The Ordered Partition Model: An Extension of the Partial Credit Model

Mark Wilson
University of California, Berkeley

An item response model, called the ordered partition model, is designed for a measurement context in which the categories of response to an item cannot be completely ordered. For example, two different solution strategies may lead to an equivalent degree of success because both strategies may result in the same score, but an examiner may want to maintain the distinction between the strategies. Thus, the data would not be nominal nor completely ordered, so may not be suitable for other polytomous item response models such as the partial credit or the graded response models. The ordered partition model is described as an extension of the partial credit model, its relationship to other models is discussed, and two examples are presented. *Index terms: ordered partition model, partial credit model, partial order model, polytomous* IRT *model, Rasch model.*

The ordered partition model (OPM) is designed to represent empirical response formats that are based on the idea of an ordered partition. Mathematically, an ordered partition divides a set into equivalence classes that then can be ordered, but the order does not extend inside the equivalence classes (Kemeny, Snell, & Thompson, 1957). Suppose that, according to some substantive theory, students' responses can be classified into a series of successive levels and that a number of different manifestations lead to classification into a particular level. This type of ordered partition scoring arises in theories of learning, cognition, and development in which (1) there is an a priori concept of progress between levels as represented by the items, and (2) different types of responses may be scored within a given level. In this conception, the relationship between the alternatives within a particular level is not compensatory. If one alternative is observed, it is the equivalent of observing any other within that level, and observing more than one within the same level is either not modeled or not possible.

Theoretical structures of this kind have been common in psychology and education for many years—in developmental and especially Piagetian studies (e.g., Flavell, 1971; Piaget, 1950; Werner, 1957), in learning hierarchies (e.g., Gagne, 1968), and in other educational domains (e.g., Van Hiele, 1986). However, the measurement techniques applied to data collected in the investigations of these kinds of theoretical structures have been much less sophisticated than those applied to simpler dichotomous data situations (Fischer, Pipp, & Bullock, 1984; Wilson, 1990). This has led to criticism—measurement has been accused of retarding the development of theory and applications (Glaser, Lesgold, & Lajoie, 1987). More recently, with the rise of the ''authentic'' or ''performance'' assessment movement in education (Wiggins, 1989), ordered partition scoring schemes are becoming commonplace (e.g., Assessment of Performance Unit, 1980; Baron, Forgione, Rindone, Kruglanski, & Davey, 1989; Baxter, Shavelson, Goldman, & Pine, 1991; California Assessment Project, 1989).

The OPM is proposed as a step toward the development of flexible measurement techniques to guide the application of good measurement practices to help theoretical progress in psychology and education. Although it does not encompass all possible complications that can arise

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 4, December 1992, pp. 309-325
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/040309-17$2.10

in these situations, the OPM does deal with one complication that is presently not handled by familiar models.

Extending the Partial Credit Model

Consider a situation in which there are I items ($i = 1, \ldots, I$). Each item has M_i response categories ($k = 1, \ldots, M_i$) that are graded into M_i possible score levels ($m = 0, \ldots, M_i - 1$). The categories are related to the levels by a function $B_i(k)$ that is given by $B_i(k) = k - 1$.

For data possessing this structure, Masters (1982, 1988) provided a derivation of the partial credit model (PCM) from certain assumptions. Here, the pattern of that derivation is extended to the ordered partition case. Masters showed that the PCM can be derived if it is assumed that the probability of person n, with parameter θ_n, responding in the first level of a pair of consecutive levels is governed by a simple logistic model,

$$\frac{p_{nim}}{p_{ni(m-1)} + p_{nim}} = \frac{\exp(\theta_n - \delta_{im})}{1 + \exp(\theta_n - \delta_{im})}\,, \qquad (1)$$

for $m = 1, 2, \ldots, M_i$,

where p_{nim} is the probability of person n responding in level m to item i, and δ_{im} is a parameter associated with the transition between levels $m - 1$ and m. Equation 1 can be rewritten (Andersen, 1973, 1983) as:

$$\frac{p_{nim}}{p_{ni(m-1)} + p_{nim}} = \frac{\exp\{\theta_n - [\eta_{im} - \eta_{i(m-1)}]\}}{1 + \exp\{\theta_n - [\eta_{im} - \eta_{i(m-1)}]\}}, \qquad (2)$$

for $m = 1, 2, \ldots, M_i$,

where η_{im} is a parameter associated with each level, $\eta_{i0} = 0$, and the relationship between Masters' δ parameter and Andersen's η parameter is given by $\delta_{im} = \eta_{im} - \eta_{i(m-1)}$. Masters derived the PCM from this assumption, and the resulting formulation for score levels is:

$$p_{nim} = \frac{\exp(m\theta_n - \eta_{im})}{\sum_{t=0}^{M_i} \exp(t\theta_n - \eta_{it})}, \qquad (3)$$

for $m = 0, 1 \ldots, \mathrm{M}_i$.

Masters (1982) originally defined p_{nim} as

$$p_{nim} = \frac{\exp\left(\sum_{j=1}^{m} \theta_n - \delta_{ij}\right)}{1 + \sum_{h=1}^{M_i} \exp\left(\sum_{j=1}^{h} \theta_n - \delta_{ij}\right)} \qquad (4)$$

for $m > 1$, and

$$p_{ni1} = \left[1 + \sum_{h=1}^{M_i} \exp\left(\sum_{j=1}^{h} \theta_n - \delta_{ij}\right)\right]^{-1} . \qquad (5)$$

Wilson & Masters (in press) have shown that, for the PCM, Equation 2 can be generalized to:

$$\frac{p_{nim}}{p_{nih} + p_{nim}} = \frac{\exp\{[B_i(m) - B_i(h)]\theta_n - (\eta_{im} - \eta_{ih})\}}{1 + \exp\{[B_i(m) - B_i(h)]\theta_n - (\eta_{im} - \eta_{ih})\}}, \quad (6)$$

for *any* $m, h \in \{0, 1, \ldots, M_i\}$.

They used this relationship for response levels in which no person responds. In this context, it will be used to expand the applicability of the PCM. Note that the relationship holds even when $m = h$.

An Illustrative Example

Consider a more complex situation in which there is not a one-to-one relationship between response categories and score levels, which represents the idea behind the OPM. Suppose there are four categories ($k = 1, \ldots, 4$) in which a person can respond to item i. On an a priori basis these categories have been scored into three levels, ranging from $m = 0$ to $m = 2$. Let the function B_i define this scoring scheme, and define B_i by the mapping: $B_i(1) = 0$, $B_i(2) = B_i(3) = 1$, and $B_i(4) = 2$.

