A Likert-type Data Analysis Using the Partial Credit Model

Sun-Geun Baek

Korean Educational Development Institute

This study is about examining the possibility of using the partial credit model to solve several problems that occur when we analyze and interpret Likert-type data by traditional methods. The problems are as follows: (i) scores are not directly interpretable and must be examined in the light of a criterion group; (ii) the absence of a zero point handicaps the direct use of individual scores; and (iii) the adequacy of integer scoring, resting upon the validity of the assumption of equal distances between response categories, is not often verified. This study shows that the partial credit model (PCM) solves these problems.

In addition, the PCM provides several advantages in the analysis and interpretation for Likert-type data (e.g., item response maps, person and item fit statistics). The PCM also might help to implement the computerized adaptive testing for Likert-type scales.

I. Introduction

Measurement in the area of aptitude and achievement deals with the cognitive characteristic of human behavior and hence must be relatively sophisticated. Most investigations, however, are related to noncognitive characteristics and involve the use of questionnaires, judges, ratings, self-report ratings, interviews, and similar procedures (Gamache, 1983). Clogg (1979), for example, reports that approximately one half of all recorded observations in the 1975 General Social Survey used a Likert-type response format. Questionnaires with ordered response categories are common in psychological, educational, and social research. Sets of questions are developed to measure underlying characteristics such as fear of crime, attitude to drugs, or liking school. In these situations the intention is to combine an

individual's responses to a number of different questions to obtain a measure of that person's standing on the single latent characteristic that the questions are intended to define (Masters, 1985).

Gamache (1983) examined whether scales constructed under procedures and criteria outlined by the various traditional and latent trait methods are varied in characteristics related to scale quality (e.g., coefficient Alpha, scale validity, score equivalence, etc.). Scales were constructed from a common pool of items analyzed in the polychotomous form according to Likert and the partial credit models, and analyzed in a dichotomous form for the Guttman, two-parameter Birnbaum, and one-parameter Rasch models. According to the study, a traditional method based on item to total score correlation produced a slightly more valid scale. All five method-defined scales, however, were remarkably similar in other characteristics related to scale quality.

The problem of the traditional method might be not in the scale construction but in the analysis and interpretation. Likert (1932) assigned successive integers to response categories and simply summed the items to obtain a questionnaire score for each respondent. However, this approach has been criticized on the grounds that it assumes equal differences between adjacent response categories. Although a variety of alternatives to integer scoring based on systems of empirically-derived weights have been proposed, these more complicated schemes have invariably proven no more useful in practice than integer scoring (Wang & Stanley, 1970). As a result, many standardized attitude and personality instruments have reverted to using the simpler integral weights.

With the Likert approach there are several problems as follows: (i) scores are not directly interpretable and must be examined in the light of a criterion group, (ii) the absence of a zero point handicaps the direct use of individual scores, and (iii) the adequacy of integer scoring, resting upon the validity of the assumption of equal distances between response categories, is not often verified (Gamache, 1983). To overcome the these problems, forms of item response models have been applied to a rating scale or ordinal data (Andrich, 1978a, 1978b, 1978c, 1982; Masters, 1985; Rost, 1985, 1988; Wright & Masters,

1982). Item response models are claimed to share the unique potential, when the data fit the models, that item parameters are estimated independently of the calibration sample and that both measurement of persons and analysis of items is freed from the specific set of items and persons used for calibration.

In the 1950's, Georg Rasch, introduced and used a measurement model for dichotomously-scored performances. This model, which is often referred to as 'the Rasch model', has been widely applied to the analysis of educational test data and to the construction and maintenance of item banks. A simple extension of right/wrong scoring is to identify one or more intermediate levels of performance on an item and to award partial credit for reaching these intermediate levels.

For an item on a Likert-type attitude questionnaire (e.g., culture shock questionnaire (Baek, 1991)), 'completing the j'th step' can be thought of as choosing the j'th alternative over the (j - 1)'th in response to the item. Thus a person who chooses 'Moderate Difficulty' with a statement on a culture shock questionnaire when given the ordered categories shown in Figure 1 to choose among, can be considered to have chosen 'Slight Difficulty' over 'No Difficulty' (first step taken) and also 'Moderate Difficulty' over 'Slight Difficulty' (second step taken), but to have failed to choose 'Great Difficulty' over 'Moderate Difficulty' (third step not taken).

