Likert Scaling Using the Graded Response Latent Trait Model

William R. Koch
The University of Texas at Austin

The two-parameter graded response latent trait model was applied to data collected from a conventionally constructed Likert-type attitude scale. Comparisons were made of both the person latent trait estimates and the item parameter estimates with their counterparts from the conventional scaling method. Also studied were the goodness of fit of the graded response model and the information function feature of the model indicating the precision of measurement at each level of the attitude trait continuum. The results demonstrated that the graded response model could be successfully used to perform attitude measurement for Likert scales.

Recently, some interest has been shown in broadening the domain of applications of latent trait theory to include the realms of attitude and personality measurement (Andrich, 1978a, 1978c; Bejar, 1977). Assuming the existence within persons of some unidimensional personality trait or attitude continuum that may be measured by means of items on an instrument, effort has focused on the development and application of latent trait models specifically for the types of item responses that may result. Instead of the dichotomous scoring typically associated with multiple-choice items on aptitude and achievement tests, responses to attitude and personality scale items are often polychotomous.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 1, Winter 1983, pp. 15–32

0146-6216/83/010015-18$1.90

Samejima's early work to develop the graded response latent trait model (Samejima, 1969) extended the dichotomously scored two-parameter normal ogive model (Lord & Novick, 1968) and the two-parameter logistic model (Birnbaum, 1968) to the case of ordered category, polychotomously scored items. Subsequently, Bock (1972) developed the nominal response model for unordered item response categories and Samejima (1973) proposed the continuous response model.

More recently, the general Rasch model for polychotomously scored item responses (Rasch, 1961) has been further developed by Andersen (1977), Andrich (1978b), and Masters (1981). Of particular interest and contrast in approach to the present research is Andrich's work investigating a Rasch rating scale model for ordered item response categories which are scored with successive integers in the usual Likert scale fashion. In this situation, respondents typically mark each item on a 5- or 7-point scale to indicate the degree to which they endorse a statement.

In the Andrich (1978b) model each item receives an estimated scale value or location on the attitude continuum, while the response thresholds or steps for the rating points of each item are estimated only once and used across the entire item set. Andrich (1978a) has presented an application of this model to demonstrate its usefulness in the context of Likert-style attitude scaling. In contrast, no applications of Samejima's graded response model to the

case of attitude measurement have appeared in the literature. The present study makes that application.

It is important to note that the Andrich model has a quite distinct motivation from the Samejima graded response model (Masters, 1981). In the Andrich model (1) although the thresholds between the ordered categories are estimated, the same thresholds are estimated for all items; (2) the discriminations at all thresholds for all items are assumed equal; (3) the summated raw score is a sufficient statistic for a person's attitude estimate; and (4) the threshold and person parameters are separable, providing the condition of "specific objectivity." However, in the Samejima model (1) different threshold values are estimated for each item, (2) a different discrimination value is estimated for each item, (3) the summated raw score is not a sufficient statistic, and (4) the parameters of the model are not separable for estimation.

The primary practical advantage of traditional Likert attitude scaling is its simplicity. While latent trait approaches do not share this advantage, there are considerable benefits to be gained in applied work that make the complexity of latent trait measurement worth the extra effort. For example, given a pool of precalibrated attitude scale items that measure the same trait, several possibilities present themselves. An attitude scale can be constructed that is optimal for a given situation because the accuracy of measurement at each point on the attitude trait continuum may be determined. Thus, a broad range scale could be devised that measured equally well across trait levels, or a peaked scale could be designed for maximum accuracy in a restricted range, say for individuals with very strong negative attitudes. The item pool would also facilitate the efficient generation of equivalent forms of an attitude scale or the adaptive measurement of persons' attitudes by tailoring the administration to only those items appropriate for each individual. Even though persons would be responding to different items, it would still be possible to estimate their attitude levels on a common scale. Finally, the item bias of attitude statements for subgroups of a population may also be effectively studied and detected using latent trait methodology.

The purposes of the present research were to investigate the applicability of the graded response model to attitude data and to determine if the model could successfully eliminate some commonly encountered problems with the conventional Likert attitude scaling methodology. For instance, traditional Likert scaling is limited to sample dependent item statistics, gross measures of scale characteristics (e.g., the reliability and standard error of measurement), and norm-referenced interpretations. Moreover, equal weights are usually assigned to each item during scoring. The graded response model, on the other hand, offers the capability of using item discriminations to weight the scoring, the advantage of specifying measurement errors at each attitude level, and the possibility of invariant parameter estimates for persons and items, within a transformation (Lord, 1980).

The specific objectives of the research included the determination of (1) the degree of correlation between latent trait ability estimates and traditional summated rating scores, (2) the fit of the graded response model to the attitude scale data, (3) the results of latent trait item analysis compared to classical item analysis, and (4) the attitude scale's precision of measurement at different points along the attitude trait continuum.

The Two-Parameter Logistic Model

In the context of mental ability testing, the two-parameter logistic model presented by Birnbaum (1968) requires the estimation of two parameters for each item and one parameter for each person to represent the interaction between test items and examinees. The exponential form of the model is given by

$$P_{ij} = P(u_{ij}=1|\theta) = \frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \qquad [1]$$

where

$P(u_{ij} = 1 \mid \theta)$ is the probability of a correct response by person j to item i;
D is a scaling constant equal to 1.7;
a_i is the item discrimination parameter;
b_i is the item difficulty parameter; and
θ_j is the ability parameter for person j.

The probability of an incorrect response, Q_{ij}, is defined simply as $1 - P_{ij}$.

The two-parameter model assumes that the items are scored dichotomously, that the latent trait being measured by the items is unidimensional, and that local independence holds (Lord & Novick, 1968). The last assumption means that the probability of a certain response to any given item on a test is unaffected by responses to any of the other items for a fixed level of ability.

The Graded Response Model

Samejima (1969) introduced an important extension of the two-parameter latent trait test model that allowed test items to be scored in an ordered, polychotomous fashion rather than dichotomously. The graded response model was appropriate for instances in which responses to items could be evaluated according to the degree of attainment of the solution to a problem or to the magnitude of agreement with statements in attitude measurement (Samejima, 1969). A major feature of the graded response model was that more information about a person's ability or attitude level could be obtained for graded responses than for binary responses.

