some more difficult, while others remained the same. The developmental stages of the children thus resulted in relative shifts in the probability that certain groups of items would be answered correctly. The saltus model is suitable for use with such sets of items (see Wilson, 1989; and Draney, 1996). A more general mixture IRT model, such as the mixed Rasch model, would require the estimation of a difficulty parameter for each item within each developmental stage (if the items are dichotomous); the saltus model can accommodate many developmental theories by estimating one difficulty for each item, plus a small number of additional parameters to describe the changes associated with developmental stage.

Although Piagetian theory has been somewhat controversial of late (e.g., Lourenço & Machado, 1996), there is still a strong interest in stage-like development in a number of areas, including moral and ethical reasoning (e.g., Dawson, 2002; Kohlberg & Candee, 1984), evaluative reasoning (e.g., Armon, 1984; Dawson-Tunik 2004), adult development (e.g., Commons, Trudeau, Stein, Richards, & Krause, 1998; Fischer, Hand, & Russel, 1984), and cognitive development (e.g., Bond, 1995a, 1995b; Bond & Bunting, 1995; Demetriou & Efklides 1989, 1994; van Hiele, 1986).

Researchers in the Piagetian tradition are using increasingly complex statistical and psychometric models to analyze their data. For example, Béland and Mislevy (1996) analyzed proportional reasoning tasks using Bayesian inference networks. Noelting, Coudé, and Rousseau (1995) discussed the advantages of Rasch scaling for the understanding of Piagetian tasks. And Bond (1995a; 1995b) discussed the implications of Rasch-family models for Piagetian theory and philosophy. In addition, psychometric researchers have begun wrestling with the problem of developing and applying models with sufficient complexity to address specific substantive issues. For example, the three-parameter model has been used diagnostically by researchers such as Yen (1985), who described patterns of problematic item fit that are sometimes observed when analyzing complex data. She asserted that these may be indicators for increasing item complexity, which could potentially be indicative of a set of items that represent more than one developmental stage.

The current chapter will discuss the basic structure and parameterization of the saltus model, and then give an example of its use, comparing it with a prior analysis conducted by the authors who collected the original set of data, in which a more exploratory mixture model was fitted. The benefits of

the saltus model, including the ability to quantify the magnitude of group membership effects on specific collections of items, will be examined, and the fit of the two models compared.

THE SALTUS MODEL

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 in proficiency. In a Piagetian context, this means that a child in, say, the concrete operational stage is always in that stage, and answers all items accordingly. The child does not show formal operational development for some items and concrete operational development for others. However, some concrete operational children may be more proficient at answering items than are other concrete operational children.

To describe the model, suppose that, as in the partial credit model (Masters, 1982), the random variable $\mathbf{X}_{ni}$ indicates the nth person's response to item i. Items have $J_i + 1$ possible response alternatives indexed $j = 0,1,\ldots,J_i$. The difference in difficulty between any two consecutive item levels is referred to as a step, as in Masters' representation of the model. The parameter indicating step j for item i will be indicated by β_{ij}; the vector of all β_{ij} by $\boldsymbol{\beta}$.

In the saltus model, a person is characterized by a proficiency parameter θ_n and an indicator vector for stage membership $\boldsymbol{\phi}_n$. If there are H potential stages, $\boldsymbol{\phi}_n = (\phi_{n1},\ldots,\phi_{nH})$, where ϕ_{nh} takes the value of 1 if person n is in stage h and 0 if not. Only one of the ϕ_{nh} is theoretically nonzero. As with θ_n, values of $\boldsymbol{\phi}_n$ 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 $\boldsymbol{\phi}_n$, $\mathbf{b}_i = (b_{i1},\ldots,b_{iH})$, 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:

$$P\left(X_{nij} = j \mid \theta_n, \phi_{nh} = 1, \boldsymbol{\beta}_i, \tau_{hk}\right) = \frac{\exp \sum_{s=0}^{j}\left(\theta_n - \beta_{is} + \tau_{hk}\right)}{\sum_{t=0}^{J_i} \exp \sum_{s=0}^{t}\left(\theta_n - \beta_{is} + \tau_{hk}\right)}, \qquad (8.1)$$

