1-2009

Using Paired Comparison Matrices to Estimate Parameters of the Partial Credit Rasch Measurement Model for Rater-Mediated Assessments

Mary Garner
Kennesaw State University, mgarner@kennesaw.edu

George Engelhard Jr.
Emory University

JOURNAL OF APPLIED MEASUREMENT, *10*(1), 30-41

Using Paired Comparison Matrices to Estimate Parameters of the Partial Credit Rasch Measurement Model for Rater-Mediated Assessments

Mary Garner
Kennesaw State University

George Engelhard, Jr.
Emory University

The purpose of this paper is to describe a technique for estimating the parameters of a Rasch model that accommodates ordered categories and rater severity. The technique builds on the conditional pairwise algorithm described by Choppin (1968, 1985) and represents an extension of a conditional algorithm described by Garner and Engelhard (2000, 2002) in which parameters appear as the eigenvector of a matrix derived from paired comparisons. The algorithm is used successfully to recover parameters from a simulated data set. No one has previously described such an extension of the pairwise algorithm to a Rasch model that includes both ordered categories and rater effects. The paired comparisons technique has importance for several reasons: it relies on the separability of parameters that is true only for the Rasch measurement model; it works in the presence of missing data; it makes transparent the connectivity needed for parameter estimation; and it is very simple. The technique also shares the mathematical framework of a very popular technique in the social sciences called the Analytic Hierarchy Process (Saaty, 1996).

Requests for reprints should be sent to Mary Garner, Department of Mathematics and Statistics, Kennesaw State University, 1000 Chastain Rd., #1204, Kennesaw, GA 30144-5591 USA, e-mail: mgarner@kennesaw.edu.

Choppin (1968, 1985) described a conditional algorithm for estimating parameters of the dichotomous Rasch model that was based on comparisons between pairs of items. That method consists of creation of a matrix with entries representing comparisons between pairs of items, and then using either a maximum likelihood or a least squares technique to extract parameters from the paired comparison matrix (Garner and Engelhard, 2000). Choppin limited his discussion of the pairwise algorithm to dichotomous data. Andrich (1988), Fischer and Tanzer (1994), Linacre (1989), van der Linden and Eggen (1986), and Zwinderman (1995) have also explored various aspects of the pairwise algorithm, always in the context of dichotomous data and always using a maximum likelihood approach for parameter estimation. Garner and Engelhard (2000) explored Choppin's least squares approach using dichotomous data and illustrated Choppin's suggestion of using powers of the paired comparisons matrix.

Wright and Masters (1982) included the pairwise algorithm in their review of estimation procedures and were the first to suggest a method for extending the pairwise algorithm to the rating scale and partial credit models by describing how the paired comparison matrix might be created. Garner and Engelhard (2002) described a new method for creating the paired comparison matrices for the rating scale and partial credit models. In the same article, the authors described an alternative to maximum likelihood and least squares methods for extracting the item parameters from the matrix of paired comparisons; the method involves calculating the eigenvector associated with the maximum eigenvalue of a matrix derived from the matrix of paired comparisons.

Fischer (1974) described a method of using eigenvectors to obtain parameters of the Rasch model for dichotomous data. The method described in this article differs from the method described by Fischer in the following way. Although the initial matrix of paired comparisons formed in both algorithms is the same, Fischer applies a series of transformations such that the eigenvector obtained from the resultant matrix contains parameters that minimize the difference between the observed matrix and an expected symmetric matrix. In this article, on the other hand, the original matrix of paired comparisons or a power of the matrix is manipulated to form a positive reciprocal matrix, and then the eigenvector associated with the maximum eigenvalue of the reciprocal matrix is calculated. The properties of reciprocal matrices are described extensively by Saaty (1996) who uses the matrices in a popular method in the social sciences called the Analytic Hierarchy Process.

The purpose of this paper is to describe a method for extending the pairwise algorithm and the eigenvector method of Garner and Engelhard (2002) to a form of the partial credit Rasch model (Masters, 1982) that incorporates a parameter for judge severity. This method for extension of the pairwise algorithm to the partial credit model differs from the one described by Wright and Masters (1982) in that the matrix of paired comparisons, which forms the basis of the algorithm, is built in a manner more consistent with Choppin's approach. In addition, Wright and Masters (1982) apply a maximum likelihood technique to the paired comparison matrix to obtain item parameters, whereas this paper applies an approach described by Garner and Engelhard (2002) in which partial credit parameters and rater severity parameters appear as eigenvectors of matrices derived from the paired comparison matrices. No description exists for the extension of the pairwise algoithm to a partial credit Rasch model that includes parameters for rater severity.