In the homogeneous case of the graded response model, the categories of possible responses to an item are arranged in order, where the $(m_g + 1)$ categories are scored as $x_g = 0, 1, \ldots, m_g$, respectively. Thus, the response to each item, g, is denoted as some x_g value, with the response pattern for an individual who answers n items designated as a vector of integers, $V = (x_1, x_2, \ldots, x_n)$.

The probability of an individual responding to an item in any particular category *or higher*, $(P^*_{x_g})$, is given by the equation

$$P^*_{x_g} = [1+e^{-Da_g(\theta_j - b_{x_g})}]^{-1} \qquad [2]$$

where

b_{x_g} is the boundary (difficulty level) for category m_g,
θ_j is the ability level, and
a_g is the discrimination parameter for the item.

Because there are m_g cutting points or boundaries resulting from the $(m_g + 1)$ categories, there are m_g equations for each item in the form of Equation 2 above.

The probability of an individual responding to an item in a particular category, $P_{x_g}(\theta)$, is defined as the operating characteristic (Samejima, 1969). In general, for graded response x_g the operating characteristic is given by

$$P_{x_g}(\theta) = P^*_{x_g}(\theta) - P^*_{(x_g+1)}(\theta) > 0 \quad . \qquad [3]$$

Therefore, as is shown in Figure 1, a graphic representation of the operating characteristics for a graded response item may be obtained by plotting the differences between successive $P^*_{x_g}$ functions for each category. The exceptions are the cases when $x_g = 0$ or $x_g = m_g$, for which $P^*_0(\theta) = 1$ and $P^*_{m_g+1}(\theta) = 0$, respectively (see Figure 1).

A very significant contribution by Samejima (1969, 1976) was her comparison of the information provided by items scored in a graded response manner to the information of items scored dichotomously. In the context of latent trait theory, Birnbaum (1968) had previously defined information as the precision of measurement of an item or set of items for each point on the ability scale or trait continuum. Another interpretation was that information provided an indication of the accuracy with which an item or set of items could estimate an examinee's ability or attitude level. Samejima's results demonstrated that substantial gains in information could be achieved by means of graded scoring.

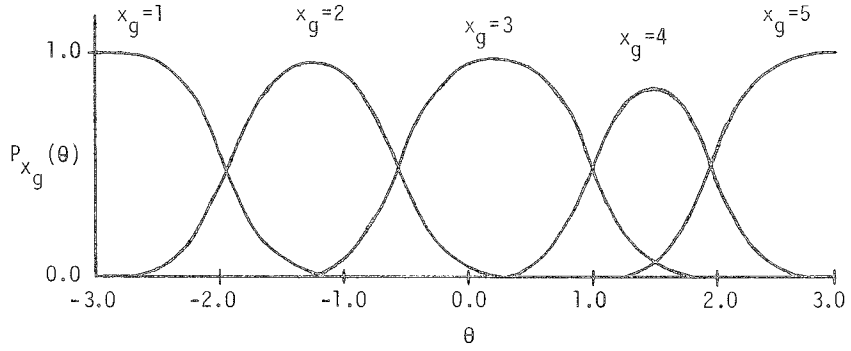
Samejima (1969) showed that the information function for a given response to a graded item has the form

$$I_{x_g}(\theta) = (P'_{x_g})^2 / P_{x_g} \quad . \qquad [4]$$

Due to the additive property of information, it followed that the information function of a single graded response item could be expressed as

$$I_g(\theta) = \sum_{x_g=0}^{m_g} I_{x_g}(\theta) \quad . \qquad [5]$$

Figure 1
Operating Characteristics for a Graded Response Item



Finally, the information of a test or scale consisting of g items was given by

$$I(\theta) = \sum_{g=1}^{n} I_g(\theta) = \sum_{g=1}^{n} \sum_{x_g=0}^{m_g} I_{x_g}(\theta) \quad . \quad [6]$$

Method

Instrumentation and Sample

The data used for the present research consisted of responses from a total sample of 491 teachers to a 40-item Likert scale. The instrument, the Audit of Administrator Communication (ADCOM; Valentine, 1978), was designed to measure the communication skills of school administrators with the teachers working under this supervision. Each of the statements described an aspect of communication between the teachers and their administrator, with the responses intended to reflect the teachers' attitudes toward the communication skills of their administrators.

As with many scales of this type, one primary dimension was being measured (communication in general) along with numerous subscales. All of the items were scored on a 5-point scale, with a score of 1 indicating an unfavorable response toward the communication skills of the administrator and a score of 5 indicating a favorable response. The respondents were requested to choose one of the following verbal descriptors for each item: *Never, Rarely, Occasionally, Usually,* or *Always*.

The responses by teachers from 10 different school systems were combined for analysis, with one administrator per school being evaluated by the teachers in that system. The schools, ranging from elementary through senior high school, with both rural and urban systems included in the data, were located in the states of Michigan, Missouri, and Pennsylvania. There was no overlap of teachers or administrators across school systems.

It was recognized that certain dimensionality problems could have resulted from lumping together the data for analysis, especially in terms of the individual differences in school systems, administrators, and teachers. However, for the exploratory and methodological purposes of the present research, where the substance of the scale itself was of little concern, the data were considered to be reasonable for use.

Item Analysis Procedures

Three different methods of item analysis (or item calibration) were compared: (1) traditional item analysis, (2) maximum likelihood item parameter estimation, and (3) factor analysis.

The traditional item analysis method consisted of the determination of item difficulty and discrimination values based on raw score responses to the items. The item difficulty values were the means of the responses to each of the items. For example, an item with a difficulty value of 4.0 was easier to agree with than an item with difficulty equal to

2.0. The traditional item discrimination values were computed as the correlations of item scores with total scores on the scale. Items with high discrimination values, therefore, measured approximately what the overall scale measured. A final procedure that is characteristic of traditional Likert scale item analysis is the removal of items that have relatively low discrimination values from the scale, which reduces the effect of dimensions other than the dominant dimension. This revision technique was used to both shorten and purify the ADCOM scale using an arbitrary criterion for item removal of correlations less than or equal to $r = .60$. The internal consistency of the resulting ADCOM scale was then determined through the calculation of the coefficient alpha reliability.

