DOCUMENT RESUME

ED 333 043 TM 016 583

AUTHOR Davison, Mark L.; Chen, Tsuey-Hwa
TITLE Parameter Invariance in the Rasch Model.
PUB DATE 3 Apr 91
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equations (Mathematics); *Estimation (Mathematics); Factor Analysis; Item Response Theory; *Mathematical Models; Mathematics Tests; *Regression (Statistics); Simulation; *Test Interpretation; Test Results
IDENTIFIERS Ability Estimates; *Invariance Principle; Item Parameters; Paired Comparisons; *Rasch Model

ABSTRACT

This paper explores a logistic regression procedure for estimating item parameters in the Rasch model and testing the hypothesis of item parameter invariance across several groups/populations. Rather than using item responses directly, the procedure relies on "pseudo-paired comparisons" (PC) statistics defined over all possible pairs of items. Methods of computing the PC statistics in non-independent and independent fashions are described. Two simulation studies were conducted. Both studies used a 2 x 2 factorial design in which the number of items (6 or 11) and the number of subjects (100 or 500) varied. There were 100 replications in each cell of the design, and for each replication, two samples of ability parameters were randomly drawn from a standard normal distribution. In the first study, the PCs were computed in a non-independent fashion. In the second study, the PCs were computed in an independent fashion; however, only the two cells involving 500 subjects had been analyzed to date. The results of these studies suggest that the procedure yields negligibly biased estimates of item difficulty parameters even with small numbers of items. The simulation data were used to compare the distribution of observed test statistics under the null hypothesis of invariant item parameters across groups to the theoretical Student's t-distribution and the theoretical chi-square distribution. An application to sixth-grade mathematics achievement data for 178 fall mathematics test takers and 153 spring mathematics test takers is presented. Five data tables and a 15-item list of references are included. (Author/RLC)

Parameter Invariance in the Rasch Model

Mark L. Davison and Tsuey-Hwa Chen

University of Minnesota

Paper presented to the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3, 1991. Requests for copies should be sent to Mark L. Davison, Department of Educational Psychology, University of Minnesota, 178 Pillsbury Dr. S.E., Minneapolis MN 55455.

Abstract

We explored a logistic regression procedure for estimating item parameters in the Rasch model and testing the hypothesis of item parameter invariance across several groups. Rather than utilizing item responses directly, the procedure relies on "pseudo-paired comparisons" (PC) statistics defined over all possible pairs of items. We describe methods of computing the PC statistics in nonindependent and independent fashions. Simulation results suggest that the procedure yields negligibly biased estimates of item difficulty parameters even with small numbers of items. The simulation data were used to compare the distribution of observed test statistics under the null hypothesis of invariant item parameters across groups to the theoretical Student's t-distribution and the theoretical chi-square distribution. An application to sixth grade mathematics achievement data is presented.

Parameter Invariance in the Rasch Model

Starting from assumptions of the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959), McGuire and Davison (in press) described a logistic regression procedure for estimating stimulus scale values in true paired comparisons (PC) data. The procedure also yields a chi-square test of the BTL model. When paired comparisons data are available from two or more populations, the logistic regression approach can be extended to a test of the hypothesis that the PC data conform to the BTL model with equal stimulus scale values in all populations.

Andrich (1978), Chen and Davison (1991), Choppin (1968), and Rasch (1960) have discussed a "pseudo"-paired comparisons statistic defined over pairs of dichotomously scored test items. For item pair (j,k), the PC statistic is the conditional probability of passing item j and not item k given that the subject passes exactly one of the two items. When the items conform to the Rasch model, the "pseudo-"paired comparisons data have the form specified by the BTL model for true PC data. The stimulus scale values in the BTL expression for the pseudo-PC statistic π_{jk} will equal the Rasch model item difficulty parameters. If the Rasch model holds with equal item difficulty parameters in two or more populations, then the pseudo-PC matrix for each population will have the form specified by the BTL model for true paired comparisons data, and the stimulus scale values will be equal for all populations.

The PC statistics for item pairs are readily computed. Unlike true PC data, the PC statistics as described by the above authors, are <u>not</u> independent. That is, for two pairs (j,k) and (j,k'), the responses to item j will be used in computing both the PC statistic for pair (j,k) and pair (j,k') and hence the estimates of PC statistics for these two pairs will <u>not</u> be independent.

In the research reported here, we extended the McGuire and Davison (in press) approach for true PC data to the analysis of pseudo-PC data for item pairs satisfying the

Rasch model. That is, we investigated using a standard logistic regression analysis of pseudo-PC statistics to estimate item difficulty parameters and to test the often discussed hypothesis of invariant item difficulty parameters across populations (Hambleton & Swaminathan, 1985; Lord, 1980).