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 a 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 $H \times H$ 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:

$$P\left(X_{nij}=j|\theta_n,\phi_n,\beta_i,\mathbf{b}_i,\mathrm{T}\right)=\prod_h\prod_k P(X_{nij}=j\,|\,\theta_n,\phi_{nh}=1,\beta_i,\tau_{hk})^{\phi_{nh}b_{ik}} \quad (8.2)$$

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

$$P\left(\mathbf{X}_n=\mathbf{x}_n\,|\,\theta_n,\phi_n,\beta_i,\mathbf{b}_i,\mathrm{T}\right)=\prod_h\prod_k\prod_i P(X_{nij}=x_{ij}\,|\,\theta_n,\phi_{nh}=1,\beta_i,\tau_{hk})^{\phi_{nh}b_{ik}} \quad (8.3)$$

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, & 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 classified into the stage for which that person's probability of membership is highest; however, it is possible to investigate the confidence with which we classify persons with various sorts of response patterns into that stage. Software for this purpose was developed by Draney (in press).

THE DEDUCTIVE REASONING TEST

Theoretical Background and Exploratory Data Analysis

The Competence Profile Test of Deductive Reasoning—Verbal (DRV; Spiel, Glück, & Gößler, 2001, 2004; Spiel & Glück, in press) was developed to assess competence profile and competence level in deductive reasoning in the course of the transition from the concrete-operational stage to the formal-operational stage as proposed by Piaget (1971). The test makes use of the four main types of syllogistic inference. Each item consists of a given premise ("if A, then B"), and a conclusion. The task is to evaluate a conclusion, assuming the premise as given. The four types of inferences are: Modus Ponens (A, therefore B), Negation of Antecedent (Not A, therefore B or not B), Affirmation of Consequent (B, therefore A or not A), and Modus Tollens (Not B, therefore not A). Table 8.1 gives examples of the four inference types. Modus ponens (MP) and modus tollens (MT) are biconditional conclusions, that is, the response to the respective items is either "yes" or "no." For negation of antecedent (NA) and affirmation of consequent (AC), however, the correct solution is "perhaps", as the premise does not allow for deciding whether these conclusions are correct. However, they provoke the choice of a biconditional, but logically incorrect conclusion ("no" for NA, "yes" for AC), which is why they are often called logical fallacies. Research has shown that individuals at the concrete operational stage treat all four inferences as biconditional (e.g., Evans, Newstead, & Byrne, 1993; Janveau-Brennan & Markovits, 1999). While solution probability for these fallacies increases with progress in cognitive development, performance on MT items and, in some cases, also on MP items decreases.

TABLE 8.1 Example Items from the DRV

Premise: *If Klaus is ill, he is lying in his bed.* (Concrete content, no negation)

Type of Inference	Item Text	Correct Solution
Modus Ponens (MP): A → B.	Tom is ill. Is Tom lying in his bed?	yes
Negation of Antecedent (NA): Not A → B or not B.	Tom is not ill. Is Tom lying in his bed?	perhaps (incorrect biconditional solution: no)
Affirmation of Consequent (AC): B → A or not A.	Tom is lying in his bed. Is Tom ill?	perhaps (incorrect biconditional solution: yes)
Modus Tollens (MT): Not B → not A.	Tom is not lying in his bed. Is Tom ill?	no

This is interpreted as developmental progress because individuals who have noticed the uncertainty of the fallacies tend to overgeneralize (e.g., Byrnes & Overton, 1986; Markovits, Fleury, Quinn, & Venet, 1998).