The maximum likelihood item calibration was performed by means of program LOGOG (Kolakowski & Bock, 1973). One option of this program was specifically intended to provide item parameter estimates using the graded response latent trait model. Thus, the program yielded latent trait discrimination and difficulty values for each item, with the difficulty expressed in terms of four category boundaries or cutting scores (because there were five response categories). The discrimination was a constant within an item, but it was free to vary across items.

In order to estimate the item parameters, the program divided the distribution of respondents into 10 fractiles or groupings of persons whose members were considered homogeneous enough to be represented by one trait level per fractile. Next, the proportions within each fractile responding to each of the score categories were determined for use in the likelihood function of the item parameters. The program used an iterative procedure that cycled back and forth from item parameter estimates to ability (attitude) estimates until stable values were obtained. (Details are provided in Kolakowski & Bock, 1973.)

Another feature of the program was the provision of chi-square tests of the fit of the graded response model to the individual items. Therefore, it was possible to use chi-square values as one criterion to judge the appropriateness of the graded response model when applied to the ADCOM data. In addition, based on chi-square values, the LOGOG program was used successively to remove poorly fitting items from the ADCOM scale. The results of using the LOGOG program to revise the scale were compared to the traditional method of scale revision described above.

The factor analysis method for obtaining item parameter estimates was suggested by Samejima (1969) and later specified in detail (Samejima, 1976). The method involved initially performing a principal axis factor analysis with iteratively estimated communalities on the item intercorrelation matrix. The factor loadings of the items on the first principal factor were then used in the estimation of the item parameters. These were calculated as

$$a_g = \rho_g / (1-\rho_g^2)^{\frac{1}{2}} \qquad [7]$$

where a_g was the discrimination estimate for item g and ρ_g was the factor loading of item g. The formula used to compute the difficulty parameters was given by

$$b_{x_g} = \gamma_{x_g} / \rho_g \quad \text{for } x_g = 1,\ldots,m_g \qquad [8]$$

where

b_{x_g} is a category boundary for item g,

γ_{x_g} is the normal deviate value corresponding to the proportion of the examinees who obtained the item score x_g or greater, and

ρ_g is the factor loading of item g.

For data with one clearly dominant factor, the loadings of the items on the first factor were equivalent to the item discrimination values obtained from the traditional item analysis. Therefore, the removal of items to revise and purify the ADCOM scale was identical for both the Samejima method and the traditional methods.

Ability Estimation Procedures

Three separate methods of trait estimation were compared, corresponding to the three item analysis methods. The trait estimation procedures consisted of (1) the traditional summated raw score, (2) the

iterative maximum likelihood trait estimate from the LOGOG program, and (3) an empirical maximum likelihood method.

The traditional raw score trait level estimates were obtained simply by adding the scores on each of the items for each person. No attempt was made to weight the item responses to reflect the variations in item difficulty.

The maximum likelihood latent trait estimates were output from LOGOG. The estimation procedure made the assumption of a normal distribution of the latent trait in the subject population. Given the response strings to the items and initial item parameter estimates, Newton-Raphson iterations determined the mode of the trait likelihood functions for each person. In turn, these trait estimates were conditioned upon for the determination of new item parameter estimates, which were used to obtain new trait estimates, and so forth. The number of program cycles was an option to be set by the user.

The empirical maximum likelihood trait estimates were obtained through the use of a computer program that computed the likelihood function of a person's specific response string, given the item parameter estimates. Once the mode of the likelihood function was bracketed, the program successively converged upon the mode, which became the latent trait attitude estimate. The program was a modified version of a procedure developed by Reckase (1974) to perform latent trait ability estimation for dichotomously scored items. The input for the graded response version consisted of the response strings and the item parameter estimates obtained from the Samejima method described above.

Analyses

An iterative principal axis factor analysis from the SAS package (Barr, Goodnight, Sall, & Helwig, 1976) was run on the ADCOM data for two reasons. First, the loadings of the items on the first factor were needed to obtain the Samejima item parameter estimates. Second, the factor structure of the data had to be determined because the latent trait model assumed that the trait being measured was unidimensional.

Correlation analyses were performed on the item parameter estimates and attitude trait estimates that were produced by the traditional, LOGOG, and Samejima methods. The goal was to determine the degree of linear relationship present among the sets of item parameters for each method, as well as among the trait estimates.

Information

The final set of data analyses were performed to investigate the feature of latent trait models in which the information that was contributed by each item toward the measurement of the respondent's attitudes may be determined. When the item information functions were summed for the whole ADCOM scale, its precision of measurement was specified for each level of the underlying attitude continuum. These information analyses were conducted for both the LOGOG and the Samejima latent trait item parameter estimates.

Finally, several individual item information plots were simulated in order to examine the effects of the item parameters on the information provided by the items. For example, the effects of high and low item discrimination values were illustrated, as well as the effects of symmetric or skewed response distributions and small or large ranges of the item difficulty boundaries.

Results

Factor Structure

The unrotated factor loading matrix for the ADCOM data is shown in Table 1. In the preliminary stage of the factoring, there were four factors present in the data with eigenvalues greater than 1.0, so four factors were retained. The eigenvalue for the first factor was equal to 14.66 and accounted for 50.5% of the total variance. However, after three iterations, only the eigenvalue for the first factor was greater than 1.0, having a value of 14.25 and accounting for 85% of the common variance. Clearly, there was a single dominant factor being

Table 1
Iterative Principal Axis Factor Analysis
Unrotated Factor Pattern for ADCOM Data[a]

Item Number	Factors			
	I	II	III	IV
1	.70	-.05	.36	.08
2	.61	.30	.12	-.07
3	.65	.15	-.13	-.08
4	.72	-.22	.05	.15
5	.79	-.02	.10	.03
7	.72	.07	.18	.03
8	.83	-.20	-.03	-.02
13	.64	.33	.08	.11
14	.61	-.25	.24	-.19
15	.65	-.08	.32	.11
16	.70	-.22	-.00	.02
17	.72	-.25	-.16	.13
18	.71	.05	.12	.19
19	.71	.08	-.19	.21
21	.65	.07	-.15	.25
22	.67	.31	.01	-.17
23	.66	.18	-.01	.04
26	.71	-.16	.13	-.33
27	.62	-.21	-.31	.01
28	.71	.03	.17	.16
32	.72	.21	-.15	-.16
33	.79	-.01	.11	-.08
34	.81	-.16	-.20	.05
35	.73	-.11	-.07	-.36
36	.82	-.22	-.23	-.00
37	.60	.03	.10	-.02
38	.64	.13	-.06	.18
39	.69	.19	-.13	-.25
40	.68	.23	-.19	-.01
Eigenvalues	14.25	.95	.82	.69

[a]Eleven items were deleted from the ADCOM scale during item analysis prior to factor analysis.

measured by the ADCOM scale, with all of the items having high loadings on this first principal factor.