Choppin's Pairwise Algorithm for Dichotomous Data

Step 1: Obtaining the Paired Comparisons Matrix

Choppin's (1968, 1985) conditional pairwise algorithm begins with the construction of a paired comparison matrix B, with entries b_{ij} representing the number of people who got item i right and item j wrong. This is the paired comparison matrix described in Figure 1. Each row represents a score of 1 on item i, and each column represents a score 0 on item j. Hence, the entry in the ith row and jth column represents the number of people who scored 1 on item i and 0 on item j. For example, suppose that a four-item test is administered to 10 people with results shown below:

		Items:			
		1	2	3	4
People:	1	1	0	1	1
	2	1	1	0	0
	3	1	0	0	0
	4	0	1	1	1
	5	1	1	1	0
	6	1	1	0	1
	7	1	1	1	1
	8	1	0	1	0
	9	1	1	1	1
	10	1	1	0	0

The paired comparison or B matrix for this data would be as follows:

		Items:			
		1	2	3	4
Items	1	0	3	4	5
	2	1	0	3	3
	3	1	2	0	2
	4	1	1	1	0

For example, the entry in row 1 and column 2 is 3, because three people (person #1, person #3, and person #8) got item 1 right and item 2 wrong. A full justification for this paired comparison matrix and its relevance to the item parameters of the Rasch model can be found in Garner and Engelhard (2000).

Step 2: Obtaining the Item Difficulties

Once the paired comparisons matrix has been established, Andrich (1988), Fischer and Tanzer (1994), Linacre (1989), van der Linden and Eggen (1986), Wright and Masters (1982), and Zwinderman (1995) then describe the likelihood of obtaining such a matrix and find the item parameters that maximize that likelihood. However, in addition to this maximum likelihood approach, Choppin (1968, 1985) described an alternative technique that is appealingly simple. In that technique, the paired comparisons matrix B is then converted to a matrix D with entries d_{ij} equal to b_{ji}/b_{ij}.

1	1/3	1/4	1/5
3	1	2/3	1/3
4	3/2	1	1/2
5	3	2	1

The matrix D is called a positive reciprocal matrix. D is then converted to ln D with entries ln (b_{ji}/b_{ij}). Choppin (1968, 1985) then showed that the item difficulties are simply the row means of the matrix ln D. For our example, the item difficulties would be –1.02, –.10, .27, and .85 for items 1 through 4.

Garner and Engelhard (2002) showed that the eigenvector associated with the maximum eigenvalue of the positive reciprocal matrix also generates the item difficulties. The properties of positive reciprocal matrices, as well as the associated eigenvectors, have been described by Saaty (1996). In particular, Saaty showed that the eigenvector can be obtained by calculating

Paired Comparisons Matrix for Dichotomous Items:

An item difficulty is generated for each item.

	Item:	1	2	3	4	
	Score:	0	0	0	0	
Item & Score:	1 1	*The entry in the ith row and jth column*				D_1
	2 1	*represents the number of examinees who*				D_2
	3 1	*got a score of 1 on item i and a score*			⟶	D_3
	4 1	*of 0 on item j*				D_4

Figure 1. The paired comparisons matrix for dichotomous data and the parameters generated by the pairwise algorithm. Each entry in the ith row and jth column represents the number of examinees who got item i correct and item j wrong.

the normalized row sums of the limiting power of the matrix. In other words, the eigenvector can be obtained by multiplying the matrix times a vector of 1's, normalizing the resulting vector (dividing each entry by the maximum value of all the entries in the vector), multiplying again by the matrix, normalizing the vector, multiplying again by the matrix, etc. The procedure stops when successive approximations of the eigenvector are close enough. The natural logarithm of the eigenvector is then calculated and the values are centered around a mean of 0. Using this procedure, the same item difficulties are obtained as above.

Harker (1987) showed that the eigenvector calculated as described above, using graph theoretical terms, represents the average of the intensities of all paths starting at a particular item. Thus, calculation of the entries in the eigenvector reflects comparisons with all the other items in the sample.

Powers of the Paired Comparison Matrix

The reciprocal matrix described above cannot be obtained when any of the B matrix entries are zero, which must be expected when the same person does not take two items or when persons always get both items right or both wrong. Thus, zero entries could appear in the matrix of paired comparisons even when there are no data missing. In the case of missing data, even more zero entries will be present. Choppin showed algebraically that the entries of B^2 rather than B may be used, thus replacing the results of the direct comparisons between i and j with the sum of the indirect comparisons of i and j through an intermediate k. If the items are adequately linked, all off-diagonal entries of the squared matrix will be non-zero. If zero entries still exist in the B^2 matrix, higher powers could be used.