To evaluate the effect of dependencies between PC statistics for item pairs, a small simulation study was conducted. The effect of the dependencies on test statistics was substantial. Consequently, pursuant to a suggestion by van den Wollenberg, Wierda, and Jansen (1988), we explored a method of computing PC statistics in an independent fashion. When computed in an independent fashion, each item response is used once and only once in computing the PC statistics.

We begin by discussing the adaptation of the logistic regression approach to the analysis of pseudo-PC data. This section describes computation of the PC statistics in both independent and nonindependent fashions. Then we report results of two small simulation studies designed to evaluate the effect of dependencies when the null hypothesis is true; that is, the item responses fit the Rasch model with equal item parameters in two populations. In Study 1, PC statistics were computed in a nonindependent fashion, whereas in Study 2 they were computed in an independent fashion. Finally, we report an illustrative application to sixth grade mathematics items. A more extensive evaluation of the logistic regression approach will be contained in Chen (in progress), including a comparison of logistic regression results to those obtained by unconditional maximum likelihood estimation.

The Logistic Regression Analysis of Pseudo-PC Statistics

Let the discrete random variable a_{vj} take the value 1 when subject v answers item j correctly and take the value 0 otherwise. Then according to the Rasch model

$$\pi(a_{vj} = 1) = \exp(\Theta_v - b_j)/[1 + \exp(\Theta_v - b_j)] \qquad (1)$$
$$\pi(a_{vj} = 0) = 1/[1 + \exp(\Theta_v - b_j)],$$

where Θ_v is the ability parameter for subject v and b_j is the difficulty parameter for item j.

The Paired Comparisons Statistic for Items

For items, the PC statistic is defined over the set of all possible item pairs. Our development of the statistic closely follows earlier work (Andrich, 1978; Chen & Davison, 1991; Choppin, 1968; Rasch 1960) and is presented only for completeness.

Defined over item pairs, the PC statistic is a function of two joint probabilities, the probability that the subject passes item j and fails item k and the probability that the subject fails item j and passes item k: $\pi(a_{vj} = 1, a_{vk} = 0)$ and $\pi(a_{vj} = 0, a_{vk} = 1)$. In terms of these two joint probabilities, the PC statistic is defined as

$$\pi_{jk} = \frac{\pi(a_{vj} = 1, a_{vk} = 0)}{\pi(a_{vj} = 1, a_{vk} = 0) + \pi(a_{vj} = 0, a_{vk} = 1).} \qquad (2)$$

In words, the PC statistic for item pair (j,k) is the probability that subject v passes item j and fails item k given that the subject passes only one item of the pair.

From the assumption of local independence in the Rasch model, these two joint probabilities have the following form:

$$\pi(a_{vj} = 1, a_{vk} = 0) = \frac{\exp(\Theta_v - b_j)}{[1 + \exp(\Theta_v - b_j)]\,[1 + \exp(\Theta_v - b_k)]} \qquad (3)$$
$$\pi(a_{vj} = 0, a_{vk} = 1) = \frac{\exp(\Theta_v - b_k)}{[1 + \exp(\Theta_v - b_j)]\,[1 + \exp(\Theta_v - b_k)].}$$

Inserting Equation 3 into the definition in Equation 2 yields

$$\pi_{jk} = \exp(b_k - b_j)/[1 + \exp(b_k - b_j)]. \quad (4)$$

Readers familiar with the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) will recognize that the probability π_{jk} in Equation 4 stands in the same relationship to item difficulties, b_k and b_j, as do the choice probabilities in the BTL model to the stimulus scale values.

As can be seen from Equation 4, the PC statistic depends only on the item difficulties, b_k and b_j, and is independent of the ability parameters. Consequently, a sample estimate of π_{jk}, say p_{jk}, can be computed by counting the number of people who passed item j but not k and dividing that count by the number who passed exactly one of the two items.

Table 1 shows the PC statistics from a fall (n = 178) and spring (n = 153) testing for ten items from a sixth grade mathematics test developed by a large suburban school district. Each data point shows the proportion of students who passed the row item among the students who passed exactly one of the two items. For instance, the data point in row 2, column 1 equals .039, because 4 of the fall subjects passed item 2 but not item 1 and 103 subjects passed either item 1 or item 2 but not both. Hence $p_{jk} = 4/103 = .039$.

