DOCUMENT RESUME

ED 398 270 TM 025 400

AUTHOR Lam, Peter; Foong, Yoke-Yeen
TITLE Calibrating Attitude Scale with Negatively Worded Items Using PARELLA and Rating Scale Models.
PUB DATE [96]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attitude Measures; Foreign Countries; Internship Programs; Item Response Theory; *Likert Scales; *Rating Scales; Scaling; *Student Teachers; Test Construction; *Test Items; Test Theory
IDENTIFIERS *Calibration; *PARELLA Model

ABSTRACT

An important principle in constructing rating scales is to develop items that reflect various degrees of the "pro" (positive) and "contra" (negative) aspects of the trait being measured. Where both positive and negative items are pooled, they can be arranged in order along the trait continuum, but for classical and item response theory analysis, scores for negatively worded items will have to be reversed in the Likert tradition. The data for this study of calibrating item scales came from 350 first-year teacher interns after their first internship who responded to 10 career commitment statements based on a 5-point Likert scale. The dataset was fitted into the Rating Scale Model (D. Andrich, 1978), and after dichotomizing, fitted into the PARELLA model (H. Hoijtink, 1991). Results showed that the Rating Scale model was able to arrange items 7 through 10 of the scale in the order as intended, but there were problems with the negatively worded items. The PARELLA model, however, was able to align the items correctly. On the other hand, the step estimates from the Rating Scale model give additional information on the way in which respondents indicate their level of agreement with the statements. Although the binary conversion of the Likert scale may result in a loss of information, the PARELLA model can provide information on item location and person separation of the trait. It is recommended that separate analysis of positive and negative items on Likert scales can be made using the Rating Scale model with the PARELLA model as a complement in establishing proximity items with item scale order. (Contains 4 tables and 12 references.) (Author/SLD)

CALIBRATING ATTITUDE SCALE WITH NEGATIVELY WORDED ITEMS USING PARELLA AND RATING SCALE MODELS

Main Author : Peter Lam
National Institute of Education
Center for Educational Research

469 Bukit Timah Road
S259756 SINGAPORE

Coauthor : Yoke-Yeen Foong
School of Science
National Institute of Education

469 Bukit Timah Road
S259756 SINGAPORE

Suggested running head: IRT, PARELLA and rating scale models, attitude scale calibration

Key words: PARELLA, Rating Scale Model, Attitude scale calibration, IRT

CALIBRATING ATTITUDE SCALE WITH NEGATIVELY WORDED ITEMS USING PARELLA AND RATING SCALE MODELS

Abstract

One important principle of rating scale construction is to develop items that reflect various degrees of the 'pro' (positive) and 'contra' (negative) aspect of the trait. Where both negative and positive items are pooled, the items can be arranged in order along the trait continuum based on the judgement of the test constructor. For classical and IRT (Rating Scale) analyses, scores for negatively worded items will have to be reversed in the Likert tradition. The data came from 350 teacher interns who responded to 10 Career Commitment statements based on the 5-point Likert scale (1=Strongly disagree, 5=Strongly Agree). The dataset was fitted into the Rating Scale Model. After dichotomizing (1,2,3=0, 4,5=1), the dataset was fitted into the PARELLA model. Results showed that the Rating Scale model was able to arrange Items 7-10 of the scale in the order as intended but there were problems with the negatively worded items. However the PARELLA Model was able to align the items correctly. On the other hand, the step estimates from the Rating Scale model gives additional information to the way in which respondents indicate their level of agreement to the statements. Although binary conversion of the Likert scale may result in a loss of information, the PARELLA model can provide information on item location and person separation on the trait. The study recommends that for Likert type scales, separate analyses of positive and negative items using the Rating Scale model can be used while the PARELLA Model is a complement in establishing proximity items with item scale order.