Using powers of the paired comparison matrix is a common practice in the literature in the presence of incomplete paired comparisons or tournament matrices (Andrews and David, 1990; Cowden, 1975; Harker, 1987; Kendall, 1955). In terms of tournaments, the power of a tournament matrix represents a reallocation of wins; that is, each time an item i "wins" over another item j, all of item j's "wins" are reallocated to item i. The use of powers of the paired comparison matrix is discussed further in Garner and Engelhard (2000).

If the original data set is not adequately connected, that is, if comparisons between pairs of items cannot be made either directly or indirectly, then the powers of the matrix of paired comparisons will contain entire rows or columns of zeros. The rows or columns of zeros are associated with a particular item that is not adequately linked to the rest of the items. This would occur, for example, when all people get a particular item right or all get it wrong.

Paired Comparison Matrices for the Partial Credit Model with a Rater Parameter

In the partial credit model (PCM) (Andrich, 1978; Masters, 1982; Wright and Masters, 1982), a unique difficulty D_{ix} is associated with each category of each item. As in the many-facet Rasch model (Linacre, 1989), a separate parameter can be introduced to the model to reflect rater severity. In the resulting model, the probability that a student n will achieve a score x rather than a score x–1 on a certain item i is described as follows:

$$P_{nrix} / P_{nrix-1} = \exp(B_n - C_r - D_{ix}), \qquad (1)$$

where

P_{nrix} = the probability of achieving a score of x by person n with ability B from judge r with severity C on item i with difficulty D.

P_{nrix-1} = the probability of achieving a score of $x - 1$ by person n with ability B from judge r with severity C on item i with difficulty D.

B_n = the ability B of person n.

C_r = the severity C of judge r.

D_{ix} = the difficulty D of obtaining a score of x rather than $x - 1$ on item i.

Another approach to conceptualizing the Rasch model for categories is through the rating scale model (RSM). The RSM differs from the

PCM in that each item has a single associated item difficulty parameter and each transition from score $x - 1$ to score x has an associated threshold parameter that reflects the probability of achieving a score of x rather than $x - 1$ on any item. In this article, the RSM is considered a special case of the PCM, in which the item difficulty of the RSM associated with a specific item can be obtained by averaging all the PCM parameters associated with that item, and the threshold parameters of the RSM can be obtained by averaging all the PCM parameters associated with that transition from score x to score $x - 1$.

Estimation of Partial Credit Parameters. Consider two items i and j on which a person has scored an x and a y respectively. The probability of a person scoring an x rather than an $x - 1$ on item i is as described in equation (1). The probability of a person scoring a y rather than a $y - 1$ on item j is

$$P_{nrjy} / P_{nrjy-1} = \exp(B_n - C_r - D_{jy}). \quad (2)$$

To obtain a ratio involving item difficulties only, divide the above equations to obtain:

$$(P_{nrix} \times P_{nrjy-1}) / (P_{nrix-1} \times P_{nrjy}) = \exp(B_n - C_r - D_{ix}) / \exp(B_n - C_r - D_{jy}). \quad (3)$$

Upon cancellation, this implies that

$$(P_{nirx} \times P_{njry-1}) / (P_{nirx-1} \times P_{njry}) = \exp(D_{jy}) / \exp(D_{ix}). \quad (4)$$

The ratio in equation (4) can be estimated by the number of people who scored x on item i and $y - 1$ on item j divided by the number of people who scored $x - 1$ on item i and a y on item j, all of whom are rated by the same judge. Note that these people have the same ability in relation to these two items, since each person's raw score on items i and j sum to $x + y - 1$.

Figure 3 shows how the paired comparison matrix may be created. Each row represents an item i and a score x on that item. Each column represents an item j and a score $y - 1$ on that item. Each entry in the matrix then represents the number of people who scored an x on item i and a $y - 1$ on item j. When step 2 of the pairwise algorithm is executed, ratios will be formed that represent the number of people who scored an x on item i and a $y - 1$ on

Paired Comparisons Matrix for Partial Credit Model

(4 items, 3 thresholds)

A difficulty is generated for each threshold of each item.

Item: 1 2 3 4
Score: 0 1 2 0 1 2 0 1 2 0 1 2

Item & Score: 1 1	*The entry corresponding to row i, x*		D_{11}
2	*and column j, y-1 represents the number of*		D_{12}
3	examinees *who scored x on item i*		D_{13}
2 1	*and y-1 on item j.*		D_{21}
2			D_{22}
3			D_{23}
3 1		⟶	D_{31}
2			D_{32}
3			D_{33}
4 1			D_{41}
2			D_{42}
3			D_{43}

Figure 2. The paired comparisons matrix for the partial credit model.