The DRV consists of 24 single items (6 different premises × 4 types of inference). The six premises were constructed based on the literature on moderator variables of syllogistic reasoning. Studies have consistently shown that concrete items are easier to solve than abstract and counterfactual items (e.g., Overton, 1985), while empirical evidence concerning differences between abstract and counterfactual items is mixed. In addition, empirical investigations show systematic increases in task difficulty when negations were used in the antecedents (e.g., Roberge & Mason, 1978). Therefore, the DRV (Spiel et al., 2001, 2004; Spiel & Glück, in press) systematically varied three item characteristics in a 4 × 3 × 2 design (see Table 8.1):

- Type of inference: Modus Ponens, Modus Tollens, Negation of Antecedent, Affirmation of Consequent;
- Content of the conditional: Concrete (premise example: "If Tom is ill, he is lying in his bed"), Abstract ("if Y belongs to group F, Y has attitude g"), Counterfactual ("if it is evening, the sun rises");
- Mode of presentation of the antecedent: with and without negation.

Previous analyses were conducted using an exploratory approach—the general mixed Rasch model. Based on the literature, we expected to identify at least four latent classes corresponding to distinct developmental stages: concrete-operational (high solution probabilities for the biconditionals, low solution probabilities for the fallacies); formal-operational (high solution probabilities for all items); and two intermediate stages with markedly higher solution probabilities for the fallacies than in the concrete-operational stage but differences depending on item content (for details see Spiel et al., 2004). Data analyses were based on a sample of 418 students in grades 7 through 12. The exploratory analysis produced a best-fitting model containing three latent classes, which were given the following names and descriptions:

- Concrete-operational (36% of participants; mean probability of class membership: 0.96): Tend to correctly solve MP and MT items, and no others.
- Intermediate (32%; mean probability of class membership: 0.93): Tend to correctly solve concrete-level fallacy items (i.e., NA and AC), but have difficulty with concrete MP and MT items. The pattern for abstract and counterfactual items is the same as in the concrete-operational class.

- Advanced intermediate (32%; mean probability of class membership: 0.95): Performed better in the fallacies than in the biconditionals for all items, independent of content.

The results are illustrated in Figure 8.1. There were too few formal-operational individuals, who tend to correctly solve most items, present in the sample to reliably estimate the relevant parameters. Therefore, this group was postulated theoretically, but has not yet been empirically identified.

The Saltus Analysis

When applying the saltus model to a set of items, it is necessary to determine which items are representative of which stages. Ideally, for each stage it should be the case that persons are first fully capable of answering those items correctly (or at their highest level of correctness) when entering that stage. It is, of course, possible for persons at lower developmental stages to perform the item correctly; however, this usually occurs because of guessing or a poorly developed strategy that happens to produce the correct answer in some cases. Similarly, it is possible for persons at higher developmental stages to miss items at or below their developmental stage—due to the usual causes such as carelessness, wrong choice of strategy, and so on. A prototypical example of this is given in the Juice Mixtures Data, as analyzed by Draney and Wilson (2007). Each of the three stages described by Noelting

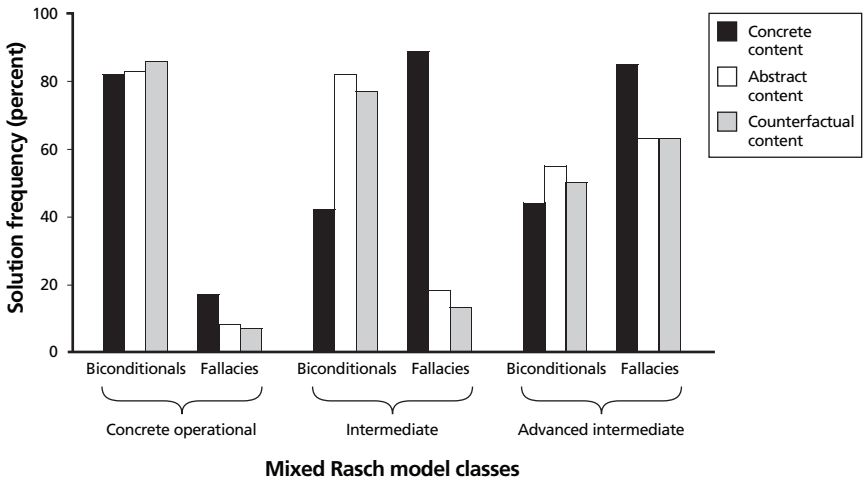

Figure 8.1 Results of prior mixed Rasch model analysis (Spiel, Glück, & Gößler, 2001, 2004).