The relative difficulties of the 'steps' in a Likert-type scale item are usually intended to be governed by the fixed set of rating points accompanying the items (Andrich, 1978a,1978b, 1978c, 1982). As the same set of rating points is used with every item, it is usually thought that the relative difficulties of the steps in each item should not vary from item to item. This expectation can be incorporated into the extended Rasch model by resolving each item step into two components so that

$$\delta_{ij} = \delta_i + \tau_j$$

Figure 1. A Typical Likert-type Scale

No Difficulty	Slight Difficulty	Moderate Difficulty	Great Difficulty	Extreme Difficulty
0	1	2	3	4

where δ_i is the location or 'scale value' of item i on the variable and τ_j is the location of the j'th step in each item relative to that item's scale value. The extended Rasch model with $\delta_{ij} = \delta_i + \tau_j$ by Andrich (1978a) is called the rating scale model (RSM). The RSM is defined as follows:

$$\Pi nix = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_j)]}$$

$$x = 0,1,\cdots, \ m \ \text{ where } \ \tau_0 \equiv 0 \ \text{ so that } \ \exp \sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 1.$$

The Πnix is the probability of person n responding in category x to item i, β is a person's level of ability in a given area, and there are m categories in the item.

Although we ideally analyze the Likert-type data using the RSM, there is at least an unsolved problem: The assumption that the relative difficulties of the steps in each item are the same is not often verified. In addition, the RSM is, not appropriate for the Likert-type data used in this study (see the detailed results section in this paper). However, it is possible to solve this problem. Masters (1982) has extended the Andrich's RSM to the situation where category alternatives are free to vary in number and structure from item to item, the so-called the partial credit model (PCM). The PCM does not impose any particular expectation of the pattern of difficulties of steps within each item. Although the PCM is more general than the RSM, it is a more parsimonious model than the other models (e. g., the graded response model (Samejima, 1969)). The PCM contains one parameter for each person and one parameter for each 'step' in an item. Consider an item with five ordered levels, 0, 1, 2, 3, and 4, provides four steps. For such an item, four parameters are estimated. First, δ_{i1}, governs the model probability of scoring 1 rather than 0. Second, δ_{i2}, governs the model probability of scoring 2 rather than 1. Third, δ_{i3}, governs the model probability of scoring 3 rather than 2. Fourth, δ_{i4}, governs the model probability of scoring 4 rather than 3. The PCM with δ_{ij} is defined as follows:

$$\Pi nix = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}$$

$x = 0,1,\cdots, m_i$ where $\delta_{io} \equiv 0$ so that $\exp \sum_{j=0}^{0} (\beta_n - \delta_{ij}) = 1.$

The Πnix is the probability of person n responding in category x to item i, β is a person's level of ability in a given area, and δ_{ij} is a parameter that governs the probability of a response being made in category x rather than in category x-1 of item i.

This study examines the possibility of using the PCM to solve three problems in Likert-type data analysis and interpretation: (i) scores are not directly interpretable and must be examined in the light of a criterion group, (ii) the absence of a zero point handicaps the direct use of individual scores, and (iii) the adequacy of integer scoring, resting upon the validity of the assumption of equal distances between response categories, is not often verified.

II. Method

Subjects: The number of subjects of the study is 104 Korean students at U.C. Berkeley (males= 89, females= 15). They were randomly selected. The mean age of the subjects was 27.33 years (range: 20 - 41 years).

Data: Seven data items are selected from a study (Baek, 1991) using five-category Likert-type culture shock questionnaire items (see Table 1).