Paired Comparisons Matrices for the Partial Credit Model With Rater Severity Parameter

(4 items, 3 thresholds, 4 raters)

An item difficulty is generated for each threshold of each item.

	Item: 1 / Score: 0 1 2	Item: 2 / Score: 0 1 2	Item: 3 / Score: 0 1 2	Item: 4 / Score: 0 1 2	
Item & Score: 1 1	*The entry corresponding to row i, x*				D_{11}
2	*and column j, y-1 represents the number of*				D_{12}
3	examinees *who scored x on item i*				D_{13}
2 1	*and y-1 on item j.*				D_{21}
2					D_{22}
3					D_{23}
3 1				→	D_{31}
2					D_{32}
3					D_{33}
4 1					D_{41}
2					D_{42}
3					D_{43}

A rater severity parameter is obtained separately.

	Rater: 1 / Score: x-1	Rater: 2 / Score: x-1	Rater: 3 / Score: x-1	Rater: 4 / Score: x-1	
Rater & Score 1 x	*The entry in the ith row and jth column*				C_1
2 x	*represents the number of examinees who*				C_2
3 x	*were rated one point higher by rater i than*			→	C_3
4 x	rater j.				C_4

Figure 3. The paired comparisons matrices for a partial credit model with a rater severity

item j divided by those who scored a y on item i and an $x - 1$ on item j. For example, the entry in row 2 and column 7 represents the number of people who scored a 2 on item 1 and a 0 on item 3; in this case, x is 2 and $y - 1$ is 0. On the other hand, the entry in row 7 and column 2 represents the number of people who scored 1 on item 3 and 1 on item 1; in this case, $x - 1$ is 1 and y is 1. Note that if the only scores are 0 and 1, this matrix reduces to exactly the same paired comparisons matrix that would exist in the dichotomous case.

Estimation of Rater Severity Parameters. The probability of rater r delivering a score x rather than an $x - 1$ to person n on item i is the same as in equation (1). The probability of a rater s delivering a score of x rather than $x - 1$ to person n on item i is:

$$P_{nsix} / P_{nsix-1} = \exp(B_n - C_s - D_{ix}). \quad (5)$$

To obtain a ratio involving rater severity only, divide equation (1) by equation (5) to obtain:

$$(P_{nrix} \times P_{nsix-1}) / (P_{nrix-1} \times P_{nsix}) = \exp(B_n - C_r - D_{ix}) / \exp(B_n - C_s - D_{ix}). \quad (6)$$

Upon cancellation, this implies that

$$(P_{nrix} \times P_{nsix-1}) / (P_{nrix-1} \times P_{nsix}) = \exp(C_s) / \exp(C_r). \quad (7)$$

The ratio described in (7) can be estimated by the ratio of the number of students who received a score of x from rater r and a score of $x - 1$ from rater s on the same item, to the number of students who received a score of $x - 1$ from rater r and a score of x from rater s on the same item.

A Rasch Model With a Single Parameter for Item, Threshold, and Rater

In this model, the probability that a student n will achieve a score x rather than a score $x - 1$ on a certain item i is described as follows:

$$P_{nrix} / P_{nrix-1} = \exp(B_n - D_{ixr}), \qquad (8)$$

where

P_{nrix} = the probability of achieving a score of x by person n with ability B from judge r with severity C on item i with difficulty D.

P_{nrix-1} = the probability of achieving a score of $x - 1$ by person n with ability B from judge r with severity C on item i with difficulty D.

B_n = the ability B of person n.

D_{ixr} = the difficulty D of obtaining a score of x on item i from judge r.

Justification for the elimination of the person parameter and construction of the matrix of paired comparisons would parallel equations (2), (3), (4) in the case of the partial credit parameters. The matrix of paired comparisons would take the form shown in Figure 4. Note that there are a greater number of matrix entries necessary than in the previous model.

Illustrative Data Analysis

Data

Consider the following example of 10 people each with scores on a scale of 0 to 2 on three items rated by each of three raters.

Items:	1			2			3		
Judges:	1	2	3	1	2	3	1	2	3
People: 1	2	1	1	2	2	0	1	0	0
2	2	2	1	1	1	1	2	1	1
3	1	1	0	0	0	1	0	0	0
4	0	0	1	1	0	1	1	1	1
5	1	1	2	2	1	1	2	1	1
6	2	1	1	2	2	0	0	0	0
7	2	2	2	2	2	2	1	1	2
8	1	1	0	0	0	1	1	1	1
9	2	2	1	2	1	0	1	2	1
10	2	1	0	1	0	0	0	0	0

Partial Credit Model with a Rater Parameter

To obtain the threshold parameters for each item, a matrix of paired comparisons would be formed as shown below and described in Figure 3.