Now make the following assumption, which is similar to Masters (1982). Suppose that a relationship analogous to Equation 2 holds between the probabilities of responding in consecutive categories, π_{nik}; the parameters associated with persons, θ_n; and the parameters associated with categories, ξ_{ik} (where ξ is used rather than η to emphasize that these parameters are not necessarily the same as those used in Equation

2, and π is used rather than p for a similar reason):

$$\frac{\pi_{nik}}{\pi_{ni(k-1)} + \pi_{nik}} =$$

$$\frac{\exp\{[B_i(k) - B_i(k-1)]\theta_n - [\xi_{ik} - \xi_{i(k-1)}]\}}{1 + \exp\{[B_i(k) - B_i(k-1)]\theta_n - [\xi_{ik} - \xi_{i(k-1)}]\}} , \qquad (7)$$

for $k = 2, 3, 4$,

where $\xi_{i1} = 0$, and K_i is the number of response categories for item i.

For the particular situation under consideration, Equation 7 becomes:

$$\frac{\pi_{ni2}}{\pi_{ni1} + \pi_{ni2}} = \frac{\exp(\theta_n - \xi_{i2})}{1 + \exp(\theta_n - \xi_{i2})} , \qquad (8)$$

$$\frac{\pi_{ni3}}{\pi_{ni2} + \pi_{ni3}} = \frac{\exp(\xi_{i3} - \xi_{i2})}{1 + \exp(\xi_{i3} - \xi_{i2})} , \qquad (9)$$

and

$$\frac{\pi_{ni4}}{\pi_{ni3} + \pi_{ni4}} = \frac{\exp[\theta_n - (\xi_{i4} - \xi_{i3})]}{1 + \exp[\theta_n - (\xi_{i4} - \xi_{i3})]} . \qquad (10)$$

Equations 8 and 10 are formally equivalent to Equation 2, with appropriate substitutions. Equation 9 is different in form—the categories have the same score; therefore, the multiplicative factor associated with θ_n is 0, and θ drops out of the equation. In substantive terms, this is what it means to give the same score to two responses. By ascribing the two responses an equal score, it is implicitly assumed that the choice between the two responses is not sensitive to differences in θ. Hence, the relative probability of one or the other response will be a constant for all θs. If this assumption could not be met in the substantive situation, then the OPM would not be appropriate.

Now use Equations 8, 9, and 10 in combination with the normalizing relationship, $\pi_{ni1} + \pi_{ni2} + \pi_{ni3} + \pi_{ni4} = 1$, to deduce the form of the probabilistic model that is a consequence of Equation 7 (Masters, 1988). Elementary algebra provides

$$\pi_{ni1} = 1/\psi , \qquad (11)$$

$$\pi_{ni2} = \exp(\theta_n - \xi_{i2})/\psi , \qquad (12)$$

$$\pi_{ni3} = \exp(\theta_n - \xi_{i3})/\psi , \qquad (13)$$

and

$$\pi_{ni4} = \exp(2\theta_n - \xi_{i4})/\psi , \qquad (14)$$

where ψ is the sum of the numerators. Note that Equations 11, 12, 13, and 14 conform to the form:

$$\pi_{nik} = \exp[B_i(k)\theta_n - \xi_{ik}]/\psi , \qquad (15)$$

which is a generalization of Equation 3, substituting $B_i(k)$ for m.

Note that the connection must be made back from the category representation to the score representation, as in Equation 3. First, the function B_i implies the following relationships:

$$\pi_{ni1} = p_{ni0} , \qquad (16)$$

$$\pi_{ni2} + \pi_{ni3} = p_{ni1} , \qquad (17)$$

and

$$\pi_{ni4} = p_{ni2} . \qquad (18)$$

Comparing Equations 11, 12, 13, and 14 with Equation 3, it can be seen that Equation 16 requires that the denominators in Equations 3 and 15 must be the same in this instance. Hence, Equation 18 will hold if

$$\eta_{i2} = \xi_{i4} , \qquad (19)$$

and Equation 17 will hold if η_{i1} is selected, such that $\exp(-\eta_{i1}) = \exp(-\xi_{i2}) + \exp(-\xi_{i3})$. Rearranging this relationship gives

$$\eta_{i1} = -\ln(e^{-\xi_{i2}} + e^{-\xi_{i3}}) . \qquad (20)$$

Thus, if the OPM parameters (ξ_{i2}, ξ_{i3}, ξ_{i4}) describe a model for the categories under the scoring scheme defined by B_i, then the PCM parameters (η_{i1}, η_{i2}), as defined by Equations 19 and 20, will describe a PCM model consistent with the OPM model (Equations 16, 17, and 18).

The Ordered Partition Model

Formal Definition of the Model

Consider a situation in which there are I items ($i = 1, \ldots, I$). Each item has K_i response categories ($k = 1, \ldots, K_i$), which are graded into $M_i + 1$ possible score levels ($m = 0, \ldots, M_i$). For item i, response k is assigned on an a priori basis to level m by the function $B_i(k)$, which is the scoring scheme for that item. That is $B_i(k) = m$.

Note that B_i is a known function for each item, but B_i need not have the same definition for all items. Each response must map onto only one level (but several responses may map onto the same level); and to avoid redundancy, each level must be represented by at least one alternative from some item in the test.

Let X_{ni} be a random variable that represents the response of person n to item i. Under the OPM, person n, with ability θ_n, has the probability of selecting the response with index k

$$P(X_{ni} = k) = \frac{\exp[\theta_n B_i(k) - \xi_{ik}]}{\sum_{h=1}^{K_i} \exp[\theta_n B_i(h) - \xi_{ih}]} , \qquad (21)$$

where ξ_{ik} is the parameter associated with response k on item i, and $\xi_{i1} \equiv 0$. Usually, the sum of the item parameters is constrained to 0, but other constraints may be used. The example above did not have multiple categories in the first and last levels. This should not be considered a limitation of the model; Equation 21 can be applied to situations in which there are multiple categories at any of the defined levels.

For the score levels $m = 1, 2, \ldots, M_i$, the parameters δ_{im} can be defined that are equivalent to Master's (1982) partial credit "step" parameters as a direct function of the OPM parameters:

$$\delta_{im} = \ln\left[\frac{\sum_{B_i(t) = m-1} \exp(-\xi_{it})}{\sum_{B_i(t) = m} \exp(-\xi_{it})}\right] , \qquad (22)$$

where t in the numerator and denominator is a dummy variable indexing the categories in levels $m - 1$ and m, respectively (Kelderman, 1989). Equation 22 is a generalization of Equation 20. Note that this expression applies when there are multiple categories in the extreme categories and when there are multiple categories in nonextreme categories.