Analysis: The data are analyzed according to the partial credit model by a computer program: TITAN (Adams & Toon, 1991). The questions to be answered are: (1) Where is the item i located on the variable? (the item's calibration δ_{ij}); (2) How precise is this calibration? (the modeled error of calibration si); (3) How well do responses to item i fit the expectations of the measurement model? (the item's fit ti); (4) Where is person n located on the variable? (the person's measure δ_n); (5) How precise is this measure? (the modeled error of measurement sn); and, (6) How

Table 1. 7 Items' Contents

Items	Degree of difficulty*				
1. Using facilities (e.g., school store, lounge, rest room).	0	1	2	3	4
2. Using educational facilities (e.g. laboratory, computer room, studio).	0	1	2	3	4
3. Making friends of the same sex and of another nationality in the school.	0	1	2	3	4
4. Making friends of the opposite sex and of the same nationality in the school.	0	1	2	3	4
5. Making friends of the opposite sex and of another nationality in the school.	0	1	2	3	4
6. Joining in circles, clubs or associations in the school.	0	1	2	3	4
7. Joining in school events.	0	1	2	3	4

*0-no difficulty, 1-slight difficulty, 2-moderate difficulty, 3-great difficulty, 4-extreme difficulty.

well do responses of person *n* fit the expectations of the model? (the person's fit *tn*).

If item's fit *ti* and person's fit *tn* are good enough, the PCM will be a candidate to solve those problems in Likert-type data analysis and interpretation.

III. Results

Table 2 shows the item fit statistics using the partial credit model. Each item's calibration δ_i (the logit scale value), the modeled error of calibration *si* , and the item's fit *ti* are described. It shows all 7 items fit the expectations of the measurement model ($-2 < ti < 2$). In contrast, the item fit statistics using the rating scale model and shows two items (item 3 and 4) do not fit the the expectations of the measurement model ($ti = -3.5$ and 2.3 respectively). It implies the PCM is more appropriate for this data than the RSM.

Table 3 shows the person fit statistics using the partial credit

Table 2. Item Fit Statistics Using the Partial Credit Model.

Item	Item's Calibreation d_i (s_i)				Item Fit
	$d1$	$d2$	$d3$	$d4$	t_i-values
1	.26(.27)	1.21(.37)	2.14(.55)		0.6
2	–.29(.27)	.46(.29)	1.98(.51)		1.6
3	–2.36(.42)	–.19(.27)	1.07(.33)	2.56(.76)	-1.5
4	–1.56(.34)	–.05(.27)	–.04(.28)	.77(.34)	0.3
5	–2.85(.55)	–.46(.29)	–.52(.27)	.75(.31)	–0.9
6	–2.11(.48)	–1.19(.31)	–.16(.27)	.83(.32)	–0.5
7	–1.63(.42)	–1.23(.31)	–.18(.27)	2.80(.62)	–0.5

model. All 104 person's measure β_n (the logit scale value), the modeled error of measurement *sn*, and the person's fit *tn* are described. It shows 88% of the subjects (91 persons) fit within the 95% expected range of the model ($-2 < tn < 2$) and 98% of the subjects fit within the 99% expected range ($-2.7 < tn < 2.7$). This is not perfect, but it is considered an acceptable level for the PCM.

Since item's fit *ti* and person's fit *tn* are acceptable, the PCM is a candidate to solve the problems in Likert-type data analysis and interpretation. Because the probability of person n reaching difficulty level k in item i depends on that person's and that item's parameters by the PCM, (i) scores are directly interpretable without a criterion group, (ii) items and persons are measured on an interval scale (a logit scale), so differences have a zero, and (iii) it has no assumption about equal distances between response categories. For instance, using the PCM, person parameters can be calculated on a logit scale that is an interval scale. If student attitude level (β) by the PCM equals 0 in logit scale, then we can interpret, without a criterion group, that his/her attitude level (β) is in the middle of the trait distribution. Also, because each step difficulty parameter for each item can be calculated on the logit scale, the assumption of equal distances between response categories is not applied.