	Item	1		2		3	
	Rating	0	1	0	1	0	1
Item	Rating						
1	1	0	0	8	3	7	6
	2	0	0	0	5	2	6
2	1	3	3	0	0	2	7
	2	0	3	0	0	3	4
3	1	3	6	4	7	0	0
	2	0	1	0	2	0	0

For example, the entry of 8 in the first row and third column represents the fact that 8 people scored a 0 on item 2 and a 1 on item 1 under the same judge. Those people are: person 1 under judge 3, person 3 under judge 1, person 3 under judge 2, person 6 under judge 3, person 8 under judge 2, person 8 under judge 1, person 9 under judge 3, and person 10 under judge 2. The entry of 0 in the sixth row and third column represents the fact that no one scored a 0 on item 2 and a 2 on item 3 under the same judge.

The square of the original paired comparison matrix, as seen below, has off-diagonal entries that are 0 (last row, first and third columns).

45	81	28	61	25	68
6	33	8	26	15	20
6	19	32	52	27	36
9	22	12	44	6	18
12	33	24	39	62	110
0	6	0	5	8	14

The cube of the matrix is then calculated, as shown below, and is found to have all non-zero entries off the diagonal.

159	485	460	851	716	1196
69	212	108	328	202	394
177	450	156	374	300	582
54	222	96	215	263	446
258	671	344	855	315	594
24	77	32	114	27	56

The reciprocal matrix is then formed:

1	69/485	177/460	54/851	258/716	24/1196
485/69	1	450/108	222/328	671/202	77/394
460/177	108/450	1	96/374	344/300	32/582
851/54	328/222	374/96	1	855/263	114/446
716/258	202/671	300/344	263/855	1	27/594
1196/24	394/77	582/32	446/114	594/27	1

The natural logarithms of the entries in the reciprocal matrix are calculated, and the item difficulties are then the averages of the rows of the matrices. The item difficulties are –1.76, .43, –.77, .73, –.76, 2.14. These values also represent the entries in the eigenvector associated with the maximum eigenvalue of the last matrix shown above.

The maximum eigenvalue associated with the eigenvector described above is 6.0624. Theoretically, the eigenvalue should be 6. Saaty relates the difference between the obtained eigenvalue and the theoretical eigenvalue with an index that reflects both the consistency of the reciprocal matrix and the error in the entries of the eigenvector. A reciprocal matrix D is consistent when, for any $1 \leq i < n$, $1 < k \leq n$, and $1 < j < n$,

$$d_{ij} \times d_{jk} = d_{ik}$$

where d_{ij} is the entry in the ith row and jth column of D. Saaty (1996) showed that a necessary and sufficient condition for the reciprocal matrix to be consistent is that the maximum eigenvalue λ_{max} be equal to N, where N is the dimension of the matrix. As a measure of deviation from consistency, the author uses a consistency index:

$$C.I. = (l_{max} - N)/(N - 1)$$

Saaty (1996) states that the matrix is satisfactorily consistent if the $C.I.$ is less than .1. This value is based on simulations with reciprocal matrices whose entries are randomly generated. This measure of consistency was applied in Garner and Engelhard (2002) and will be applied here. The dimension of the above matrix is 6 and the maximum eigenvalue is 6.0624, thus the consistency index is .012.

The paired comparison matrix for the rater severity parameters takes the form:

0	10	11
1	0	7
5	6	0

The reciprocal matrix takes the form:

1	1/10	5/11
10	1	6/7
11/5	7/6	1

The judge severity parameters can be obtained by taking the natural logarithm of each entry in the matrix and then averaging the rows. The parameters are then –1.03, .72, and .31. These values also represent the eigenvector associated with the maximum eigenvalue of the last matrix shown above. The maximum eigenvalue is 3.3173; hence the consistency index is .159.

Rasch Model With a Single Parameter for Item, Threshold, and Rater

The paired comparison takes the form shown below.