(1980a, 1980b) is characterized by a set of skills that children acquire at that stage, and he has developed items representing each of those sets of skills.

In some cases the relation between item performance and stage membership is not so clear. This can happen for a number of reasons. One example is that persons answer certain item types correctly because they are disregarding part of the information available, information which actually makes the item more difficult. This is the case in children (and some adults) who tend to treat all syllogisms as biconditional because they do not understand that "if A then B" does not mean "if B then A". Once children understand this, they have a higher probability to solve the fallacies, but as was described above, for some time they tend to make more errors in the biconditional syllogisms (e.g., Byrnes & Overton, 1986; Markovits, Freury, Quinn, & Venet, 1998). A similar example concerning balance scale tasks was given by Siegler (1981).

Another difficulty that occurs when working with developmental theorists is that it is not always possible to develop a single group of items associated with each developmental stage. Instead, as in this case, there are several aspects (in this case type of inference, content, negation) which are fully crossed to form a set of items, for which it is then possible to predict the performance of each of the developmental stages. In such a case, it is more difficult to associate a set of items with each person group. We were not able to identify one class of items to be associated with each of four developmental groups, to attempt to replicate the results of the mixed Rasch analysis. However, we were able to propose two other saltus analyses: a two-group and a three-group analysis, as follow. Analysis 1 is a two-level analysis, in which the concrete person group is represented by all MP/MT items, and the formal person group is represented by all Fallacy items. Analysis 2 is a three-level analysis, in which the concrete person group is again represented by the MP/MT items, the intermediate person group is represented by the Concrete Fallacy items, and the formal person group is represented by Abstract and Counterfactual Fallacy items.

Saltus Results: Analysis 1

In this first analysis, the proportion of persons in the concrete group is estimated to be approximately 43%, and in the formal group 57%. Item difficulties ranged from –3.15 to –0.85 for the MP/MT items, and from 0.17 to 3.52 for the Fallacy items. This is as expected given that the MP/MT items are quite a bit easier than the Fallacy items, and the easiest Fallacy item is more difficult than the hardest of the MP/MT items. In addition, the estimated τ parameter was 4.28, suggesting that the formal group of persons has a substantial advantage over the concrete when performing the Fallacy items.

The estimated mean proficiency for the concrete group was –0.39 (with a standard deviation of 0.41), and for the formal group was –1.62 (with a standard deviation of 1.05). Thus, as was expected from the literature, although the formal group of persons has a substantial advantage on the Fallacy items, their performance on the MP/MT items was actually less good on average than that of the concrete group.

An examination of individual response patterns can shed light on such an analysis. Some illustrative response strings for the two groups, along with their classification probabilities, and proficiency estimates given membership in each group, are shown in Table 8.2. In this table, Persons A and B are most typical of persons assigned to the concrete group – persons who do well on the MP/MT items, and poorly on the Fallacy items. Persons similar to Person G, who do well on both groups of items, are most typical of persons assigned to the formal group. As persons C and D show, the tendency to begin to do well on some of the Fallacy items, especially on the more difficult items, is associated with an increasing probability of being in the formal group.

Persons like H, who do poorly on the MP/MT items, and well on the Fallacy items, are also not uncommon, and are also assigned to the formal group. Finally, persons who do poorly on both sets of items, such as persons E and F, are most likely to be in the formal group.

Examination of the score frequencies on both classes of items by classification group reveals that no one who scored over 5 on the Fallacy items was classified into the concrete group; most scored 0, 1, or 2. No one who scored under 6 on the MP/MT items was classified into the concrete group; most scored 9 or above. The persons who scored similarly (either low or high) on both sets of items were classified into the formal group. This helps to account for higher variance and lower mean of this group.