CALIBRATING ATTITUDE SCALE WITH NEGATIVELY WORDED ITEMS USING PARELLA AND RATING SCALE MODELS

PETER LAM AND YOKE-YEEN FOONG
National Institute of Education, Singapore

An important aspect of rating or Likert scale construction in the measurement of attitude is to treat the attitude as a latent trait ranging from 'contra' to 'pro' (Hoitjink, Molenaar & Post, 1994). Traditionally, the development of such 'negative' and 'positive' statements pertaining to the trait was done to minimize the occurrence of a response or acquiescence set (e.g. Mehrens & Lehmann, 1984). On the psychometric perspective, the locations of the set of items with different degrees of 'contra' and 'pro' on the trait are indicative of 'more' or 'less' of the trait. Lam and Stevens (1994) used the term 'leniency' to describe the way in which an item can be written that reflects the positive end of the scale. The term, 'stringency' was used to reflect items that are placed on the more negative end of the scale. Responses to these items are affected by the way the intensity of wording is used (Lam & Stevens, 1994). A negatively worded or stringent item can be made more negative by using intense wording such as: 'I hate the teachers in this school' compared to: 'I do not like the teachers in this school'. In this study, the content of the wording, rather than its intensity is used to reflect the location of the items on the trait. In the Teacher Intern Commitment Scale designed for this study, the statement, 'I had considered leaving the teaching

---

Paper presented at the Annual Meeting of the American Educational Research Association, New York, U.S.A. (April 8-12, 1996).

profession' will be more negative compared to the statement, 'I did not like the long hours of teaching preparation. The statement, 'I am motivated to work in the teaching profession' is positive but the statement, 'If I am rich enough without working, I would still work in the teaching profession' is more positive.

The scoring of a Likert scale involves assigning numerical weights (e.g. Strongly Disagree = 1, Strongly Agree = 5) for each position on the scale. Negative statements have their weights reversed. A person's total score is the sum of the scores on all items, with the higher score indicating a favorable attitude. The classical model of error variance involving item analytic techniques and item-total score correlations are traditionally used.

Development of polychotomous IRT models has opened the possibilities of such applications in Likert scales. For items where responses are scored using more than two ordered categories to represent varying degrees of the trait, the Graded Response Model (Samejima, 1969) and the Partial Credit Model (Masters, 1982) have been shown to be appropriate.

The Rating Scale Model (Andrich, 1978) is an extension of the Rasch Model for items with ordered categories to reflect varying degrees of the attitude level. In this model, a location parameter (scale value) which indicates the location of the item on the attitude continuum is estimated together with a set of response thresholds. The Rating Scale Model has been shown to be a special case of the Partial Credit Model (Wright & Masters, 1982). The probability of responding in a given category is:

$$P_{x_i}(\theta) = \frac{\exp\{\sum_{j=0}^{x_i} [\theta - (\delta_i + \tau_j)]\}}{\sum_{k=0}^{m_i} \exp\{\sum_{j=0}^{k} [\theta - (\delta_i + \tau_j)]\}}$$

In the equation above, θ is the attitude estimate, δ_i is the location parameter for item i, and τ_j terms are the response threshold parameters for the set of items.

Basically, the model which is a variant of the Rasch cumulative model specifies that the probability of a person responding positively to an item increases with respect to increasing location of the person on the latent trait.

An issue with the use of the Rating Scale model in calibrating Likert scales has earlier been pointed out by Wright and Masters (1982). They had demonstrated the use of rating scale analysis in calibrating a scale comprising 10 statements for drugs and 10 statements against drugs. They separated the scale into two halves based on the positive and negative statements and were able to show in each half of the scale, the item ordering separating the strong and weak *against* statements and the strong and weak *for* statements.

The researchers discussed the possibility of reversing the scoring of the *against* statements and the pooling of both *for* and *against* statements to analyze them simultaneously. However, the researchers had shown that the *for* and *against* statements produced substantially differently attitude estimates for some respondents. They cautioned the pooling of both negative and positive statements and suggested analyzing them separately for the purpose of diagnostic comparisons.

Recent development of the PARELLA Model (Hoijtink, 1991) has shown the possibility of Likert scale calibration by operationalizing the latent trait of interest as a parallelogram model. The PARELLA Model specifies that the probability of a positive response decreases with increasing distance between the location of the person and the item:

$$(X=1|\beta,d) = \frac{1}{1+|\beta+d|^{2\gamma}}$$

The degree in which the person-item-distance determines a person's response to an attitude item depends on the γ parameter which is the strength of the relation between the person-item-distance of the person and his/her response. A person located at an arbitrary distance of 1.0 from an item always has a probability of 0.50 of giving a positive response to the item. The ICC of a PARELLA item has a single peak with a choice probability equal to 1 for $\beta = d_i$ for a person responding to the item.

While a normal procedure in analyzing such scales in the cumulative IRT model is to reverse the scoring of the response alternatives of the negatively worded items, this is not necessary in the PARELLA model. In the PARELLA model, negatively worded items are arranged in order of decreasing negative affect on the latent trait and positively worded items are arranged in order of increasing positive affect to establish a sequence of proximity items.