0	0	0	3	2	0	2	0	2	1	0	3	1	1	1	2	1	2
0	0	0	3	1	4	0	2	1	2	4	1	2	3	3	2	3	2
0	3	0	0	3	2	2	1	3	1	3	3	3	2	4	2	4	2
0	0	0	0	0	2	0	1	0	2	1	1	0	2	0	2	0	2
1	0	1	2	0	0	0	2	1	2	3	2	1	3	2	2	2	3
0	1	0	1	0	0	0	0	0	1	0	1	0	1	0	2	0	1
1	0	1	1	1	2	0	0	2	1	1	2	1	1	1	2	1	2
0	1	0	3	0	3	0	0	0	2	3	1	1	3	2	2	2	2
0	1	0	1	0	2	0	1	0	0	1	2	0	1	0	2	0	3
0	0	0	2	0	2	0	0	0	0	2	0	1	2	2	1	2	0
1	3	1	3	2	2	2	2	3	2	0	0	1	2	1	4	1	4
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
1	1	1	2	1	3	1	1	2	1	2	2	0	0	1	3	1	3
0	1	0	1	0	1	0	1	0	2	0	2	0	0	0	2	0	2
1	2	1	2	1	2	1	2	2	2	0	4	0	3	0	0	0	4
0	0	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	1
1	2	1	2	1	3	1	2	2	3	1	4	0	3	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0

The entries in the matrix are described in Figure 4. For example, the entry in the 10th row and 6th column is 2. This represents the fact that person #1 and person #6 scored a 2 on item 2 under judge 2, and a 1 on item 1 under judge 3.

The square of the paired comparison matrix has non-zero entries, so the cube of the original matrix is formed. The eigenvector associated with the largest eigenvalue of that matrix is shown below.

Parameter	Item	Rater	Threshold
–1.61	1	1	1
–1.12	1	1	2
–2.67	1	2	1
0.86	1	2	2
–1.59	1	3	1
1.52	1	3	2
–1.39	2	1	1
–0.77	2	1	2
0.03	2	2	1
0.32	2	2	2
–0.78	2	3	1
3.34	2	3	2
–1.33	3	1	1
1.22	3	1	2
–0.75	3	2	1
1.92	3	2	2
–0.94	3	3	1
3.73	3	3	2

The eigenvalue is 18.5778; hence the value of the consistency index is .034.

Simulation

Parameters

The parameters used are shown in Table 1. These parameters were obtained from a population of students described by Engelhard (1994). The ability distribution of the original population had a mean of 1.19 and standard deviation of 2.75. The parameters are consistent with a many-facet model that includes an item difficulty parameter, a threshold parameters, and a rater parameter. Each of 16 raters scored an essay on a scale of 0 to 3, on each of 5 domains.

Simulated Data

Person abilities were obtained by random sampling from a normal distribution with mean 0 and standard deviation 1. Responses from 200 persons, 500 persons, and 1000 persons were simulated. A random number between 0 and 1 was selected and compared to the probabilities generated by a given person ability, item difficulty, judge severity, and threshold parameter for each rating category. If the random number selected was greater than the cumulative probability for a certain rating n but less than the cumulative probability for the next category, the rating n was assigned.

Each simulation was conducted 10 times; in other words, a set of responses from 200 persons was generated 10 times, and on each occasion the parameters were recovered. Similarly, a set of responses from 500 persons was generated 10 times, and a set of responses from 1000 persons was generated 10 times. Those parameters were averaged over the 10 trials and compared to the original parameters.

Methods

Two methods were used for recovering the parameters. In the first method, described in Figure 3, a separate parameter was generated for each judge and for each threshold of each item; thus, 16 rater parameters and 15 item parameters were generated. To recover the 3 threshold parameters and the 5 item difficulty parameters of the original model, the 15 item parameters are averaged across thresholds and across items. In the second method, described in Figure 4, a separate parameter was generated for each combination of judge, item, and threshold; thus, 240 parameters were generated. To recover the 16 rater parameters, 3 threshold parameters and 5 item difficulty parameters of the original model, the 240 parameters were averaged across raters, across thresholds, and across items.

The simulation and all estimation methods were accomplished using SAS programs written by the authors.

Results

For the partial credit model with 3 threshold parameters for each item and a separate rater parameter, the results are shown in Table 1. For $N = 200$, the correlation between the original parameters and the recovered parameters was .9999; the correlations for $N = 500$ and $N = 1000$ were both 1.0000. The root mean square error (RMSE) improves as the size of the population increases. The mean of all parameters was 0 for both original and recovered sets; and the standard deviation for all sets was 1.60. Consistency indices for all data sets were all less than .1.

For the model with a parameter for each item, judge, and threshold combination, a fully connected matrix with non-zero entries could be obtained only by simulating a population with mean 0 and standard deviation 2.75. The results are shown in Table 2. Parameters were again successfully recovered. For $N = 200$, the correlation between the original parameters and the recovered parameters was .9995; the correlations for $N = 500$ and $N = 1000$ were both .9999. For $N = 1000$ and $N = 500$, the standard deviations of the parameters were 1.62, compared to the standard deviation of the original parameters which was 1.60. The standard deviation for $N = 200$ was 1.68. The means of the separate sets of parameters were all 0.00. Consistency indices for all estimates were less than .1.