TABLE 8.2 Example Person Response Strings and Classifications for Analysis 1

Person	Responses MP/MT	Responses Fallacy	P(concrete)	P(formal)
A	111101111111	000000000000	1.00	.00
B	111011101100	110000000000	.98	.02
C	011011110100	100100000000	.86	.14
D	101111111011	111001001000	.36	.64
E	100000101111	101010000000	.08	.92
F	000010001100	100000000000	.01	.99
G	111111101100	111101111110	.00	1.00
H	000000000000	111111111000	.00	1.00

Thus, it does appear that more than one developmental level of subject has been classified into the formal group by this two-level analysis. There are persons who simply do poorly on all items (either students who are not yet developmentally capable of understanding the task, or perhaps students who are not paying particular attention); there are persons who, as Glück and Spiel (2007) describe, have begun to answer the Fallacy items correctly, but now are making errors on the MP/MT items; and there are those who are capable of correctly solving both types of items. This would indicate the need for a more sophisticated model. Thus, we fit a three-class model in Analysis 2, to see if this would accommodate such an effect.

Saltus Results: Analysis 2

Item difficulties in this analysis are similar to those in Analysis 1. Other results are shown in Table 8.3. It can be seen that several of the patterns from the first analysis hold in this second analysis. For example, the intermediate and formal groups have substantial advantages over the concrete group in answering the higher-level items (although the standard error for τ_{32} is large enough that this parameter is not statistically different from zero). In addition, the group means are ordered in the opposite way from which one might expect—the higher the group, the lower the overall mean proficiency. This shows that overall scores are often (in cases where the Rasch model does not hold across the whole set of items) less informative than score profiles across homogeneous groups of items. A profile contrasting the different types of inferences would show that participants' scores in the fallacy items increase whereas their scores in the biconditional items decrease. Overall, the largest proportion of persons (approximately half the sample) was classified into the intermediate group; 30% into the concrete group, and 20% into the formal.

The substantive interpretation here is more complex. Examination of individual response patterns shown in Table 8.4, as well as score cross-tabu-

TABLE 8.3 Parameter Estimates for Analysis 2

(a) Saltus parameters	Item class 1	Item class 2	Item class 3
Person group 1	—	—	—
Person group 2	—	4.88 (0.82)	7.84 (0.26)
Person group 3	—	3.31 (5.65)	7.20 (0.69)
(b) Person parameters	**Group 1**	**Group 2**	**Group 3**
Mean	–0.29	–1.65	–2.75
Standard Deviation	0.78	0.83	1.61
Proportion	0.31	0.51	0.18

TABLE 8.4 Example Person Response Strings and Classifications for Analysis 2

	Responses					
Person	MP/MT	Concrete Fallacy	Abstract/ Counterfactual Fallacy	P(group 1)	P(group 2)	P(group 3)
A	111111110000	0000	00000000	.99	.01	.00
B	100111111011	0000	00001000	.99	.01	.00
C	110010111000	0000	00000001	.43	.57	.00
D	111111100000	0000	00000101	.29	.71	.00
E	111011101100	1100	00000000	.10	.90	.00
F	110100001111	0010	00100000	.08	.92	.00
G	101010111111	1111	10101001	.00	.99	.01
H	110111111111	1111	11000000	.01	.99	.00
I	111110001111	0111	11000010	.00	1.00	.00
J	000011011111	1111	00000000	.00	1.00	.00
K	111010101111	1111	10101111	.00	.50	.50
L	100010001111	1110	10110101	.00	.28	.72
M	111111101111	1111	11111111	.00	.04	.96
N	101010101011	1111	11111111	.00	.00	1.00
O	000000000000	1111	11111111	.00	.00	1.00
P	000000001000	0001	10011111	.00	.00	1.00
Q	000000000100	1010	11100001	.00	.00	1.00

lation by group classification, shows that the concrete group contains persons who scored low on both class 2 items (concrete fallacies) and class 3 items (abstract and counterfactual fallacies) and high on class 1 items (all biconditional syllogisms). This is just as in Analysis 1—such persons scored 6 or above on the class 1 items, and 4 or below on the class 3 items (the majority scored below 2); in addition, none correctly answered all 4 class 2 items, and most got one or less of these items correct. Persons A and B in Table 8.4 are examples of concrete group members.