The purpose of this study is to compare the effectiveness of the PARELLA model and the Rating Scale Model in evaluating Likert scales where negatively worded items are present. Such

comparisons will take into account, the model fit of the dataset and the ordering of the item locations.

METHOD

Data came from responses to an attitude scale as part of a study on job satisfaction of first year teacher interns. The 10-item Teacher Intern Career Commitment Scale (TICCS) was administered to 350 teacher interns after their first teaching internship. The term, 'Career Commitment' is defined as the employee's willingness to exert effort on behalf of the organization, and the desire to remain as an employee of the organization (Mowday, Porter & Steers, 1982). This construct includes intrinsic motivation and a sense of personal investment in teaching. Items were constructed based on this definition. The items addressed perceptions of leaving/staying on in the teaching profession, as well as motivation. Table 1 shows the items of the TICCS.

Insert Table 1 about here

The first five items were negatively worded. That is, these items were phrased to reflect the 'contra side' of the unidimensional representation of job satisfaction. The second five were positively worded, reflecting the 'pro side' of the same construct. The items were arranged in order from measuring the most negative affect to the most positive affect toward career commitment, based on the judgement of the test constructors and the opinion of fellow researchers. The items were rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

The scale was originally developed for analysis using the Rating Scale Model, that is, responses to all negatively worded items were reverse scored before analysis. For analysis using the PARELLA model, responses to the negatively worded items were not reversed scored but the response matrix was recoded dichotomously (1, 2, 3 = 0; 4, 5 = 1).

To establish unidimensionality of the scale, factor analysis was done using the polychoric correlation matrix of the response matrix meant for the Rating Scale Model. The same analysis was done using the tetrachoric correlation matrix of the response matrix meant for the PARELLA Model. The computer program, MicroFACT (Waller, 1994) was used for both factor analysis procedures.

The original response matrix was fitted into the Rating Scale Model using the computer program, QUEST (Adams & Khoo, 1994). The dichotomotized response matrix was fitted into the PARELLA Model using the computer program, PARELLA (Hoitjink, Molenaar & Post, 1994).

RESULTS

Results of the factor analyses showed both response matrices to be essentially unidimensional, thus meeting a condition for IRT analyses. Table 2 shows the percentage item responses to the matrix to be fitted into the Rating Scale Model and the matrix to be fitted into the PARELLA Model.

Insert Table 2 about here

Analysis of item fit was based on the weighted and unweighted residual based statistics (Wright & Masters, 1982).

These are mean squares that are normalised to the infit and outfit t statistics respectively. In general, the weighted statistic is more reliable as an overall measure of measurement quality since it is also less susceptible to aberrant responses (Wright & Masters, 1982). When the data conform to the model, the t values have a mean near to zero and a SD near to 1. In the first run, Item 5 showed a bad fit in which the infit and outfit t were 15.3 and 16.9 respectively, beyond the third standard deviation. This item was dropped in the second run. In the second run, Item 4 was identified as misfitting with both infit and outfit t values of 6.3, beyond the 2nd standard deviation. This item was dropped from the dataset in the third run. Results showed satisfactory fit statistics for both case estimates (mean infit t = -0.13, mean outfit t = -0.05) and item estimates (mean infit t = -.09, mean outfit t = 0.15). The case weighted mean square was 1.03 (SD = 0.86) and the item weighted mean square was 1.00 (SD = 0.17). Table 3 shows the results of the final QUEST run. The results show the weighted t statistic, the item location parameter values and the item threshold values.

Insert Table 3 about here

The sample reliability of item separation R was 0.81. Based on Wright and Masters' (1982) definition where $R = G^2/1-G^2$ and G is the Item Separation Index, the 8 items defined 3 statistically distinct attitude strata.

Results of the Rating Scale analysis showed that:

a) The location parameters do not indicate the order of the items on the trait as originally intended. The removal of the two negatively worded items left only 3 negatively

worded items (Items 1 - 3) with little item separation between them. Items 1 - 3 appeared to show more positive scale values when the negatively worded items actually defined a more negative effect towards career commitment. Item 7 and item 8 which defined a positive affect towards career commitment had lower location values compared to those of Items 1 - 3.

b) The positively worded items were arranged in order of increasing trait level from Items 7 - 10 based on the location parameter values.

c) Respondents with greater negative attitude estimates were likely to indicate 'Strongly Disagree' to positive statements represented by Items 6 - 8. These positive items appeared to be difficult to strongly disagree unless one possesses very negative attitude towards career commitment.

d) With the exception of Item 1, all items required high respondent attitude estimates in order to strongly agree to statements represented by these items.

e) Parameter values for step levels 2 and 3 in Item 1 were equal. Parameter value for step level 2 was less than that in step level 1 for Item 2.