Table 1

Parameter recovery for the partial credit model with a rater severity parameter. Each set of parameters represents an average over 10 trials.

		Parameter Values Recovered from Simulations					
		N = 200		*N* = 500		*N* = 1000	
	Original Parameter Values	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Raters	**0.97**	1.01	0.08	0.97	0.04	0.98	0.03
	0.91	0.92	0.06	0.94	0.04	0.92	0.03
	0.76	0.78	0.04	0.78	0.03	0.76	0.02
	0.76	0.77	0.09	0.77	0.06	0.76	0.02
	0.68	0.67	0.05	0.69	0.04	0.68	0.02
	0.45	0.44	0.07	0.45	0.05	0.45	0.02
	0.38	0.39	0.06	0.37	0.05	0.38	0.02
	0.25	0.26	0.07	0.27	0.05	0.25	0.02
	–0.15	–0.18	0.05	–0.15	0.03	–0.13	0.02
	–0.21	–0.18	0.06	–0.20	0.04	–0.22	0.02
	–0.30	–0.30	0.05	–0.29	0.02	–0.31	0.02
	–0.61	–0.62	0.05	–0.61	0.04	–0.61	0.03
	–0.68	–0.67	0.05	–0.68	0.04	–0.67	0.03
	–0.80	–0.80	0.07	–0.81	0.04	–0.79	0.03
	–1.10	–1.10	0.04	–1.13	0.04	–1.10	0.03
	–1.37	–1.40	0.05	–1.38	0.04	–1.37	0.03
Threshold	**–5.15**	–5.14	0.12	-5.15	0.07	–5.15	0.05
	0.34	0.34	0.08	0.35	0.03	0.34	0.02
	4.81	4.80	0.10	4.80	0.06	4.81	0.05
Items	**–0.32**	–0.28	0.14	–0.33	0.09	–0.36	0.07
	0.48	0.54	0.09	0.48	0.04	0.48	0.05
	–0.51	–0.53	0.14	–0.49	0.10	–0.49	0.06
	0.05	–0.02	0.10	0.02	0.05	0.05	0.07
	0.3	0.29	0.12	0.32	0.06	0.33	0.05
RMSE		.025		.014		.012	

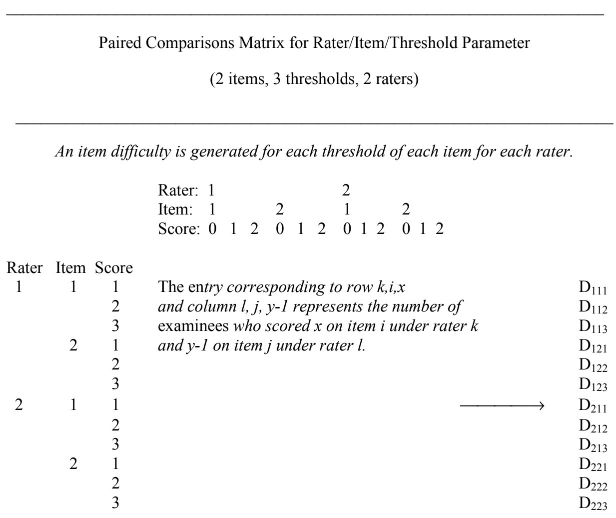

Figure 4. The paired comparisons matrix for a Rasch model with one parameter that reflects the difficulty associated with a combination of rater, item, and threshold.

Table 2

Parameter recovery for the model with one parameter for rater, threshold, and item. Each set of parameters represents an average over 10 trials.