Item Parameter Estimation

The results of the traditional item analysis are reported in Table 2. The difficulties and discrimination values of all 40 items from the ADCOM scale are shown in the table. One objective of the item analysis was to reduce the length of the scale somewhat; another was to make it more unidimensional. The 11 items that were removed from the scale due to their relatively low item discrimination values are identified in Table 2. The criterion for item removal, item discriminations less than or equal to .60, was arbitrary and was used only for illustrative purposes.

The mean response of the ADCOM scale items was 3.69, which reflected the fact that the items tended to have fairly low difficulty values overall. Upon removal of the 11 items, the mean item difficulty changed slightly (to 3.65), while the mean item discrimination changed from .66 to .71. The coefficient alpha reliability of the final 29-item scale was $\alpha = .96$.

The results of the latent trait item parameter estimation for the ADCOM scale are shown in Table 3. Both the item parameter estimates resulting from the LOGOG program item calibration and those from Samejima's method of item parameter estimation are presented. Inspection of the respective

Table 2
Traditional Item Statistics

Item Number	Diff.	Disc.	Item Number	Diff.	Disc.
1	3.10	.70	21	4.08	.68
2	3.80	.61	22	3.32	.67
3	3.86	.66	23	3.57	.68
4	4.08	.73	24	2.31	.56[a]
5	3.78	.79	25	2.53	.55[a]
6	4.46	.42[a]	26	2.99	.71
7	3.17	.72	27	4.04	.63
8	3.87	.82	28	3.19	.71
9	4.02	.55[a]	29	4.20	.47[a]
10	3.95	.49[a]	30	4.04	.54[a]
11	4.02	.53[a]	31	4.08	.55[a]
12	4.12	.57[a]	32	3.63	.71
13	3.68	.65	33	3.33	.79
14	3.11	.62	34	3.95	.80
15	3.18	.67	35	3.85	.73
16	3.67	.72	36	4.08	.81
17	4.24	.72	37	2.84	.61
18	3.50	.71	38	3.56	.67
19	4.12	.72	39	3.92	.69
20	4.06	.60[a]	40	4.22	.69

[a]Indicates items removed from the ADCOM scale during item analysis due to their relatively lower discrimination values.

LOGOG and Samejima difficulty values for each item reveals a very close correspondence. The item discrimination parameters from the two methods are also quite comparable.

The Samejima method of item parameter estimation was applied only to the revised 29-item ADCOM scale resulting from the traditional item analysis, which explains the absence of item parameter estimates for some of the items in Table 3. However, as was previously mentioned, the LOGOG program was also used to perform revisions of the full ADCOM scale. The chi-square criterion of fit of the graded response model to the item response data was used to successively remove items from the initial 40-item scale. After each program run, the items with obvious lack of fit were removed, which explains the absence of LOGOG item parameter estimates for some of the items in Table 3. Note that different items were removed with the LOGOG item analysis than with the traditional item analysis. This difference demonstrated that the LOGOG lack of fit reflected something more than just low item discriminations. In fact, the poorly fitting items tended to have high discrimination values. Detailed examination of the actual content of the deleted statements failed to reveal any obvious patterns.

Item Parameter Correlations

The correlations among the item parameter estimates yielded by the three different methods are reported in Table 4. It is quite evident that the three procedures produced highly related estimates of item difficulty. For the ADCOM scale, only 20 items in common remained after the traditional and LOGOG item analysis revisions (see Table 3). Since each item had four difficulty boundary estimates, the correlation between the LOGOG and Samejima difficulties was based on 80 pairs of values. However, only one difficulty estimate was provided by the traditional item analysis method. Therefore, the traditional item difficulty values were correlated with the means of the four difficulty boundaries per item for the LOGOG and Samejima methods. Hence, these correlations were based on only 20 pairs of values.

The correlations among the item discrimination values yielded by the three estimation methods were also relatively high. The correlations in the upper off-diagonal portion of the matrix are based on the full ADCOM scale prior to revision. However, because the LOGOG program was unable to accurately estimate the discrimination parameter for Item 29, it is not included in two of the correlations. Across program iterations or cycles, the estimated values of the discrimination for this particular item were very unstable and were associated with extremely large standard errors. Examination of the statement's content showed that it was double-barreled, interpretable in various ways. The detailed item responses indicated that both persons with high and low total attitude scores responded similarly to the item. This item was deleted from subsequent runs of LOGOG, and it was also deleted during traditional item analysis due to its low discrimination value.

For the full scale it is apparent that the discriminations from the two latent trait procedures correlated highly with each other as well as with the traditional values. The correlations reported in the lower off-diagonal part of the matrix, which are based on the common items remaining in the revised ADCOM scale, are substantially lower. In particular, the correlations between the LOGOG discriminations and both the traditional and Samejima estimates are only moderate, being $r = .54$ and $r = .64$, respectively. As will be discussed later, this result was likely due to the restricted variance of the discrimination estimates for these remaining items. The high correlation ($r = .98$) between the traditional and the Samejima discriminations was expected because both essentially represent the loadings of the items on the first factor of the data.

Person Attitude Estimate Correlations

Table 5 reports the correlations among the attitude trait estimates yielded by the three different methods of person parameter estimation. It is quite evident from the high correlations that all three procedures produced highly related person attitude estimates.