In addition, the PCM provides several advantages in the analysis and interpretation for Likert-type data. The PCM provides a framework for assessing the validity of attempting to summarize a trait level (β) on the basis of different aspects in a single global measure. The PCM is able to construct several 'maps' to show the relationships between a person's parameter

Table 3. Person Fit Statistics Using the Partial Credit Model

ID#	$b_n(s_n)$	t_n	ID#	$b_n(s_n)$	t_n	ID#	$b_n(s_n)$	t_n
1	-.57(.43)	-1.3	36	-1.43(.50)	-.9	71	.34(.44)	-1.0
2	.16(.43)	-.3	37	-.03(.43)	-.9	72	1.19(.49)	-.1
3	-.57(.43)	-1.3	38	.74(.46)	.7	73	.34(.44)	.7
4	-.57(.43)	2.3	39	-1.19(.48)	-1.4	74	-1.19(.48)	-1.8
5	.54(.45)	-.1	40	-.03(.43)	-1.4	75	-.57(.43)	-.9
6	.34(.44)	-.9	41	-1.68(.72)	-1.4	76	-1.19(.48)	.2
7	-.03(.43)	-1.1	42	.34(.44)	-.4	77	.16(.43)	-.2
8	.96(.47)	1.0	43	-.97(.46)	-1.0	78	1.69(.51)	-.4
9	.96(.47)	1.0	44	1.69(.51)	-1.6	79	-.03(.43)	1.0
10	.16(.43)	-.7	45	-.57(.43)	-1.3	80	-.57(.43)	2.2
11	-.57(.43)	1.6	46	.16(.43)	.9	81	-.57(.43)	-2.0
12	-.57(.43)	-1.4	47	.74(.46)	-.3	82	-.76(.44)	2.7
13	-.57(.43)	-1.4	48	-.39(.43)	-1.7	83	-.21(.42)	-.7
14	1.19(.49)	-1.2	49	.96(.47)	.8	84	-2.40(.66)	-.5
15	1.19(.49)	-1.2	50	.54(.45)	.4	85	.34(.44)	-2.3
16	.16(.43)	-.7	51	.54(.45)	.8	86	.34(.44)	-1.0
17	.34(.44)	.8	52	-.21(.42)	-1.1	87	-.03(.43)	1.8
18	-.97(.46)	1.1	53	-2.40(.66)	.3	88	.16(.43)	-.9
19	-.57(.43)	-.3	54	-1.43(.50)	2.4	89	.16(.43)	-1.1
20	-2.01(.59)	-.2	55	.74(.46)	.6	90	.74(.46)	.0
21	-.97(.46)	.9	56	1.69(.51)	-1.6	91	-1.70(.54)	-2.6

Table 3. (Continued)

ID#	$b_n(s_n)$	t_n	ID#	$b_n(s_n)$	t_n	ID#	$b_n(s_n)$	t_n
22	1.19(.49)	–.5	57	–.57(.43)	.5	92	.34(.44)	–.5
23	–1.19(.48)	–1.0	58	–.57(.43)	–.0	93	–1.70(.54)	–.7
24	.54(.45)	–1.3	59	.34(.44)	–1.0	94	–.21(.42)	–1.4
25	–.21(.42)	–.8	60	–1.19(.48)	1.4	95	–.97(.46)	–.8
26	–1.43(.50)	–1.7	61	–2.01(.59)	1.3	96	–1.43(.50)	–1.6
27	–.76(.44)	–1.2	62	–1.19(.48)	.7	97	–.21(.42)	2.7
28	–1.70(.54)	–2.6	63	–.76(.44)	2.3	98	1.19(.49)	4.2
29	1.44(.50)	–1.0	64	–1.70(.54)	1.8	99	.16(.43)	–.4
30	–.21(.42)	1.1	65	.34(.44)	–2.3	100	–2.40(.66)	–1.1
31	–.03(.43)	–1.1	66	–.39(.43)	1.6	101	–.97(.46)	4.0
32	–.57(.43)	1.0	67	–.76(.44)	.1	102	–.76(.44)	–.6
33	.34(.44)	–.9	68	–.39(.43)	–2.4	103	–.03(.43)	–1.5
34	–.21(.42)	.1	69	1.69(.51)	–1.6	104	–1.19(.48)	–1.8
35	–.39(.43)	–.2	70	.16(.43)	–.2			