Relationship to Other Models

The nominal model. To clarify the relationship of the OPM to other item response models, it is convenient to start with the taxonomy suggested by Thissen & Steinberg (1986). They discussed several extensions of the PCM, each of which "fills in the parameter space" (p. 572) between the PCM and Bock's nominal model (Bock, 1972). The OPM can be construed as being between the nominal model and the PCM. The response model in Equation 21 for the OPM and the nominal model are of the same form. The difference is that the weights represented by the scoring function B_i are fixed a priori in the OPM, but the equivalent function is estimated in the nominal model. Understanding the reason for this difference is crucial to understanding the usefulness of the OPM. The same set of data could be analyzed with both the nominal model and the OPM, but the OPM assumes that there is extra information that the nominal model does not need—substantive knowledge about the structure of the item responses in the shape of the scoring scheme represented by the B_i. If that knowledge is not available, then it would be misleading to "make up" the function B_i; therefore, the nominal model would provide a good starting point. If that knowledge is available, it can be ignored at some risk, and the OPM would be an obvious starting point under such circumstances. Thus, the OPM incorporates a priori substantive knowledge about scoring schemes into the calibration of the items; the nominal model leaves that to empirical estimation.

Samejima's graded response model. Another model that has been used in this context is the Samejima graded response model (SGRM;

Samejima, 1969). Unfortunately, no generalization from the SGRM to the OPM exists; therefore, the SGRM is generalized differently. (The following method of introducing an ordered partition into the SGRM was suggested by an anonymous reviewer.)

Let m_k denote the score level of category k for item i. The SGRM is based on probabilities of scoring at or above a given level. The probability of selecting any response at or above level m_k is given by

$$P^*_{im_k}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{im_k})]}{1 + \exp[\alpha_i(\theta - \beta_{im_k})]} \; . \qquad (23)$$

The probability of any response at level m_k is given by

$$\pi_{im_k} P_{im_k}(\theta) = P^*_{im_k}(\theta) - P^*_{im_{k-1}}(\theta) \; . \qquad (24)$$

For each score level at which there are one or more response categories, the conditional probabilities for those categories within their respective levels can be defined as π_{imk}. These conditional probabilities sum to 1 within a level and do not depend on θ. The conditional probability of responding within category k is $\pi_{imk}P_{imk}(\theta)$. This is not as elegant as the OPM formulation, nor is it an instance of a general form.

Interpreting the Parameters

The OPM parameters, ξ, can be used to interpret the results. For any given person (i.e., for constant θ), the odds for item i of that person being in category k, as opposed to being in k', are proportional to

$$\mathrm{O}_i^{k'k} = \frac{\exp(-\xi_{ik})}{\exp(-\xi_{ik'})} \; . \qquad (25)$$

Equation 25 is the same between and within score levels. As an extension to the PCM, the interpretive techniques developed by Wright & Masters (1982) also are available. In particular, the odds (for constant θ) of being in a higher category, m, as opposed to being in a lower category $m - 1$ are proportional to:

$$\mathrm{O}_{im} = \sum_{B_i(k) = m} \mathrm{O}^k_{im} \; , \qquad (26)$$

where

$$\mathrm{O}^k_{im} = \frac{\exp\xi_{ik}}{\sum_{B_i(t) = m-1} \exp\xi_{it}} \qquad (27)$$

can be interpreted as proportional to the odds of being in category k (in level m), rather than level $m - 1$.

Estimation

Equation 21 shows that the OPM can be considered a special case of Kelderman's (1989) polytomous loglinear item response model. Formulating the OPM in this manner has advantages: (1) the computer program LOGIMO (Kelderman & Steen, 1988) can be used to estimate parameters, and (2) the extension to a multidimensional framework can be made in the same way as outlined by Kelderman (1989). Parameters are estimated by modified iterative proportional fitting. The algorithm is a relatively straightforward modification of the algorithm used for the dichotomous loglinear Rasch model (Kelderman, 1984).

The OPM model also may be estimated by other means. Wilson & Adams (in press) described a marginal maximum likelihood algorithm that may be used to estimate the OPM parameters and provided a monte carlo study to investigate the behavior of the algorithm under differing numbers of quadrature points. Conditional maximum likelihood estimation is also theoretically possible. Adams & Wilson (1992) proposed an algorithm for a somewhat more general class of models, and a computer program is available to implement the procedure.

Example Applications

Application to the SOLO Taxonomy

Tests and data. The OPM was applied to measuring learning outcomes based on the neo-Piagetian theory of learning called SOLO (Structure of the Learning Outcome; Biggs & Collis, 1982). SOLO levels are defined in Table 1. It is expected that, for a given topic, learners will move through each level—from the prestructural to the extended abstract—as their comprehension and maturity improve. Furthermore, the majority of responses should be classifiable into one of the levels in the SOLO taxonomy indicating the learner's location on a latent dimension.

Table 1
SOLO Levels

Level	Description
P	A *prestructural* response is one that consists only of irrelevant information.
U	A *unistructural* response is one that includes only one relevant piece of information from the stimulus.
M	A *multistructural* response is one that includes several relevant pieces of information from the stimulus.
R	A *relational* response is one that integrates all relevant pieces of information from the stimulus.
E	An *extended abstract* response is one that not only includes all relevant pieces of information, but extends the response to integrate relevant pieces of information not in the stimulus.

SOLO may be used to construct item clusters that consist of some stimulus material followed by items that are keyed to successive levels in the SOLO structure (Collis & Davey, 1986). The simplest of these consists of one item at each level, but more accurate measurement can be attained by having several items at each level (Wilson, 1989). However, this results in a complicated decision rule for determining which level is indicated by any particular response vector. These decision rules can be represented by ordered partitions of the response vectors. The OPM allows for these different decision rules and compares their suitability.

This method of test construction is a special case of the superitem strategy described by Cureton (1965)—such subsets of items also are known as "item clusters," "item bundles," or "testlets." In the superitem strategy, a subset of items is linked by common stimulus material or other substantive feature, and the superitem is scored by the sum of the item scores. An example is shown in Table 2. The test illustrated in the top half of Table 2 is composed of two subtests—Subtest 1 had three dichotomous items, and Subtest 2 had two dichotomous items. The levels are scores on the subtests, and the response vectors are the different ways of achieving each of those scores, that is, the superitem response patterns. It is assumed that the two subtests start from an equal difficulty level, so that succeeding on two items is equally indicative of ability for either subtest. A somewhat less familiar situation also is illustrated in the second example in Table 2 in which there are a priori grounds to suppose that the three-item subtest includes an item from a lower level, so that persons of equal ability would be expected to score one point higher on the three-item subtest (Subtest 1) than on the two-item subtest (Subtest 2). In this case, the levels and the scores differ. In general, the terms "score" and "level" are synonymous, but the second example indicates that in some circumstances it is worthwhile distinguishing them.