	Original Parameter Values	Parameter Values Recovered from Simulations					
		N = 200		*N* = 500		*N* = 1000	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Raters	**0.97**	1.04	0.13	1.05	0.07	1.02	0.06
	0.91	1.01	0.13	0.91	0.08	0.94	0.07
	0.76	0.87	0.18	0.75	0.08	0.82	0.06
	0.76	0.92	0.15	0.74	0.08	0.78	0.06
	0.68	0.69	0.10	0.69	0.06	0.69	0.05
	0.45	0.53	0.17	0.47	0.09	0.45	0.04
	0.38	0.39	0.13	0.40	0.11	0.45	0.05
	0.25	0.22	0.14	0.26	0.07	0.26	0.05
	–0.15	–0.16	0.09	–0.11	0.04	–0.18	0.03
	–0.21	–0.23	0.11	–0.19	0.09	–0.22	0.07
	–0.30	–0.26	0.07	–0.29	0.07	–0.30	0.05
	–0.61	–0.63	0.17	–0.60	0.08	–0.62	0.06
	–0.68	–0.82	0.12	–0.70	0.11	–0.70	0.06
	–0.80	–0.84	0.12	–0.86	0.11	–0.81	0.07
	–1.10	–1.23	0.14	–1.14	0.06	–1.17	0.06
	–1.37	–1.50	0.14	–1.37	0.10	–1.35	0.03
Threshold	**–5.15**	–5.37	0.07	–5.20	0.03	–5.19	0.04
	0.34	0.37	0.05	0.35	0.02	0.36	0.02
	4.81	5.00	0.05	4.85	0.03	4.83	0.03
Items	**–0.32**	–0.38	0.05	–0.33	0.04	–0.32	0.02
	0.48	0.52	0.06	0.49	0.06	0.50	0.03
	–0.51	–0.55	0.07	–0.55	0.03	–0.52	0.03
	0.05	0.05	0.10	0.08	0.04	0.04	0.02
	0.3	0.37	0.08	0.31	0.02	0.30	0.02
RMSE		.087		.029		.030	

Discussion

This paper presents a new technique for estimating parameters of two formulations of the Rasch model for rater-mediated assessments involving ordered categories. It was shown that parameter values are contained in the eigenvector associated with the maximum eigenvalue of a reciprocal matrix derived from a matrix of paired comparisons between items or raters. The technique is very simple, yet it very effectively recovered simulated parameters. Furthermore, parameter estimation improved with increasing population size. The mathematical basis for the algorithm rests soundly in the methods of a popular technique for ordering preferences called the Analytic Hierarchy Process (Saaty, 1996).

Rasch (1966, 1977) repeatedly pointed out that a key characteristic of the Rasch measurement model is its specific objectivity—the property that the relative difficulty of any two items does not depend on the values of other parameters in the model. In other words, estimation of each parameter in the model could be performed separately from the other parameters. This specific objectivity is exploited by using paired comparisons matrices. Furthermore, Rasch stated that any good measurement model is based on objective pairwise comparisons.

> It is my opinion that only through systematic comparisons—experimental or observational—is it possible to formulate empirical laws of sufficient generabiliy to be—speaking frankly—of real value. (Rasch, 1977, p. 68-69)

Not only is the use of paired comparisons matrices close to the original intent of the Rasch measurement model, but the use of these matrices can make transparent connectivity within parameters. This aspect must be explored in future studies.

In this paper, standard errors for the parameter estimates were not obtained, nor were person ability estimates. Garner and Engelhard (2000, 2002) describe a bootstrap procedure for obtaining standard error estimates, and they also describe a maximum likelihood procedure for obtaining person parameters, thus completing the approach to parameter estimation outlined in this paper. Since item and rater parameter estimates are so close to original parameters, estimates of the ability parameters would be the same.

References

Andrews, D. M., and David, H. A. (1990). Nonparametric analysis of unbalanced paired-comparison or ranked data. *Journal of the American Statistical Association*, *85*(412), 1140-1146.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Choppin, B. (1968). Item banking using sample-free calibration. *Nature*, *219*(5156), 870-872.

Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education*, *9*, 29-42.

Cowden, D. J. (1974). A method of evaluating contestants. *The American Statistician*, *29*(2), 82-84.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall, International Thomson Publishing.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-facet Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Fischer, G. H. (1974). *Einfuhrung in die theorie psychologischer tests* (pp. 263-268). Berne: Verlag Huber.

Fischer, G. H., and Tanzer, N. (1994). Some LBTL and LLTM relationships. In G. H. Fischer and D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 277-303). New York: Springer Verlag.

Garner, M., and Engelhard, G. (2000). The method of paired comparisons, Rasch measurement theory, and graph theory. In M. Wilson and G. Engelhard (Eds.) (Vol. 5, pp. 259-286). *Objective measurement: Theory into practice.* Stamford, CT: Ablex.

Garner, M., and Engelhard, G. (2002). An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models. *Journal of Applied Measurement*, *3*, 107-128.

Harker, P. T. (1987). Incomplete pairwise comparisons in the Analytic Hierarch Process. *Mathematical Modelling*, *9*(11), 837-848.

Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, *11*, 43-62.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Psychology*, *4*, 321-333.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarfeld and N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89-107). Chicago: Science Research Associates.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58-94.

Saaty, T. L. (1996). *The analytic hierarchy process.* Pittsburgh: RWS Publications.

van der Linden, W. J., and Eggen, J. H. M. (1986). An empirical Bayesian approach to item banking. *Applied Psychological Measurement*, *10*(4), 345-354.

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

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*(4), 369-375.