The intermediate is a mixed group. Persons such as G, H, and I are one type in this group; they have answered most of the class 1 and class 2 items correctly, and few of the class 3 items. Persons C and D, although they still have some probability of being in the concrete, are most likely also in this group. In addition to responding correctly to the majority of the class 1 items, they have answered one or two of the most difficult class 3 items correctly, although they have answered none of the class 2 items correctly. Person J, while missing a significant number of the class 1 items and all of the class 3 items, has answered all of the class 2 items correctly, and is

classified solidly in this group. Persons E and F are less intuitively obvious; although their probabilities clearly place them in the intermediate group, they look at first glance much like Persons C and D. Person E missed half of the class 2 items and several of the class 1 items, and thus could still be in the concrete group, but is most likely in the intermediate group; person F missed almost half of the class 1 items and most of the class 2 items, but perhaps because this person correctly answered one of the class 3 items, is still most likely to be in the intermediate group.

Persons who scored high on class 3 items (regardless of other scores) were in the formal group. Persons such as M—who answered nearly all the items correctly—typify what we would expect to see in this group. However, when examining the score distributions for persons with high scores on class 3 items, one tended in most cases to find the full range of possible scores on class 1 items (i.e., among persons who got more than 50% of the class 3 items correct, there were persons who had scores of 0 or 1 on the class 1 items). Hence, we see persons like O and P, who have missed many of the other items (especially those in class 1) but correctly answered more than half of the class 3 items, as well as persons like K, L, and N, who did well on class 2 and class 3 items, but missed significant numbers of the class 1 items. Finally, we see persons like Q, who has done uniformly poorly on all item types. This might account for the large standard error accompanying the advantage assigned to this group when taking class 2 items—some persons in this group did quite well on these items, while some did poorly.

Comparisons of the two models fit in Analyses 1 and 2, and to the one-group Rasch model (which was fit for comparison purposes only), are shown in Table 8.5. Comparisons across models with differing numbers of latent groups cannot be done using likelihood ratio tests, because of boundary problems (Böhning, 2000). Model comparisons are thus based on Akaike's information criterion (AIC; Akaike, 1974). In this table, it can be seen that the three-group saltus model does fit better than both the two-group saltus and the Rasch model, and thus is the preferred model of the three. Similar results were found in the exploratory analysis.

TABLE 8.5 AIC for One-, Two-, and Three-Class Models

Model	–2*log	# parameters	AIC
Rasch	11,456.12	25	11,506.12
2-class saltus	10,190.06	29	10,248.06
3-class saltus	9,751.21	35	9,820.21

DISCUSSION

The application of the more confirmatory saltus model to the deductive reasoning data, and the comparison of this set of analyses to the more exploratory mixed Rasch model, show a number of things. For example, positive saltus parameters show that persons at higher developmental levels did have a substantial advantage when performing the various types of fallacy items, and that the fallacy items decreased substantially in their difficulty as one attained the higher developmental groups, when compared to the performance of the concrete group. However, group means ordered such that the higher the developmental group, the lower the overall proficiency, indicates that the performance of the higher groups on the items associated with a lower developmental level (the MP/MT items) showed a reversal, and this reversal persisted at the highest developmental level. Indeed, consistent with theoretical assumptions as well as with the exploratory analysis, it was relatively unusual for persons classified at the higher developmental levels to achieve a perfect score on the lower level items.

Second, we see that while there are clearly latent groups of persons, with clear differences in their patterns of performance (as was also shown in the exploratory analysis), the relation between groups of persons and classes of items is more complex than can easily be shown by associating one class of items with each developmental group, and expecting better performance on that class of item by the group in question. The fact that the higher developmental groups tended to contain greatly varying subtypes of response patterns suggests that there are more than three differing groups. While there are groups that do successively better on more complex items, there are also groups that show interesting reversals in performance on particular sets. In some cases, this is predicted by theory (e.g., that intermediate-level persons would do poorly on the MP/MT items on which the concrete persons, assumed to be in a lower developmental level, do well). However, there are also persons who do well only on those items predicted to be the most difficult, and poorly on all other items. Finally, there are doubtless groups of persons who are simply guessing at random, or otherwise not paying close attention to the task.