Table 3
Latent Trait Item Parameter Estimates for ADCOM Data

Item Number	LOGOG Program					Samejima's Method				
	b_1	b_2	b_3	b_4	Disc.	b_1	b_2	b_3	b_4	Disc.
1	-2.04	-1.00	.43	2.39	.92	-1.95	- .95	.45	2.26	.97
2	-2.85	-1.69	- .72	.68	.78	-2.81	-1.70	- .71	.69	.76
3	-2.79	-1.96	- .89	.92	.85	-2.67	-1.93	- .91	.87	.86
4	-2.66	-].64	- .88	.09	1.13	-2.68	-1.66	- .91	.08	1.05
5						-2.09	-1.33	- .53	.62	1.27
6	-5.96	-4.14	-2.56	- .63	.55					
7						-2.13	- .81	.35	1.73	1.04
8						-2.10	-1.35	- .65	.49	1.49
9	-4.55	-2.62	-1.23	.80	.78					
10	-3.78	-2.58	-1.28	1.02	.72					
11	-4.00	-2.71	-1.69	1.25	.91					
12	-3.48	-2.52	-1.29	.32	.84					
13	-2.67	-1.62	- .56	1.15	.92	-2.61	-1.69	- .58	1.20	.83
14	-2.13	- .73	.50	1.64	.66	-1.92	- .68	.46	1.49	.76
15	-2.15	-1.06	.29	2.07	.81	-2.01	-1.01	.28	1.95	.86
16	-2.52	-1.47	- .37	.99	1.00	-2.48	-1.47	- .40	.95	.99
17	-2.92	-1.97	-1.20	.04	1.24	-3.12	-2.09	-1.30	.01	1.04
18	-2.37	-1.23	- .30	1.37	1.06	-2.35	-1.23	- .30	1.39	1.00
19	-2.77	-2.02	-1.12	.43	1.36	-3.17	-2.31	-1.26	.48	1.00
20	-3.78	-2.54	-1.50	.90	.89					
21	-3.54	-1.84	-1.18	.51	1.19	-4.06	-2.14	-1.37	.57	.86
22						-1.98	- .94	- .03	1.45	.70
23	-2.69	-1.42	- .41	1.51	.96	-2.76	-1.49	- .42	1.53	.88
24	- .94	.54	1.68	3.10	.65					
25	-1.67	.01	1.68	3.80	.59					
26	-1.59	- .47	.48	1.74	.94	-1.55	- .48	.45	1.65	1.00
27	-3.09	-1.86	- .94	.27	1.06	-3.25	-2.08	-1.07	.34	.79
28						-2.05	-1.02	.24	2.09	1.01
29										
30	-4.20	-3.00	-1.49	.96	.83					
31	----[a]	-2.93	-1.69	1.02	.89					
32	-2.34	-1.37	- .37	1.04	1.03	-2.29	-1.37	- .38	1.02	1.02
33						-1.69	- .83	.04	1.11	1.30
34						-2.39	-1.67	- .87	.61	1.36
35						-2.40	-1.54	- .67	.63	1.07
36						-2.32	-1.70	-1.06	.38	1.42
37	-2.05	- .58	1.01	3.05	.81	-2.10	- .58	1.07	2.95	.75
38	-2.69	-1.31	- .45	1.56	1.00	-2.85	-1.44	- .51	1.66	.84
39	-2.58	-1.78	-1.02	.71	.99	-2.56	-1.78	-1.03	.71	.96
40	-3.42	-2.41	-1.31	.20	1.07	-3.52	-2.52	-1.38	.22	.93

[a] Indicates the inability of the LOGOG program to estimate the lowest difficulty cutting point for this item, since no respondent chose category 1 for item 31.

Although the ADCOM data initially consisted of 491 cases, several cases were deleted during latent trait attitude estimation. One reason was that maximum likelihood estimates could not be calculated when all item responses by a person were either in the highest score category or all in the lowest score category. Another reason was that the maximum likelihood trait estimation procedure was unable to converge for cases in which there were response inconsistencies. For example, the likeli-

Table 4
Correlations Among the ADCOM Item Parameter Estimates Yielded by Three Different Methods[a]

Estimation Method	Methods of Difficulty Estimation		
	1. Traditional	2. LOGOG Program	3. Samejima
1	1.00	- .99(20)	- .99(20)
2		1.00	.99(80)
3			1.00

Estimation Method	Methods of Discrimination Estimation		
	1. Traditional	2. LOGOG Program	3. Samejima
1	1.00	.87(39)	.97(40)
2	.54(20)	1.00	.93(39)
3	.98(29)	.64(20)	1.00

[a]The numbers in parentheses indicate the number of pairs of item parameter estimates upon which the correlations were based.

hood function may be virtually flat for a person who endorses the *Never* category to some items that describe good communication behavior, but who also endorses the *Always* category to similar items.

Fit of the Graded Response Model

As has been mentioned previously, the LOGOG program provided chi-square tests of the fit of the graded response model to each item on the scale. The tests were the usual Pearsonian chi-square statistics computed from the observed and expected frequencies of responses in the item categories. Table 6 shows the results of the fit of the model to the ADCOM items. Of the 30 items remaining in the ADCOM scale after the LOGOG item analysis, four items were still found to have a significant lack of fit to the model. The probability of chi-square values as large as were obtained for these four items was less than .01. In general, based on the chi-square criterion, the graded response model fit the ADCOM data moderately well.

It should be noted that chi-square statistics have frequently been criticized and judged to be inadequate for use to determine the fit of a model to data for two main reasons: the effects of sample size on statistical significance and the problem of inaccurate estimation of frequencies of response in item categories due to insufficient sample sizes in certain ranges of attitude levels. With the LOGOG program, the present author has found substantial incidence of chi-square lack of fit even when simulation data were generated deliberately to fit the graded response model.