The SOLO taxonomy was used to generate superitems (Romberg, Collis, Donovan, Buchanan, & Romberg, 1982; Romberg, Jurdak, Collis, & Buchanan, 1982) in the domain of mathematics. An example of one of the superitems is given in Figure 1. In discussing the results, individual items within a superitem will be referred to as "questions." Each successive question was linked to the unistructural, multistructural, or relational level, respectively. The responses were judged as acceptable or unacceptable according to an agreed set of criteria. The seven items examined were part of a larger study of the usefulness of the SOLO superitem format for assessment of mathematics ability (Romberg, Collis, Donovan, Buchanan, & Romberg, 1982; Romberg, Jurdak, Collis, & Buchanan, 1982). The data were gathered from 257 students from grades four, six, and eight in a central Wisconsin school district. Because of the age of the students, only the first four levels

Table 2
Two Examples of Scoring for Tests With Two Subtests Each

	Subtest 1			Subtest 2		
Level	Score	Response Vector	Response Index	Score	Response Vector	Response Index
Example 1						
3	3	(1,1,1)	8			
2	2	(1,1,0)	7	2	(1,1)	4
		(1,0,1)	6			
		(0,1,1)	5			
1	1	(1,0,0)	4	1	(1,0)	3
		(0,1,0)	3		(0,1)	2
		(0,0,1)	2			
0	0	(0,0,0)	1	0	(0,0)	1
Example 2						
3	3	(1,1,1)	8	2	(1,1)	4
2	2	(1,1,0)	7	1	(1,0)	3
		(1,0,1)	6	1	(0,1)	2
		(0,1,1)	5			
1	1	(1,0,0)	4			
		(0,1,0)	3			
		(0,0,1)	2			
0	0	(0,0,0)	1	0	(0,0)	1

(i.e., excluding extended abstract) were assessed.

Because of the hierarchical nature of the taxonomy, it was expected that the majority of the responses would be in the form of "Guttman true-types" (Guttman, 1941). That is, for most learners, success on one level of the taxonomy would be preceded by success on all lower levels. Clearly, when the response to a superitem conforms to one of these patterns, it can easily be assigned to one of the SOLO levels. Superitems that conform to the Guttman type could be scored by their usual score (i.e., simply the sum of the dichotomous items), and non-Guttman response patterns could be assigned to another dimension, in accordance with a *Guttman scoring scheme.* In contrast, the superitems could be scored in the traditional way, which will be called the *standard scoring scheme.* Students also could be mapped to the level of the highest question on which they succeed; this was called the *maximal scoring scheme.* A variant of the Guttman scheme would be to ignore success on any questions that are at levels above a failed question; this was called the *minimal scoring scheme.* These scoring schemes are given in Table 3.

Model fit. Each scoring scheme can be compared by examining the fit of the data to the OPM represented by the scheme. Because the models are not nested, Akaike's (1973) information criterion (AIC) was used for comparison:

$$\mathrm{AIC} = G^2 + 2p + C \ , \qquad (28)$$

where p is the number of estimated parameters, C is an arbitrary constant, and G^2 is the likelihood ratio statistic:

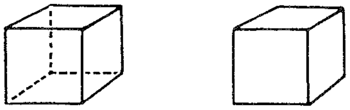
$$G^2 = -2\textstyle\sum f_{xt} \ln(m_{xt}/f_{xt}) \ , \qquad (29)$$

where f_{xt} is the observed frequency of persons with response pattern x and score t, m_{xt} is the modeled frequency, and the summation is taken over all logically nonzero cells. The comparison between models is achieved by calculating the difference between their respective AICs. All of these models, and those discussed below, are based on sparse contingency tables; therefore, the fit comparisons cannot be interpreted as firm indicators of model fit.

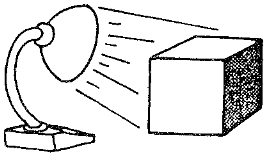
The fit of the standard scoring scheme was compared with the maximal and minimal

Figure 1
A Sample SOLO Superitem (Superitem 5)
(From Romberg et al., 1982; Reproduced by permission.)

Here are two pictures of a box. One has every face shown while the other shows a view only of those faces that could be seen if the box was solid. For the solid box, you can see 3 faces.



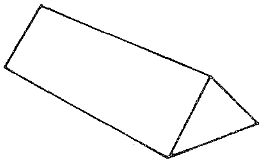
The drawing is of a light shining on a square-based box placed on a table. Three faces of the box are in the direct light. One of the three faces not in the direct light has been shaded.



A. What is the total number of faces that a square-based box has?

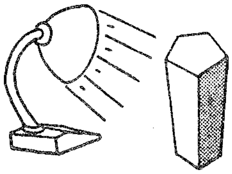
ANSWER 6

B. The picture at the side is a box with triangular ends. How many faces does it have altogether?



ANSWER 5

C. The picture at the side shows a light shining on a box with pentagonal (or five-sided) ends. Some faces are in the light. If the other boxes referred to in the table below were laid in the same position as the box in the picture, fill in the table with the number of faces that would be in the direct light and the total number of faces the box has.



	Number of faces in the direct light:	Total number of faces:
Triangular (3)	3	5
Square (4)	3	6
Pentagonal (5)	3	7
Hexagonal (6)	4	8
Septagonal (7)	4	9
Octagonal (8)	4	10
Nonagonal (9)	5	11

Table 3
Assignment of Responses to Levels for Each of the Scoring Schemes in the SOLO Example

Response			Scoring Scheme			
U	M	R	Standard	Guttman	Maximal	Minimal
0	0	0	0	0	0	0
1	0	0	1	1	1	1
1	1	0	2	2	2	2
1	1	1	3	3	3	3
0	1	0	1	*	2	0
0	0	1	1	*	3	0
1	0	1	2	*	3	1
0	1	1	2	*	3	0

* Scores could not be assigned to these response patterns.

schemes. The results of the estimations of these models are provided in Table 4. Clearly, the minimal and maximal scoring models fit almost equally well, but were somewhat less well-fitting than the standard scoring scheme.