The summary of the final runs of the PARELLA analysis of the TICCS after deletion of the misfitting item is shown in Table 3.

The sum of difference (SUM) diagnostic statistic was used to compare the fit of the empirical ICC with the corresponding theoretical (PARELLA) ICC. This is the sum of the differences between the empirical and the PARELLA taken over each node of the person location non-parametric density estimate. In any

PARELLA run, the computed standardized or SUM/SQR(N) values were compared with the distribution table (Hoijtink, Molenaar & Post, 1994). In the first PARELLA run, Item 5 showed a bad fit (SUM = 20.9, SUM/SQR(N) = 1.12). This item was removed for the second run. After the second run, the small SUM values and corresponding standardized values of the items (see Table 3) indicate good agreement between the empirical and the PARELLA model.

The order of the statements corresponded with the initial order. The estimated power (γ) was 1.34 which indicates the presence of a fairly strong parallelogram structure. That is, the data matrix shows not too many *0*s between the *1*s.

Based on the step function estimate (see Table 4), many respondents were located between -0.10 and -1.30. This is an indication that they generally expressed feelings of lower career commitment levels.

DISCUSSION

In the calibration of the TICCS, both models identified Item 10 as the item measuring the most positive affect towards career commitment based on the location parameters relative to other items. Although the scale was originally designed for calibration using the rating scale model, the ordering of the items did not appear as expected. The almost correct ordering of the positively worded items, the difficulty of fitting the two negatively worded items and discrepancies in step estimation of the negatively worded items in the Rating Scale Model appeared to lend support to Wright and Masters' (1982) recommendation that positive and negative sets of items should be calibrated

separately. Location of the items on the latent trait based on the estimated parameters by PARELLA showed congruence with that of the theoretical order of the items in the scale. It must be pointed out that the order of items in the middle of the scale is debatable and differences established between the two measurement models are relative. Although the dataset fit the PARELLA Model satisfactorily, it must be pointed out that conversion of the 5-point scale to a binary scale may result in loss of information. The step estimates given by the Rating Scale Model give additional information to the way in which respondents indicate their level of agreement to the statements as a function of their attitude estimate.

CONCLUSION

The study showed that given the need to establish proximity items with item scale order where negatively worded items are used, the PARELLA model showed promise although recoding may lead to loss of information. The rating scale model indicates logit thresholds for responses to scale categories, giving more information especially to positively worded attitude items. The study recommends that for Likert scales, separate analysis of positive and negative items using the Rating Scale Model should be used provided there are sufficient number of items. Although binary conversion of the Likert scales may result in a loss of information, the PARELLA Model could be used to provide information on item location and person separation on the trait. The results showed that the parallelogram model serves as a good complement to the cumulative IRT Rating Scale Model for Likert type items with positive and negative item sets.

Table 1. Teacher Intern Career Commitment Scale

1.	I had considered leaving the teaching profession	[-]
2.	I would take up a different profession rather than teaching if paid the same.	[-]
3.	I felt teaching frustrated me.	[-]
4.	I did not like the long hours of teaching preparation.	[-]
5.	It was difficult balancing my teaching responsibilities and social life.	[-]
6.	I felt teaching allowed me to utilize my fullest abilities.	[+]
7.	I am motivated to work in the teaching profession.	[+]
8.	The teaching profession fulfils my job values.	[+]
9.	Teaching is the ideal profession for me.	[+]
10.	If I am rich enough without working, I would still work in the teaching profession.	[+]