It would thus be useful to fit a model which, for example, allowed estimation of the effect of the various aspects of the items (form, negation, complexity) on the performance of the various groups. This might allow us to investigate particularly troublesome effects, such as the tendency of persons who score uniformly poorly across all item types to be classified into the highest group, when this is quite counter-intuitive. It is also possible that various types of individual item and person fit analysis might prove useful in further understanding (e.g., helping us to detect persons who are likely to be random guessers, or using unusual solution strategies). For ex-

ample, von Davier and Molenaar (2003) presented person fit statistics that can be used with latent class and discrete mixture Rasch models.

In addition, it may be helpful to develop a series of additional models to fit, and to compare to the results of the current model. For example, it might be useful to develop a saltus-like model with variable item slopes, as models with equal slopes for all items are often too restrictive to fit well. In addition, it might be the case that models which included saltus parameters indexed by individual item or step within item, rather than simply associating saltus parameters with items as a whole, and estimating a single parameter across all items within an item class, might yield interesting differences by item and/or step.

Von Davier and Rost (1995) discussed the estimation of mixed Rasch models, including models for polytomous data, with and without constraints using conditional maximum likelihood methods. These model are all members of the class of finite mixture distribution models (e.g., Everitt & Hand, 1981; Titterington, Smith, & Makov, 1985); the investigation of various such models could prove quite useful in understanding the subtleties of data such as the deductive reasoning data. Perhaps the most general of such models to have been discussed in an educational context is the Mixture Multidimensional Random Coefficients Multinomial Logit (M^2RCML) model described by Pirolli and Wilson (1998). There is currently no available software which has been programmed to estimate parameters for such a general class of models.

One promising method for such parameter estimation, however, is through their expression as generalized nonlinear mixed models. Statistical software packages are being developed which can estimate a wide variety of such models. An example of how this could be done using SAS was given by Fieuws, Spiessens, and Draney (2004); other software packages, such as GLLAMM (see Rabe-Hesketh & Skrondal, 2005) could also be used. In general, systematic comparison of various methodological approaches is recommended (see, e.g., Glück & Spiel, 1997, 2007).

In sum, the saltus model has shown potential for aiding researchers, especially in the fields of cognitive science and Piagetian or neo-Piagetian theory, as do other extended models able to reflect the complexities of polytomous data and latent classes. However, theories concerning the development of competence do not always indicate simple linear increases in performance. In such cases, one needs to combine complex item response models with careful item construction as in the DRV. Other promising applications should follow as researchers in psychometrics continue their collaboration with educational and psychological researchers.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Armon, C. (1984). *Ideals of the good life: A longitudinal/cross-sectional study of evaluative reasoning in children and adults.* Unpublished doctoral dissertation, Harvard University, Boston.

Béland, A., & Mislevy, R. J. (1996). Probability-based inference in a domain of proportional reasoning tasks. *Journal of Educational Measurement, 33,* 3–27.

Böhning, D. (2000). Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others. *Monographs on Statistics and Applied Probability, 81.* Boca Raton, FL: Chapman & Hall/CRC.

Bond, T. G. (1995a). Piaget and Measurement I: The twain really do meet. *Archives de Psychologie, 63,* 71–87.

Bond, T. G. (1995b). Piaget and Measurement II: Empirical validation of the Piagetian model. *Archives de Psychologie, 63,* 155–185.

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

Byrnes, J. P., & Overton, W. F. (1986). Reasoning about certainty and uncertainty in concrete, causal, and propositional contexts. *Developmental Psychology, 22,* 793–799.

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

Dawson, T. (2002). New tools, new insights. Kohlberg's moral judgment stages revisited. *International Journal of Behavior Development, 26,* 154–166.

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,* 4–112.

Demetriou, A., & Efklides, A. (1989). The person's conception of the structures of developing intellect: Early adolescence to middle age. *Genetic, Social, and General Psychology Monographs, 115,* 371–423.