Table 4
Goodness-of-Fit Statistics for the Scoring Schemes in Example 1

Scoring Scheme	Likelihood Ratio G^2	Number of Parameters	AIC
Standard	1,521.23	56	1,633.23
Maximal	1,558.78	57	1,672.78
Minimal	1,561.76	57	1,675.76
Guttman	1,581.07	61	1,703.07

As should be clear from Table 3, because several of the response categories have no scores, the Guttman scoring scheme is not appropriate for the ordered partition framework. To assign these responses a score that does not conflict with the Guttman scheme is equivalent to applying the scores to another latent dimension. This can be accomplished by going outside the ordered partition framework, but staying within the framework of Kelderman's (1989) multidimensional model for polytomously scored items. This can be done by introducing the possibility of incorporating s dimensions into the basic loglinear Rasch model:

$$P(X_{ni} = k) = \frac{\exp\left[\sum_{q=1}^{s} \theta_{nq} B_{iq}(k) - \xi_i(k)\right]}{\sum_{h=1}^{K} \exp\left[\sum_{q=1}^{s} \theta_{nq} B_{iq}(h) - \xi_i(h)\right]} , \qquad (30)$$

where θ_{nq} is the qth latent variable, and B_{iq} is the corresponding scoring function. This more complex model also can be estimated by LOGIMO (Kelderman & Steen, 1988).

In this example, the non-Guttman responses were assigned to a second dimension with the same scores as in the standard scheme, and Equation 30 was used as the basis for estimation using LOGIMO. The fit of the modified Guttman model also is given in Table 4, which shows that the standard scheme provided the best fit for the four models.

Parameter estimates. For the standard scoring model, the relationship of the OPM to the PCM can be illustrated, and the OPM parameter estimates can be interpreted according to Equations 25, 26, and 27.

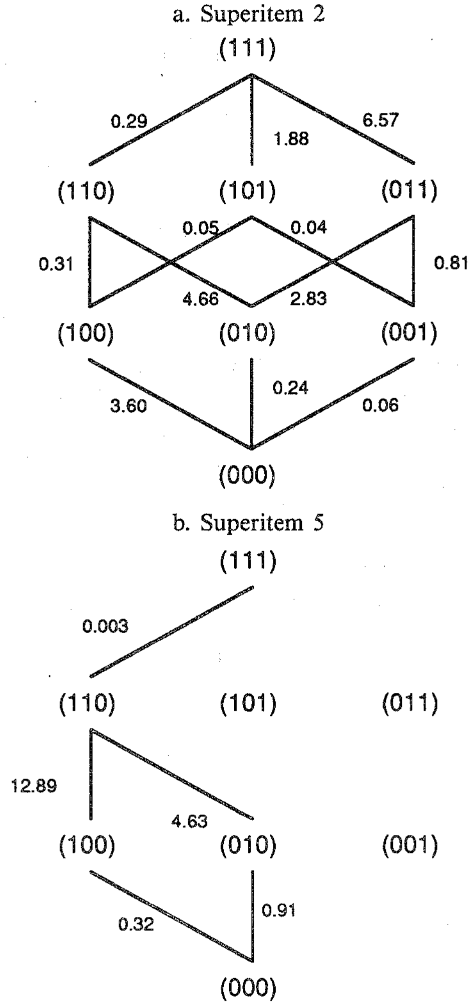
The partial credit "step" parameters were calculated (using the OPM estimates substituted into Equation 22) for Superitem 2 as –1.360, 1.079, and 1.432 for the steps from Level 0 to 1, 1 to 2, and 2 to 3, respectively. Direct estimation of the PCM parameters using LOGIMO provided almost identical values of –1.360, 1.079, and 1.433, respectively. Thus, using the interpretation common to partial credit analyses, the first step can be interpreted as relatively "easy" (the odds of being in Level 1 compared to Level 0 are 3.90 to 1), and the next steps are relatively more difficult (the odds of being in the higher category of each pair are .34 to 1, and .24 to 1, respectively).

The odds of being in a higher category (as opposed to being at the level below) were calculated and are shown in the Figure 2a. For the response (111), the odds were .24 to 1 (calculated above), because there was only one response category in Level 3. For Level 2, the odds of being in category (110), (101), or (011) compared to Level 1 were .28 to 1, .04 to 1, and .01 to 1, respectively. Thus, it is probable that a student in Level 2 will be in the first category (110), which is the Guttman response. For Level 1, the odds of being in category (100), (010), or (001), compared to Level 0 (000) were 3.60, .24, and .06, respectively. Once again, the Guttman response was the most likely. Some odds were not calculated because they did not correspond to interpretable changes in response patterns. The most likely sequence of response patterns, starting from (000), was the Guttman sequence. For each case in which a Guttman response was merely one of the possibilities, a Guttman option was always more likely than a non-Guttman response. Thus, for Superitem 2, the pattern of responses conformed to the SOLO expectations; that is, the Guttman responses were predominant in a probabilistic sense.

These results can be compared to Superitem 5, which had a pattern of results quite different than expected. The ordered partition analysis gave equivalent PCM steps of −.214, 1.217, and 5.679 (which also were indistinguishable from the directly estimated PCM estimates to two decimal places) for the first, second, and third steps, respectively. These results indicate that the first step for Superitem 5 was somewhat more difficult than the first step for Superitem 2, but that the third step was considerably more difficult than that for Superitem 2. This is not inconsistent with expectations—the PCM estimates give no indication of a problem with the superitem.

The OPM results for Superitem 5 (Figure 2b) are quite different from those for Superitem 2, and are inconsistent with expectations. First, several categories were not observed at all, so their respective odds are not shown. Second, and more importantly, the odds of the Guttman option (100) were less than that of the non-Guttman option (010; see Figure 1).

Figure 2
Odds of Being in a Higher Response Category Compared to the Category Below


a. Superitem 2

b. Superitem 5

The inconsistency of this item with expectation has already been noted elsewhere (Wilson & Iventosch, 1988, p. 327) in which a dichotomous Rasch analysis was compared to a partial credit analysis to locate the discrepancy. These results indicated that the multistructural question is empirically much easier than the unistructural question. The text of the item suggests that the illustration most near to the unistructural question is misleading—it might be taken to imply that the question is to be answered in the way that the relational question is meant to be answered. This is not the case for the multistructural question. Additionally, the classification of these two questions can be questioned. They appear to be similar, although most students would be expected to be more familiar with a "square box" than a "triangular box." Thus, the ordered partition results for this item suggest that the presentation of the stimulus material needed to be revised and that a problem might exist with the question design.