Information Analyses

The plots in Figure 2 illustrate the comparison of the information provided by the ADCOM scale for both the Samejima and LOGOG item parameter estimates. From the figure it can be seen that the ADCOM scale was most precise in its measurement toward the lower end of the attitude trait continuum. Both total scale information curves peaked near -1.0, and both provided their highest amounts

Table 5
Correlations Among the Person Attitude Parameter Estimates Yielded by Three Different Methods[a]

Attitude Estimation Method	Methods of Attitude Estimation		
	1. Traditional Summated Raw Scores	2. Maxlike. from Output of LOGOG Program	3. Empirical Maxlike. Based on Samejima Item Parameters
1	1.00	.94(477)[b]	.98(490)[c]
2		1.00	.95(477)
3			1.00

[a]The numbers in parentheses indicate the number of persons upon which the correlations were based.

[b]The LOGOG program failed to converge to attitude trait estimates for 13 cases.

[c]Latent trait estimates could not be computed for 1 case because this response string consisted of scores of 5 on all items.

Figure 2
ADCOM Scale Information Based on Samejima and LOGOG Item Parameter Estimates

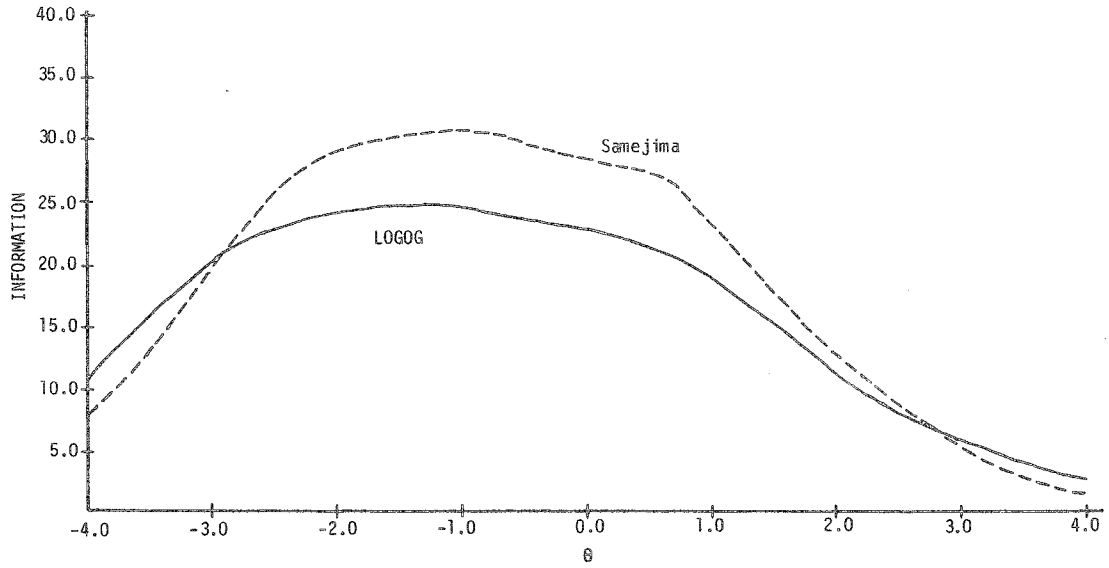


Table 6
Fit of the Graded Response Model to Items on the Revised ADCOM Scale[a]

Item Number	Chi-Square[b]	Item Number	Chi-Square[b]
1	64.65*	19	28.10
2	27.82	20	37.13
3	44.53	21	29.45
4	71.20*	23	40.78
6	47.75	24	31.47
9	55.15	25	54.82
10	35.16	26	27.78
11	31.93	27	70.26*
12	55.14	30	33.77
13	46.58	31	30.20
14	36.95	32	35.79
15	69.83*	37	39.81
16	34.72	38	55.58
17	46.91	39	29.32
18	49.11	40	37.25

[a]The chi-square values resulted after six cycles of the LOGOG program. Ten items had been deleted from the ADCOM scale due to poor fit on previous runs. Also, n = 477 because 13 cases were deleted due to nonconvergence of ability estimation.

[b]All chi-square values have 35 degrees of freedom.

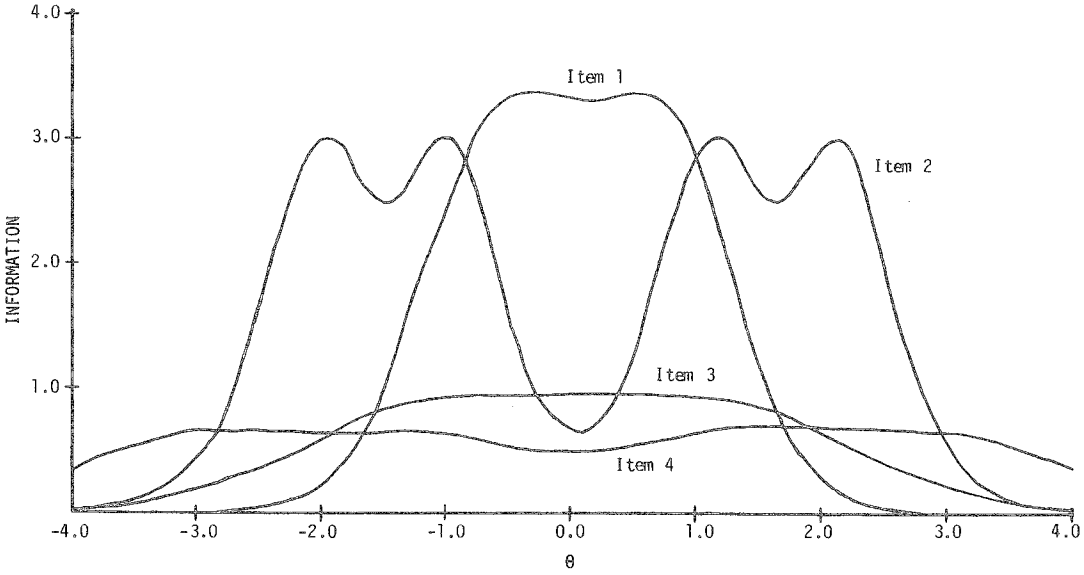
*$p < .01$

of information about attitudes for persons between −3.0 and +1.0 on the trait (θ) scale. However, the attitudes held by persons located at the upper end of the scale were not measured nearly as accurately. For both the LOGOG and the Samejima methods, the information curves are based on 29 items, although only 20 were in common. Because information is additive, one item was deleted from the LOGOG scale (an item that had low discrimination) to make the two curves comparable.