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

Dempster, A. P., Laird, N. M., & 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. (1996). The polytomous Saltus model: A mixture model approach to the diagnosis of developmental differences. Unpublished doctoral dissertation, University of California, Berkeley.

Draney, K. (in press). The saltus model applied to proportional reasoning data. *Journal of Applied Measurement.*

Draney, K., & Wilson, M (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. von Davier & C. H.

Carstensen (Eds.), *Multivariate and mixture distribution Rasch models—Extensions and applications*. New York: Springer.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Mahwah, NJ: Erlbaum.

Everitt, B. S., & Hand, D. J. (1981). Finite mixture distributions. New York: Chapman and Hall.

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

Fischer, G. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3–26.

Fischer, K. W., Hand, H. H., & Russel, S. (1984). The development of abstractions in adolescence and adulthood. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development* (pp. 43–73). New York: Praeger.

Fischer, K. W., Pipp, S. L., & Bullock, D. (1984). Detecting discontinuities in development: Methods and measurement. In R. N. Emde & R. Harmon (Eds.), *Continuities and discontinuities in development*. Norwood, NJ: Ablex.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. *Methods of Psychological Research—Online, 2 [Online Journal]; http://www.mpr-online.de*.

Glück, J., & Spiel, C. (2007). Using item response models to analyze change: Advantages and limitations. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models—Extensions and applications*. New York: Springer.

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

Janveau-Brennan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology, 35,* 904–911.

Kohlberg, L., & Candee, D. (1984). The six stages of justice development. In L. Kohlberg (Ed.), The psychology of moral development: The nature and validity of moral stages (Vol. 2, pp. 621–683). San Francisco: Jossey-Bass.

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

Markovits, H., Fleury, M.-L., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69,* 742–755.

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

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

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71

Noelting, G. (1980a). The development of proportional reasoning and the ratio concept—part I: Differentiation of stages. *Educational Studies in Mathematics, 11,* 217–253.

Noelting, G. (1980b). The development of proportional reasoning and the ratio concept—part II: Problem-structure at successive stages; problem-solving strategies and the mechanism of adaptive restructuring. *Educational Studies in Mathematics, 11,* 331–363.

Noelting, G., Coudé, G., & Rousseau, J. P. (1995, June). *Rasch analysis applied to multiple-domain tasks.* Paper presented at the twenty-fifth annual symposium of the Jean Piaget Society, Berkeley, CA.

Overton, W. F. (1985). Scientific methodologies and the competence-moderator-performance issue. In E. D. Neimark, R. de Lisi, & J. L. Newman (Eds.), *Moderators of competence* (pp. 15–41). Hillsdale: Erlbaum.

Piaget, J. (1950). *The Psychology of Intelligence.* (M. Piercy, Trans.) London: Lowe & Brydone. (Original work published 1947).

Piaget, J. (1971). *Biology and knowledge.* Chicago: University of Chicago Press.

Pirolli, P., & Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105,* 58–82.

Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using Stata.* College Station, TX: Stata Press.

Roberge, J. J., & Mason, E. J. (1978). Effects of negation on adolescents' class and conditional reasoning abilities. *The Journal of General Psychology, 98,* 187–195.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monograph of the Society for Research in Child Development, 46*(1, Serial No. 189).

Spiel, C., Glück, J. & Gößler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Adolescent Research, 16,* 150–168.

Spiel, C., Glück, J., & Gößler, M. (2004). Messung von Leistungsprofil und Leistungshöhe im schlussfolgernden Denken im SDV—Die Integration von Piagets Entwicklungskonzept und Item-Response Modellen. *Diagnostica, 50,* 145–152.

Spiel, C., & Glück, J. (in press). A model based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

van Hiele, P. M. (1986). *Structure and insight: A theory of mathematics education.* Orlando, FL: Academic Press.

von Davier, M., & Molenaar, I. W. (2003). *A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. Psychometrika, 68,* 213–228.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In Fischer, G. H., Molenaar, I. W. (Eds.), *Rasch models—Foundations, recent developments, and applications.* New York: Springer.

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

Yen, W. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional Item Response Theory. *Psychometrika, 50,* 399–410.