Application to Problem-Solving Strategies

Strategies and data. Siegler (1987) reported a study in which students were presented a series of elementary addition problems and then were asked "How did you figure out the answer to that problem?" The answers were classified into one of the following five categories according to a scheme based on earlier research:

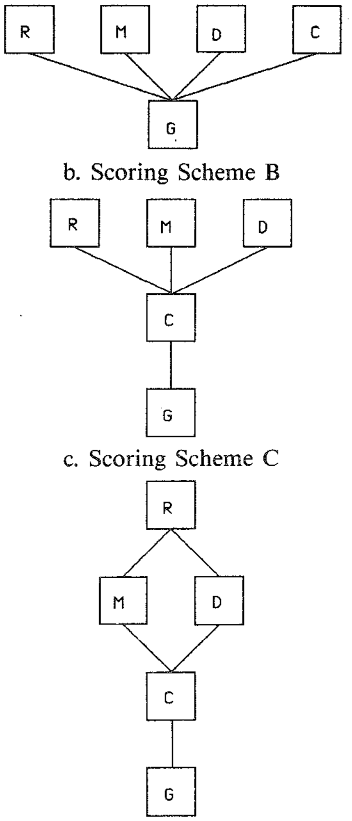
1. Retrieval (R), in which the student retrieves the answer from memory;
2. The Min strategy (M), in which the student counts up from the larger addend the number of times indicated by the smaller addend;
3. Decomposition (D), in which the student transforms the original problem into two or more simpler problems;
4. The Counting-All strategy (C), in which the student counts from one the number of times indicated by the sum; and
5. Guessing and "other" (G), in which the student says that she/he guessed or did not know the answer.

Siegler (1987) concluded (1) that dependent variables such as solution time and error rate should not be "averaged over" these strategies as done in the past, because it has led to some contradictory results in studies of addition; and (2) that students do not use one strategy exclusively, but tend to show substantial variation. His analyses show that some strategies are better than others in the sense that they are quicker and/or are associated with a lower error rate. There is also considerable evidence in the literature for a developmental sequence among the strategies. For example, using a chronometric approach, Ashcraft (1982) found that although first graders are fairly consistent in their use of the Min strategy (M), fourth graders consistently use Retrieval (R), and third graders use a mixture of the two.

Many questions remain about the development of strategy use. These questions may best be answered by considering the strategies one at a time. However, whether the strategies compose a strategy-use continuum remains an interesting question, which could assist in not only the development of strategies but also in finding any regularities in their distribution within and between individuals. For example, it could be hypothesized that any strategy is better than Guessing (G)—this is illustrated as Scoring Scheme A in Figure 3a. A scoring scheme corresponding to this would be dichotomous, with a score of 0 awarded for G and a score of 1 given for the remainder. On the basis of the developmental literature referred to in Siegler (1987), there is a strong tendency for Count-All (C) to develop first. Therefore, Scheme A might be modified to Scheme B, in which the score for G and C remain the same, but the three remaining strategies receive a score of 2. A further modification is suggested by noting that R is usually considered the superior strategy (at least in the class of items used by Siegler), so Scheme B might be modified to Scheme C (Figure 3c) in which R alone receives the highest score. As is often the case, the appropriate scoring scheme is not known a priori. The context usually provides some indication of the most probable alternatives, as in this case, but will seldom provide a definitive resolution. Then, other things being equal, the OPM that provides the better overall fit among these alternatives will have the greater empirical support.

There were 68 students with complete data records, from kindergarten, grade 1, and grade 2. The items ranged from easy (e.g., 4 + 1 = ?) to more difficult (e.g., 17 + 6 = ?). For illustrative purposes, the following subset of the

Figure 3
Scoring Schemes for Siegler's Addition Strategies



a. Scoring Scheme A

b. Scoring Scheme B

c. Scoring Scheme C

original item set is discussed below: (1) 12 + 2, (2) 14 + 1, (3) 3 + 14, (4) 1 + 14, (5) 17 + 4, and (6) 16 + 6. These were combined into three pairs. The first pair was taken from Siegler's (1987) Problem Type 1 in which the larger addend is first, and the smaller addend is relatively small (i.e., from 1 to 3). The second pair was taken from Siegler's Problem Type 2 that is the same as Problem Type 1 except that the larger addend is second. The third pair was taken from Siegler's Problem Type 4 in which the larger addend is first, and the smaller addend is relatively larger (from 4 to 6), which means that the sum is also relatively larger. The three scoring schemes discussed above are provided in Table 5.

Results. Table 6 gives the AIC for each scoring scheme. Note that for these three scoring schemes, the number of item parameters is a constant, but the total number of parameters to be estimated varies because the total number of scores with nonzero frequencies varies from scheme to scheme. Following the argument presented above, it could be expected that if the use of strategies is a developmentally-ordered variable as described above, then the fit should improve moving from Scheme A to Scheme C, and this is indeed the case.

Table 5
Assignment of Responses to Levels for Each of the Scoring Schemes in Example 2

	Scoring Scheme		
Response	A	B	C
G	0	0	0
C	1	1	1
D	1	2	2
M	1	2	2
R	1	2	3

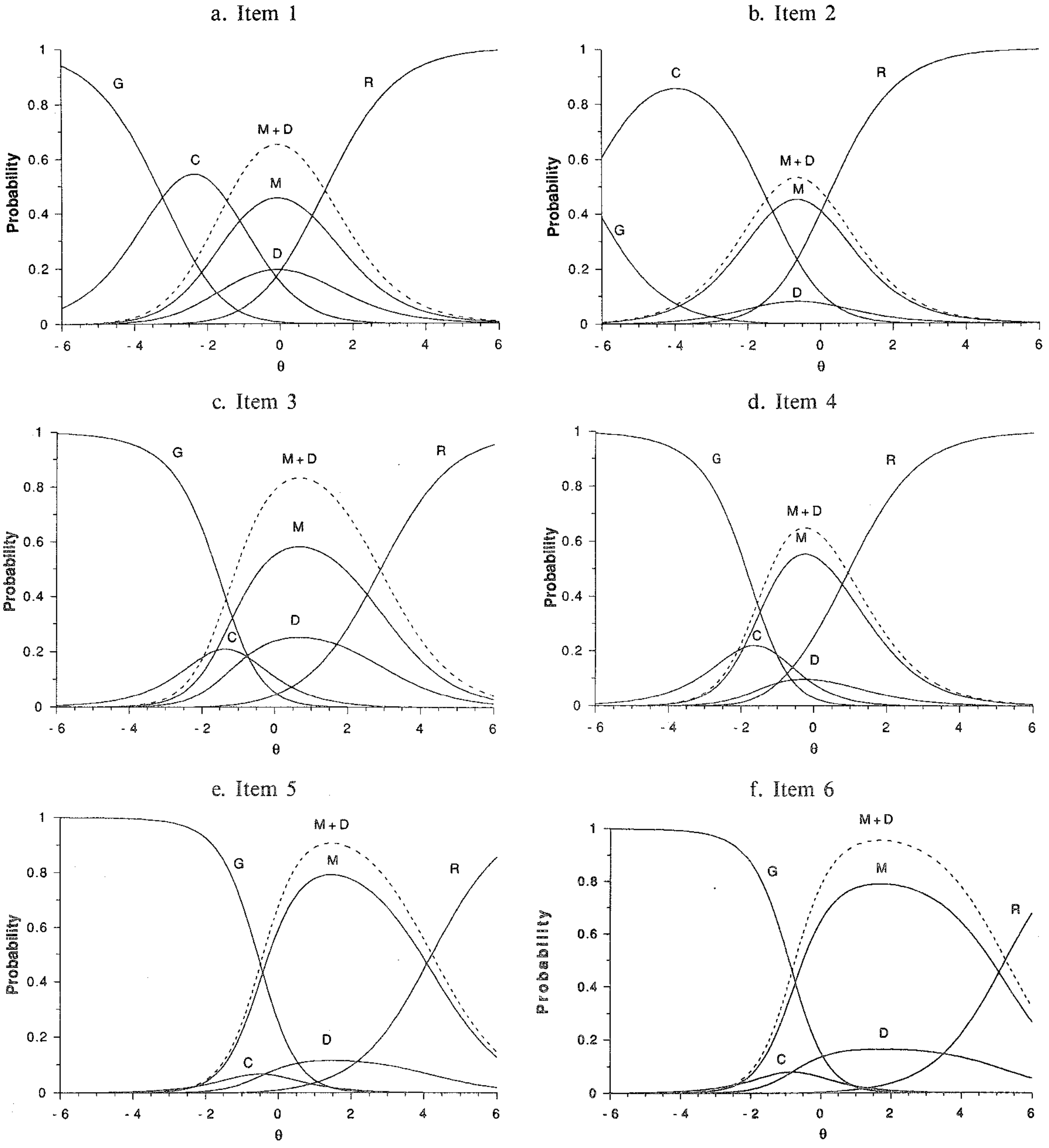
Table 6
Goodness-of-Fit Statistics for the Scoring Schemes in Example 2