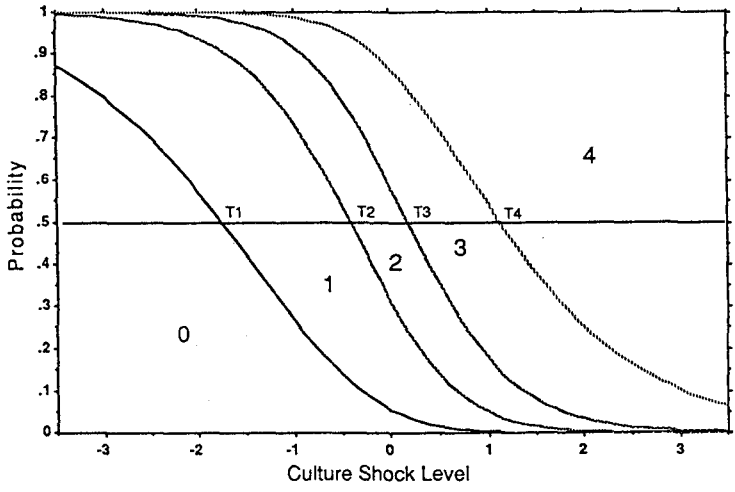

Figure 2. Item Response Map for Item 4.

and probability (see Figure 2), and between a person's parameter and item thresholds (see Table 4 and Figure 3). These maps help one to understand and to interpret those relationships (Masters & Wilson, 1991; Wilson, 1990).

Figure 2 shows how the probability of a student's response in each category of item 4 (Making friends of the opposite sex and of another nationality in the school) changes with increasing 'culture shock level'. For example, from Figure 2 we can see that a student with an estimated 'culture shock level' of 0.0 logits (middle of the picture) has an estimated model probability of about 0.05 of scoring 0 (no difficulty) on item 4; 0.26 of scoring 1 (slight difficulty); 0.27 of scoring 2 (moderate difficulty); 0.28 of scoring 3 (great difficulty); 0.13 of scoring 4 (extreme difficulty). The relative values of these model probabilities change with the changing 'culture shock level'. As the 'culture shock level', for example, increases above this level, the estimated model probability of scoring 4 (extreme difficulty) increases.

In addition, Figure 2 shows the 'thresholds (T)' of item 4. The 'thresholds' that are analogous to Thurstonian thresholds provide a way to summarize information about several partial credit items. A 'threshold' for an item step is the 'culture shock level' that is required for an individual to have a 50% probability of choosing that level. The 'thresholds (T)' can be interpreted as

Table 4. Items' Thresholds.

Item	Thresholds			
	T1	T2	T3	T4
1	-.08	1.24	2.20	
2	-.66	.50	2.06	
3	-2.50	-.38	1.02	2.64
4	-1.80	-.47	.12	1.05
5	-2.97	-.97	-.28	.92
6	-2.42	-1.23	-.20	1.03
7	-2.09	-1.17	-.03	2.75

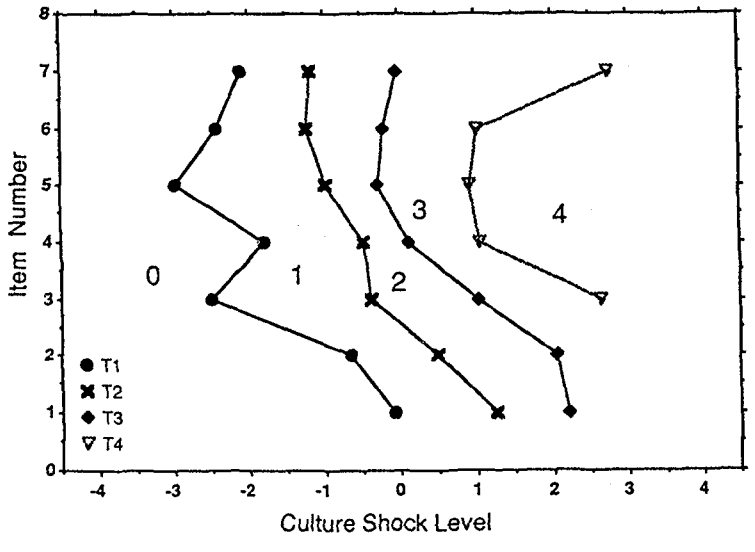

Figure 3. A Summary Item Response Map for 7 Items.

the crest of a wave of predominance of successive dichotomous segments of the set of categories.