Note that the total information of the scale resulting from the Samejima item analysis was somewhat higher than the scale information of the LOGOG item analysis method. This was not surprising, given the influence of the discrimination parameter on item information. Thus, the use of the traditional item discrimination criterion for scale revision resulted in the retention of more informative items than the use of the chi-square fit criterion for these data. However, some of the high discriminations may have capitalized on extreme responses and chance. Therefore, these high values may not have held up in a new set of data.

The information functions for selected graded response attitude items are illustrated in Figures 3 and 4, with four separate item information functions represented in each figure. All four of the items in Figure 3 are symmetric around 0.0 of the θ scale, which can also be ascertained from the item difficulty boundaries reported. Items 1 and 2

Figure 3
Information Plots for Four Items with Symmetric Difficulty Boundaries and Varying Discriminations



Item parameters (Discrimination; 4 Difficulty Boundaries):

Item 1 -- 1.98; -.83, -.29, .34, .89
Item 2 -- 2.00;-2.03, -.99, 1.09, 2.12
Item 3 -- .90;-1.02, -.37, .46, 1.14
Item 4 -- .88;-2.93,-1.39, 1.30, 2.89

have item discrimination values that are more than twice as high as those for Items 3 and 4, which makes Items 1 and 2 much more informative. Notice that Item 1 is most informative for persons with θ in the range from -1.0 to $+1.0$, while Item 2 has four separate information peaks. The difficulty boundaries for Item 2 span a much wider range on θ than those for Item 1, and the four peaks correspond to the four boundary values. The information functions for Items 3 and 4 are much lower due to their low discrimination values. However, unlike Items 1 and 2, the information provided is fairly constant over a wide range of the attitude trait θ scale.

The item information plots in Figure 4 demonstrate that information need not necessarily be symmetric around the 0.0 point but may be shifted and sometimes fairly skewed, depending on the item parameters. Again, Items 1 and 2 have much higher peaks of information than Items 3 and 4 because of higher item discrimination values.

Discussion

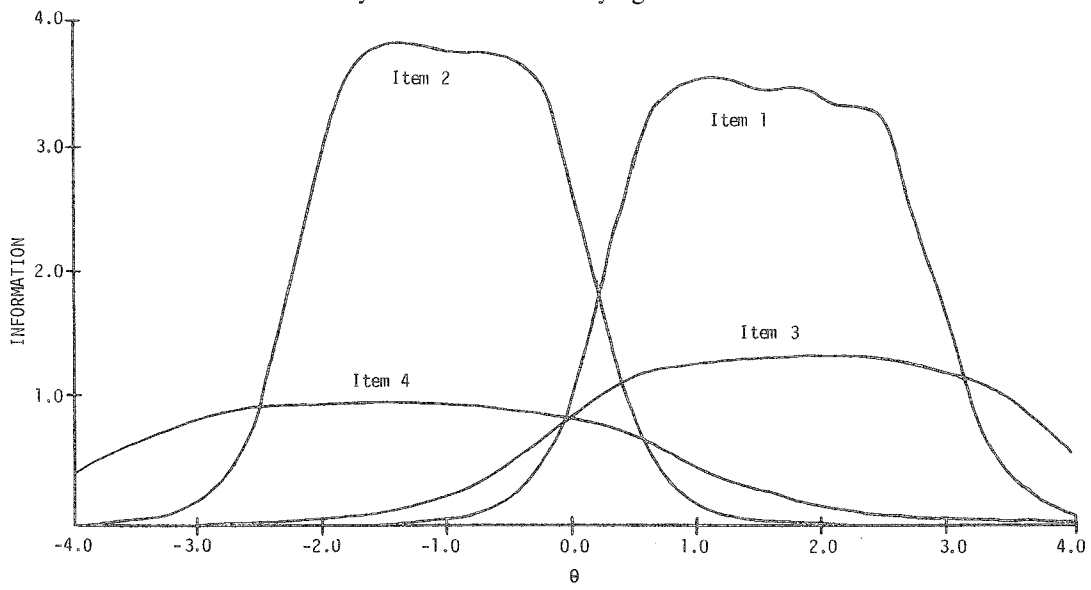
ADCOM Factor Structure

The ADCOM scale was found to have a factor structure with one very dominant factor, along with several smaller factors. Thus, the scale approximately met the assumption that the underlying trait being measured was unidimensional. Also, the scale was typical of the type of Likert attitude scale commonly in use for various attitude measurement tasks.

Item Parameter Estimation

The results of the traditional item analysis were about as expected. The item discrimination values were moderately high for the ADCOM scale. However, the item difficulties were mostly at the upper end of the 5-point response scale. This result meant that the majority of the respondents rated their administrators very favorably. The most likely explanation is that the items were fairly easy to agree

Figure 4
Information Plots for Four Items with Asymmetric Difficulty Boundaries and Varying Discriminations



Item parameters (Discrimination; 4 Difficulty Boundaries):

Item 1 -- 2.04; .64, 1.18, 1.78, 2.45
Item 2 -- 2.13;-1.83,-1.33, -.81, -.25
Item 3 -- 1.11; .64, 1.38, 2.10, 3.05
Item 4 -- .90;-2.76,-1.90,-1.03, -.25

with, although it is alternately possible that the 10 administrators evaluated by the teachers all were outstanding communicators.

The results of item difficulty parameter estimation using the LOGOG and Samejima methods were quite similar to each other. The corresponding four difficulty boundaries for each item were nearly identical for these two latent trait procedures. The explanation lies in the fact that both methods of estimation pertain to the same model and, therefore, similar estimates would be expected. Also, both procedures used the assumption that the respondents were normally distributed in terms of their attitude trait levels. The normality assumption was required for the Samejima method of item difficulty estimation and was a program option for the LOGOG program.