Scoring Scheme	Likelihood Ratio G^2	Number of Parameters	AIC
A	423.56	37	497.56
B	345.79	42	429.79
C	324.67	47	418.67

If Scheme C is accepted as appropriate, the item parameter estimates for the corresponding OPM can be examined and interpreted. The modeled probabilities for Item 1 (12 + 2) are shown in Figure 4a (the solid lines in these figures represent the curves for each category, and the dashed curve represents the sum of the two categories that receive a score of 2). The G, C, M + D, and R curves provide the simpler PCM equivalent of the OPM estimates.

The modeled probabilities for Item 2 (14 + 1) are shown in Figure 4b, and those for the remaining four items are shown in Figures 4c–4f. If the θ scale is interpreted as representing the development of strategy use, note that the centers of the curves move to the right from Problem Type 1

Figure 4
Modeled Probabilities of Strategy Use



a. Item 1

b. Item 2

c. Item 3

d. Item 4

e. Item 5

f. Item 6

to Problem Type 4. Note also that, of the two strategies comprising the third strategy level, D is at all times less likely than the M strategy. (This is necessarily a feature of the way that the model has been constructed, given that fewer students overall exhibit D. Deviations from this would be observed as item misfit.) For items in Problem Type 1 (Items 1 and 2, Figures 4a and 4b), G is superseded by the next strategy at an earlier point than for Problem Types 2 and 3, which is true for Item 2 sooner than for Item 1. But there is also another difference between the Problem Type 1 (Figures 4a and 4b) items and the items for Problem Types 2 and 3 (Figures 4c–4f). For the first pair of items, the major successor to G is the C strategy; however, for the remaining items the C strategy is not the most likely—the next most likely strategy is the M strategy.

These qualitative differences arise even though two of the items differed only in their order of presentation of the addends (Items 2 and 4). Comparing the Problem Type 2 items (Figures 4c and 4d) with the Problem Type 4 items (Figures 4e and 4f), the major difference is that, although the point at which M and D succeed G is approximately the same, the point at which R becomes most likely is much later for the items in Problem Type 4; in other words, M and D persist as strategies later for Problem Type 4 than for Problem Type 2.

These differences can be summarized as follows. In the context of the three problem types, the C strategy will not be prominent developmentally if either (1) the first addend is larger than the second, or (2) the smaller addend is larger than 1–3. Also, when the smaller of the two addends is in the range 4–6 rather than 1–3 (or, equivalently, when the sum is in the range 19–23 rather than 13–17), R is developmentally more "difficult."

Two problems per problem type were chosen for this illustrative example because (1) LOGIMO would not estimate all the parameters for a sampling rate of three per problem type (i.e., nine items with four free parameters per item, resulting in 36 item parameters, to which must be added the number of person parameter estimates), and (2) the sample size was clearly too small for this number of parameter estimations. To determine whether the above interpretations generalized beyond these items, parallel analyses were done with 10 other sets of six items drawn randomly from the 27 items comprising the three problem types. The results were essentially the same.

Table 7 provides, for Item 1, the predicted probabilities of being in each of the strategy classes at four developmental points—the minimum observed θ, the average of the kindergarten students, the average of the grade 1 and 2 students, and the maximum observed θ. These proportions clearly display the development of strategy use and, at the same time, depict the diversity of strategy choice at any given location. The pattern shows some interesting symmetries. In comparing the rows for the two most extreme θs (−2.36 and 2.62), the D and M probabilities are almost exactly the same; however, the probabilities for R and those for G and C combined virtually interchange in value, with the majority of the latter going to C. This also can be observed by comparing the category response functions in Figure 4a. The same pattern holds for the two more moderate θs, but with correspondingly larger proportions for D and M and smaller proportions for the rest.

Table 7
Predicted Probabilities of Being in Each of the Strategy Classes for Item 1 at Four Values of θ

	Strategy Classes				
θ	G	C	D	M	R
−2.36	.23	.55	.06	.15	.01
−.71	.03	.30	.18	.42	.08
.80	0.00	.06	.17	.40	.36
2.62	0.00	0.00	.06	.14	.80

These qualitative features echo certain findings that Siegler (1987, p. 255) emphasized: Students, even at the extremes, are not predicted to use just one strategy exclusively. The ordered partition analysis revealed that patterns of results are sensitive to differences in item types. The finding that some types develop later than others is predictable and consistent with Siegler's account. The finding that C is never a predominant strategy for two of the problem types is not evident in Siegler's analysis, because Siegler averaged across problem types. The ordered partition analysis, however, indicated that there are some distinctive patterns in the data that were not clear in the original analysis.

Discussion

The OPM offers a useful way to conceptualize and implement measurement of (1) achievement domains in which there is some underlying structure built into the items (such as SOLO), (2) data arising from cognitive science investigations involving strategy use, and (3) other educational and social science applications in which ordered partitions of the response categories are suggested by substantive theory (such as might arise in performance assessment). Because the OPM fits into the partial credit framework, it allows a straightforward means of integrating a more complicated view of data that arise in measurement contexts with a now well-established item response model for ordered polytomous data.