For example, T1 is the estimated point at which categories 1, 2, 3, and 4 become more likely than category 0, T2 is the estimated point at which categories 2, 3, and 4 become more likely than categories 0 and 1, T3 is the estimated point at which categories 3 and 4 become more likely than categories 0, 1, and 2, and T4 is the estimated point at which category 4 become more likely than categories 0, 1, 2, and 3.

Table 4 and Figure 3 show the 'thresholds (T)' of all 7 items.

From Figure 3, the students have less difficulty with items 1 and 2 than with items 3, 4, 5, 6, and 7. For example, a student with an estimated 'culture shock level' of 2.0 logits will more likely respond in category 2 (moderate difficulty), 1 (slight difficulty), and 0 (no difficulty) on items 1 and 2 than in categories 3 (great difficulty) and 4 (extreme difficulty). In contrast, the student will more likely respond in category 4 on items 4, 5, and 6 than in categories 0, 1, 2, and 3 and will more likely respond in categories 4 and 3 on items 3 and 7 than in categories 0, 1, and 2.

For another example, a student with an estimated 'culture shock level' of –1.0 logits will more likely respond in category 0 on items 1 and 2 than in categories 1, 2, 3, and 4. In contrast, the student will more likely respond in categories 1 and 0 on items 3, 4, and 5 than in categories 2, 3, and 4, and will more likely respond in categories 2, 1, and 0 on items 6 and 7 than in categories 3 and 4. In addition, we can see how students' response to each item changes with increasing 'culture shock level'.

From the Figure 3, we can see that the distances among the 'thresholds (T)' of each item are different. It means there are differences in changes in students' response categories for each item with increasing 'culture shock level'. For example, item 4's distances among Thurstonian thresholds are relatively smaller than those of item 3. That is, according to increasing 'culture shock level', students' response categories to item 4 (Making friends of the opposite sex and of another nationality in the school) more rapidly change than those to item 3 (Making friends of the same sex and of another nationality in the school).

The PCM, furthermore, helps to identify particular sources of misfit (e.g., particular misfitting items or persons). In case of item misfit, for example, the misfitting items do not work together well enough to be treated as indicators of a latent variable estimated by the PCM. In this study, there is no misfit item among the seven analyzed items. In case of person misfit, for example, it helps to find easily individual differences in response style. If *tn* is less than –2, it means the person has an unusually regular pattern of responses - more regular than the PCM expects for persons responding to items. The standardized residuals for this person are very close to zero (e.g., student #

28's culture shock level (β) = –1.70, tn = –2.6, and the observed responses are 0, 0, 1, 1, 1, 1, 1, and the modeled expected responses are the same).

In contrast, if tn is more than 2, it means the person has an erratic pattern of responses. The standardized residuals for this person are relatively large (e.g., student # 98's culture shock level (β) = 1.19, tn = 4.2, and the observed responses are 4, 3, 4, 4, 1, 3, 2, but the modeled expected responses are 1, 2, 3, 4, 4, 4, 3). Furthermore, these person-fit statistics will be used by counselors. For example, student # 98 had abnormally extreme difficulty in item 1 'Using facilities' (e.g., school store, lounge, rest room). Information like this could be used by counselors to help students individually overcome personal difficulties.

IV. Discussion

Testing is a very important instrument for optimum decision-making in education as well as in society. Questionnaires with ordered response categories are common in psychological, educational, and social researche. There are benefits from examining the Likert-type data analysis using the partial credit model.