Table 2. Percentage Item Responses

Item	Actual Item Responses					Binary Scored (PARELLA Model)	First Five Items Reversed Scored (Rating Scale Model)				
	SD	D	U	A	SA	% Endorsed	SD	D	U	A	SA
1	31.7	34.3	16.0	13.7	4.3	18.0	4.3	13.7	16.0	34.3	31.7
2	20.0	31.2	35.4	8.3	5.1	14.6	5.1	8.3	35.4	31.2	20.0
3	23.4	43.7	18.3	10.9	3.7	13.4	3.7	10.9	18.3	43.7	23.4
4	5.1	27.1	20.6	31.5	15.7	47.2	15.7	31.5	20.6	27.1	5.1
5	15.7	23.4	15.4	34.3	11.1	45.5	11.1	34.3	15.4	23.4	15.7
6	1.4	7.4	28.9	46.9	15.4	52.0	1.4	7.4	28.9	46.9	15.4
7	0.6	4.9	19.4	52.3	22.8	75.1	0.6	4.9	19.4	52.3	22.8
8	1.1	6.6	26.0	48.9	17.4	66.3	1.1	6.6	26.0	48.9	17.4
9	2.9	7.1	32.3	40.0	17.7	57.7	2.9	7.1	32.3	40.0	17.7
10	8.0	11.7	35.4	32.6	12.3	44.9	8.0	11.7	35.4	32.6	12.3

N = 350

Table 3. Fit Statistics and Location Parameters of Final QUEST and PARELLA Runs

Item	Rating Scale Analysis Infit t	δ	τ_1	τ_2	τ_3	τ_4	PARELLA Analysis SUM	SUM/SQR(N)	χ^2 (df=8)	Loc
1	1.6	-0.05 (0.07)	-1.44 (0.30)	-0.03 (0.17)	-0.03 (0.14)	1.51 (0.14)	8.8	0.47	10.34	-2.18 (0.07)
2	-0.9	0.25 (0.07)	-1.11 (0.28)	-1.38 (0.19)	0.62 (0.13)	1.87 (0.18)	6.0	0.32	8.38	-1.64 (0.04)
3	1.9	0.00 (0.07)	-1.47 (0.32)	-0.44 (0.18)	-0.17 (0.14)	2.08 (0.16)	6.6	0.35	7.62	-1.36 (0.04)
4	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	6.2	0.34	8.65	0.63 (0.04)
6	3.0	-0.09 (0.08)	-2.06 (0.49)	-1.13 (0.22)	0.32 (0.13)	2.87 (0.19)	8.4	0.45	12.33	0.76 (0.05)
7	-1.8	-0.61 (0.09)	-2.19 (0.75)	-0.83 (0.27)	0.20 (0.14)	2.81 (0.16)	8.1	0.43	3.64	0.76 (0.05)
8	-2.1	-0.24 (0.08)	-2.06 (0.54)	-1.05 (0.23)	0.28 (0.13)	2.84 (0.18)	10.2	0.55	12.53	0.96 (0.05)
9	-2.6	0.07 (0.08)	-1.47 (0.36)	-1.36 (0.21)	0.43 (0.13)	2.39 (0.18)	10.9	0.59	7.04	1.70 (0.09)
10	0.2	0.66 (0.07)	-1.29 (0.23)	-1.35 (0.16)	0.33 (0.13)	2.32 (0.23)	5.7	0.31	4.12	2.09 (0.10)

χ^2 = 34.71 (df=72), γ^2 = 1.34 (0.05)

(Figures in parenthesis are standard errors)

Table 4. Step Functions Estimates for Statements in the TICCS

Node	Weight
-1.78	0.02
-1.24	0.17
-0.70	0.36
-0.16	0.28
0.39	0.08
0.93	0.03
1.47	0.02
2.01	0.00
2.56	0.02
3.10	0.02

BIBLIOGRAPHY

Adams, R. & Khoo S.T. (1994). QUEST Manual. Hawthorn: Australian Council of Educational Research.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement, 38, 665-680.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. Measurement and Evaluation in Counselling and Development, 25, 172-188.

Gronlund, N.E. & Linn, R.L. (1990). Measurement and Evaluation in Teaching (pp. 412). New York: MacMillan

Hoijtink, H. (1991). The measurement of latent traits by proximity items. Applied Psychological Measurement, 15, 153-169.

Hoijtink, H., Molenaar, I. & Post, W. (1994). PARELLA Manual. Groningen: iec Progamma.

Lam, T.C.M & Stevens, J.J. (1994). Effects of content polarization, item wording and rating scale width on rating response. Applied Measurement in Education, 7, 141-158.

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

Mehrens, W.A. & Lehmann, I.J. (1984). Measurement and Evaluation in Education and Psychology (pp. 238-241). New York: Holt, Rinehart & Winston.

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

Waller, N.G. (1995). MicroFACT 1.0 Manual. St Paul, Minn: Assessment Systems Corporation.

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