The correlations among the three sets of item discrimination parameter estimates were not nearly as strong as they were for the item difficulty estimates, although there was a fairly strong correspondence between the LOGOG, the Samejima, and the traditional item discrimination parameter estimates prior to the scale revisions resulting from the item analysis methods. However, the correlations dropped to only moderate levels when they were based on the 20 common items remaining after item analysis. Items were dropped from the ADCOM scale either because they had low traditional item discrimination values or because they had significantly high chi-square values in the LOGOG program, indicating lack of fit. Examination of the poorly fitting items from the LOGOG program revealed that 9 of the 10 deleted items had very high discrimination parameter estimates. Thus, the two item analysis procedures in effect removed items with extreme discrimination estimates in either direction, leaving only the moderately discriminating items to be correlated. The low variance of these remaining discrimination parameter estimates resulted in the moderate correlations.

Attitude Trait Estimate Correlations

The results of comparing the attitude trait estimates indicated that all three estimation procedures produced highly correlated scores. The traditional total raw scores, the maximum likelihood estimates from the LOGOG program, and the empirical maximum likelihood estimates based on the Samejima item parameter estimates were all very highly related. Thus, the latent trait estimation procedures apparently were unable to effectively utilize the item difficulty boundaries and discriminations to weight the item responses enough that the trait estimates were discrepant from the summated raw scores.

However, it would be a misconception to view the summated raw scores and the latent trait attitude estimates as being equivalent. It is not the case in the graded response model that the raw score is a sufficient statistic for estimating a person's attitude level. Moreover, it is perhaps not surprising that in this study the summated scores and latent trait attitude estimates were highly related. One reason is that the discrimination values did not vary much across the items, so that relatively equal weights were given to item responses in scoring. Secondly, research dating back to Likert in 1932 has usually found that no great advantages accrue from using empirically derived category weights instead of ordinary equidistant integers.

Fit of the Graded Response Model

The results of the chi-square goodness-of-fit statistics from the LOGOG program were inconclusive in determining if the graded response model was appropriate for the attitude data. For the revised ADCOM scale only a few items were found to have significant lack of fit after the cases of nonconvergent attitude trait estimation were deleted from analysis. However, during the ADCOM scale revision, 10 items were deleted due to significant chi-square values. Because chi-square tests of fit are insensitive to the direction of misfit, it would have been expected that an approximately equal number of very high and very low discriminating items would misfit the model. Disturbingly, though, it was observed that all but one of the 10 deleted items had very high traditional and LOGOG item discrimination values. It appeared, therefore, that the high discrimination parameter estimates contributed substantially to the high chi-square values. This effect could have been due to an artifact of the estimation in the LOGOG program.

However, other explanations are possible. In theory at least, it is possible in the graded response model for item discrimination parameter estimates to become infinitely high, resulting in very steep (almost vertical) category response curves. This situation would result in unusual item operating characteristics plots where the model probability of a response in any particular category would usually be quite close to either 0.0 or 1.0. Thus, any empirical deviations from the expected probabilities would dramatically inflate the chi-square fit statistic. The same result would hold for very low (near zero) discriminating items. However, the data used in the present study did not consist of any such items, so they could not be detected.

Information Analyses

The main purpose of the information function analyses was simply to demonstrate a useful feature of latent trait models. Knowledge of the precision of measurement of an attitude scale at each point on the attitude continuum can be quite valuable during both the initial construction of the scale and its later administration. For example, the results of the information analyses showed that the measurement properties of the ADCOM scale could be greatly improved by writing and including new items in the scale that were informative for high levels of the trait being measured.

The information analyses also illustrated that the item discrimination parameters were the primary determinants of the amount of information provided by an item. Items with high discriminations provided high amounts of information but only over a restricted range of the attitude trait continuum. Items with moderate discrimination parameters provided less information over a much broader range of the continuum.

Summary and Conclusion

The primary objective of the present research was to investigate the applicability of the graded response latent trait model to attitude measurement. A Likert-type scale was chosen for the study because of its popularity for usage and because the scoring of items on Likert scales corresponds to the rationale behind the graded response model.

The design of the study consisted of a series of comparisons between the results obtained from the graded response latent trait approaches and those from the traditional method of Likert scale analysis. The comparisons included both the item parameter estimation and the person parameter estimation components of the scaling approaches. Also, the latent trait feature showing the amount of information provided by the attitude scale for each level of the attitude trait continuum was studied.

The results showed that the traditional and graded response methods of attitude measurement yielded highly correlated item parameter estimates and person parameter estimates for real data from a typical Likert scale, indicating that the graded response model was appropriate for attitude measurement. Furthermore, the information function analyses demonstrated the advantage of the graded response model in its ability to determine the precision of the measurement attained by the attitude scale for each level of the attitude trait continuum. Thus, the results showed that latent trait theory could be extended successfully to the domain of attitude measurement.

References

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika,* 1977, *42,* 69–81.

Andrich, D. Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement,* 1978, *2,* 581–594. (a)

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

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

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. *A user's guide to SAS 76.* Raleigh NC: SAS Institute, 1976.

Bejar, I. I. An application of the continuous response level model to personality measurement. *Applied Psychological Measurement,* 1977, *1,* 509–521.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

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

Kolakowski, D., & Bock, R. D. *Maximum likelihood item analysis and test scoring: Logistic model for multiple item responses.* Ann Arbor MI: National Educational Resources, 1973.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Masters, G. N. *A Rasch model for partial credit scoring.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Rasch, G. On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV.* Berkeley: University of California Press, 1961.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation,* 1974, *6,* 208–212.

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

Samejima, F. Homogeneous case of the continuous response model. *Psychometrika,* 1973, *38,* 203–219.

Samejima, F. Graded response model of the latent trait theory and tailored testing. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (U.S. Civil Service Commission, Personal Research and Development Center, PS-75-6). Washington DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Valentine, J. W. *Audit of administrator communication.* Columbia MO: Jerry W. Valentine, 1978.

Acknowledgments

This article is based on parts of the author's doctoral dissertation conducted at the University of Missouri-

Columbia. The author thanks Mark D. Reckase for his supervision and guidance. Appreciation is also extended to an unknown reviewer who provided detailed constructive suggestions to improve the manuscript.

Author's Address

Send requests for reprints or further information to William R. Koch, Measurement and Evaluation Center, The University of Texas at Austin, P.O. Box 7246, University Station, Austin TX 78712, U.S.A.