This analysis has capitalized on the framework allowed by Kelderman's (1984, 1989) loglinear Rasch model. Some of the power of that approach is evident in the relative ease with which alternative scoring schemes could be fitted in the two examples. The loglinear Rasch framework allows for the routine practical application of measurement models that would under other circumstances demand the creation of an entirely new computer program. However, there are some limitations to the number of item parameters that can be estimated with LOGIMO. Therefore, Wilson & Adams (in press) developed a special purpose program for the OPM that decreases the restrictions on the number of parameters, provides more user-friendly output, and more efficient estimation. This program also incorporates special cases of the OPM that are not attainable with the LOGIMO model, such as the imposition of constraints on category parameters between items [similar to the rating scale constraints (Andrich, 1978) that can be placed on the PCM].

The first example could be viewed as a way to use the ordered partition approach to model the role of individual items in measurement situations in which there is expected to be some local dependence among subsets of items, such as the SOLO case in which the items share a common stimulus; therefore, it could be viewed as an improvement on the earlier work by Wilson (1988) in which the results of two analyses, one that used a dichotomous Rasch model and a second that used a PCM, were combined to explore the pattern of local dependency. The ordered partition approach streamlines this by modeling both the subtest response vector and subtest score levels of information in one analysis. This is not quite an integration of the dichotomous item and subtest score level as was attempted by Wilson (1988), but it nevertheless allows for an interpretation that is centered on the item level.

The OPM can be contrasted with a purely latent class analysis for which each response category would be considered separately, but in which no latent trait is postulated. Recently, there have been a number of other approaches described that consider ordered latent classes as a possible way to analyze such data (Dayton & Macready, 1989; Goodman, 1990; Haertel & Hativa, 1986; Yamamoto, 1988). Conceptually, the ideas of ordered latent classes, which introduce a latent order into nominal classes, and the OPM, which introduces latent classes into a latent trait model, can be seen as approaching a common goal. A valuable next step will be to compare these models with the OPM, to explore the relative merits of the two approaches, and possibly to attempt a unification of the different perspectives.

References

Adams, R. J., & Wilson, M. (1992, April). *A random coefficients multinomial logit: Generalizing Rasch models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N.

Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiadó.

Andersen, E. B. (1973). Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

Andersen, E. B. (1983). A general latent structure model for contingency table data. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 117–138). Hillsdale NJ: Erlbaum.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2,* 213–236.

Assessment of Performance Unit. (1980). *Mathematical development: Secondary survey* (Research Rep. No. 1). London: Assessment of Performance Unit, Department of Education and Science.

Baron, J. B., Forgione, P. D., Jr., Rindone, D. A., Kruglanski, H., & Davey, B. (1989, March). *Toward a new generation of student outcome measures: Connecticut's common core of learning assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1991). *Evaluation of procedure-based scoring for hands-on science assessment.* Santa Barbara: University of California, Graduate School of Education.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

California Assessment Project. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento: California State Department of Education.

Collis, K. F., & Davey, H. A. (1986). A technique for evaluating skills in high school science. *Journal of Research in Science Teaching, 23,* 651–663.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25,* 326–346.

Dayton, C. M., & Macready, G. B. (1989, March). *Partial knowledge models.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco.

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* (pp. 95–121). Norwood NJ: Ablex.

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

Gagne, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6,* 1–9.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing* (Vol. 3) (pp. 41–85). Hillsdale NJ: Erlbaum.

Goodman, L. A. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional tables, with specified and/or unspecified order for response categories. In C. C. Clogg (Ed.), *Sociological Methodology* (pp. 249–294). San Francisco: Jossey-Bass.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method for scale construction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 319–348). New York: Social Science Research Council.

Haertel, E., & Hativa, N. (1986). *Measuring stages of concept attainment in geometry: Latent class analysis of the van Hiele levels* (Research Report). Stanford CA: Department of Education, Stanford University.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223–245.

Kelderman, H. (1989, April). *Loglinear multidimensional IRT models for polytomously scored items.* Paper presented at the Fifth International Objective Measurement Workshop, Berkeley CA.

Kelderman, H., & Steen, R. (1988). *LOGIMO: A program for loglinear IRT modeling* [Computer program]. Enschede, The Netherlands: Department of Education, University of Twente.

Kemeny, J., Snell, J., & Thompson, G. (1957). *Introduction to finite mathematics.* Englewood Cliffs NJ: Prentice-Hall.

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

Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–29). New York: Plenum.

Piaget, J. (1950). *The psychology of intelligence* (M. Piercy & D. E. Berlyne, Trans.). New York: Harcourt Brace Jovanovich.

Romberg, T. A., Collis, K. F., Donovan, B. F., Buchanan, A. E., & Romberg, M. N. (1982). *The development of mathematical problem solving superitems* (Report of NIE/ECS Item Development

Project). Madison: Wisconsin Center for Educational Research.

Romberg, T. A., Jurdak, M. E., Collis, K. F., & Buchanan, A. E. (1982). *Construct validity of a set of mathematical superitems* (Report of NIE/ECS Item Development Project). Madison: Wisconsin Center for Educational Research.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 49,* 501–519.

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

Werner, H. (1957). The concept of development from a comparative and organismic point of view. In D. B. Harris (Ed.), *The concept of development* (pp. 125–148). Minneapolis: University of Minnesota Press.

Wiggins, G. (1989, May). A true test: Toward more authentic assessment. *Phi Delta Kappan, 71,* 703–713.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12,* 353–364.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education, 32,* 125–138.

Wilson, M. (1990). Measuring a van Hiele geometry sequence: A reanalysis. *Journal for Research in Mathematics Education, 21,* 230–237.

Wilson, M., & Adams, R. J. (in press). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational Statistics.*

Wilson, M., & Iventosch, L. (1988). Using the partial credit model to investigate responses to structures subtests. *Applied Measurement in Education, 1,* 319–334.

Wilson, M., & Masters, G. N. (in press). The partial credit model and null categories. *Psychometrika.*

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

Yamamoto, K. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation, University of Illinois, Champaign.

Acknowledgments

The author thanks Henk Kelderman for supplying the program that was used to analyze the data from the examples. Earlier versions of this paper were presented at the Fifth International Objective Measurement Workshop, Berkeley CA, March 1989, and at the annual meeting of the National Council for Measurement in Education, Boston MA, April 1990. Preparation of sections of this paper was supported by a National Academy of Education Spencer Fellowship.

Author's Address

Send requests for reprints or further information to Mark Wilson, Graduate School of Education, Berkeley CA 94720, U.S.A.