First of all, the partial credit model will be a candidate to solve these problems in Likert-type data analysis and interpretation if item's fit ti and person's fit tn of the Likert-type data are acceptable. If the probability of person n reaching performance (or attitude) level k in item i depends on that person's and that item's parameters according to the PCM, (i) scores are directly interpretable without a criterion group, (ii) items and persons are measured on an interval scale (a logit scale), so differences have a zero, and (iii) it has no assumption about equal distances between response categories.

Second, the PCM provides several advantages in the analysis and interpretation for Likert-type data. The PCM is able to construct several 'maps' to show the relationships between a person's parameter and probability, and between a person's parameter and item thresholds. These maps help one to understand and to interpret those relationships. The PCM, furthermore, helps to identify particular sources of misfit (e.g.,

particular misfitting items or persons).

The PCM also might help to implement the computerized adaptive testing (CAT) for Likert-type scales (Baek, 1993; Koch & Dodd, 1989). With the item information using each item's calibration δ_{ij} and the person's background information (e.g., gender, academic status, marital status.), a CAT might provide the optimal item to a person in order to estimate his attitude level up to the desirable accuracy that is decided before he takes the questionnaire. It might reduce not only the number of items for the questionnaire but also the measurement error.

References

Adams, R.J. & Khoo, S.T. (1991) *TITAN: The Interactive Test Analysis System*, Hawthorn, Australia: ACER.

Andrich, D. (1978a). "Scaling Attitude Items Constructed and Scored in the Likert Tradition," *Educational and Psychological Measurement*. 38, 665-680.

———— (1978b). "Application of a Psychometric Rating Model to Ordered Categories which are Scored with Successive Integers," *Applied Psychological Measurement*. 2 (4), 581-594.

———— (1978c). "A Rating Formulation for Ordered Response Categories," *Psychometrika*, 43(4), 561-573.

———— (1982). "An Extension of the Rasch Model for Ratings Providing both Location and Dispersion Parameters, " *Psychometrika*, 47(1), 105-113.

Baek, S.G. (1993). *Computerized Adaptive Attitude Testing Using the Partial Credit Model*, Doctoral dissertation, University of California at Berkeley.

———— (1991). *The Culture Shock of Korean Students at U. C. Berkeley*, (unpublished).

BMDP. (1990). *BMDP Statistical Software Manual*, 1, University of California Press.

Gamache, L.M. (1983). *Comparison of Traditional and Latent Trait Procedures in Analysis and Selection of Rating Scale Items* (ERIC. ED230578).

Hambleton, R.K. (1989). "Principles and Selected Applications of Item Response Theory," In R.L. Linn (Ed.), *Educational Measurement*. New York: American Council on Education, Macmillan.

————, & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

———— (Ed.). (1983). *Applications of Item Response Theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Kock, W.R. & Dodd, B.G. (1989). "An Investigation of Procedures for Computerized Adaptive Testing Using Partial Credit Scoring," *Applied Measurement in Education,* 2(4), 335-57.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ.: Erlbaum.

Masters, G.N. (1985). "A Comparison of Latent Trait and Latent Class Analysis of Likert-type Data," *Psychometrika,* 50(1), 69-82.

————, & Wilson, M. (1991). "The Measurement of Understanding," Paper presented at a Workshop on Partial Credit Model, ETS, Princeton, N.J.

Rost, J. (1985). "A Latent Class Model for Rating Data," *Psychometrika,* 50(1), 37-49.

———— (1988). "Rating Scale Analysis with Latent Class Models," *Psychometrika,* 53(3), 327-348.

Samejima, F. (1969). "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometrika,* [Monograph], 34(Suppl. 4).

Wang, M.W. & Stanley, J.C. (1970). "Differential Weighting: A Review of Methods and Empirical Studies," *Review of Educational Research,* 40, 663-705.

Wilson, M. & Mislevy, R.J. (1989). "Test Theory for Measuring Understanding and Learning," Paper presented at the ETS International Symposium on Language Acquisition and Language Assessment. Princeton, N.J.

———— (1991). (in press), *The Partial Order Model: An Extension of the Partial Credit Model..*

———— (1990). *Measuring Levels of Mathematical Understanding,* (unpublished).

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