If the PC statistics are computed as described above -- counting the number of people who passed item j but not k and dividing by the number who passed exactly one of the two items -- then the PC statistics will <u>not</u> be independent. For item pairs (j,k) and (j,k'), exactly the same responses to item j are used to compute sample estimates p_{jk} and $p_{jk'}$. To insure the independence of the PC statistics, they need to be computed so that each data point is used no more than once.

Let n be the total number of items on the test and let t be the number of items correctly answered by a given subject. According to van den Wollenberg et al. (1988), a subject's data can be used to estimate a maximum of min[t, (n-t)] PC statistics without using any of the subject's responses more than once. That is, min[t, (n-t)] equals the maximum

number of item pairs which can be formed in which the subject passed exactly one item of the pair. The van den Wollenberg et al. result led us to the following approach for estimating PC statistics.

If min[t, (n-t)] = t, then we randomly paired each item answered correctly with one and only one of the items answered incorrectly. If pair (j,k) were one of these random item pairings for the current subject, then we would increase by 1 our count of the number of people passing exactly one member of the item pair. If the person had passed j but not k, we would also increment our count of people who passed j but not k. After repeating this process for each subject, then for every item pair (j,k), we would divide our count of subjects passing j but not k by our count of people passing exactly one of the two items. If min[t, (n-t)] = (n-t), the process would proceed in the same fashion, except that (n-t) item pairs were created by randomly matching an incorrectly answered item to a correctly answered item. By following this process, one can use each subject's data to estimate as many PC statistics as possible without using any response more than once.

The Logistic Regression Model

Since $\pi_{jk} + \pi_{kj} = 1$, J items will define J(J - 1)/2 unique PC probabilities. Thus, we need to work only with the lower triangular elements of the PC matrix for which the column is less than the row; i.e. k < j. Throughout the remainder of this section we shall proceed as if the lower triangular elements of the PC matrix were independent.

The logit of the probabilities in Equation 4 can be expressed as a linear function of known predictor variables in a design matrix. The unknown regression weights on the predictor variables are the item difficulties. In other words, the logits of the J(J - 1)/2 item PC statistics define a system of J(J - 1)/2 equations which are linear functions of elements in a design matrix. Hence, given sample estimates of the PC statistics, logistic regression can be used to obtain maximum likelihood estimates of the item difficulty parameters and

to estimate the fit of the logistic regression model to the PC statistics. We begin by developing the design matrix for a single population. Subsequently, we extend the PC model and the design matrix to more than one population.

The Single Group Model. From the equality in Equation 4, it follows that the logit, $L_{jk} = \ln[\pi_{jk}/(1 - \pi_{jk})] = (b_k - b_j)$. To uniquely determine the item difficulties, let's impose the restriction that the sum of the item difficulties equals zero: $\Sigma_j b_j = 0$. This restriction implies that

$$b_1 = -\Sigma_{i \neq 1} b_i. \tag{5}$$

In developing subsequent expressions for the logit L_{jk}, such as Equation 6, the subscripts j and k will designate the two stimuli corresponding to the logit, and the subscript i (i = 1,...,J) will be used to designate stimuli in summations running over stimuli. Equation 5 suggests that the logits L_{jk} can be written in terms of difficulties for items 2,...,J as follows:

$$L_{jk} = -2b_j - \Sigma_{i \neq 1,j} b_i \text{ when } k = . \tag{6}$$

and

$$L_{jk} = b_k - b_j \text{ when } k \neq 1.$$

The design matrix will have J(J - 1)/2 rows, one for each unique PC statistic; and it will have J - 1 columns, one column for items 2,...,J. Let the subscript i designate a column of the design matrix, and let d_i refer to the i^{th} element in the row corresponding to item pair (j,k). Then the elements of the design matrix in the row corresponding to item pair (j,k) are set equal to the following values when k = 1:

$d_i = -1$ for $i = 2,...,J$ and $i \neq j$, (7a)

$d_i = -2$ for $i = j$.

When $k \neq 1$, the elements of the design matrix in the row corresponding to item pair (j,k) are set equal to the following values:

$d_i = 1$ for $i = k$, (7b)

$d_i = -1$ for $i = j$, and

$d_i = 0$ otherwise.

With the elements of the design matrix defined as in Equations 7a and 7b, Equation 6 can be rewritten as

$$L_{jk} = \Sigma_{i \neq 1} b_i d_i . \quad (8)$$

Table 2 shows the design matrix for all possible pairs of four items. For the first three rows, $k = 1$, and hence the elements in these three rows follow the pattern described in Equation 7a. For the last three rows, $k \neq 1$, and hence the elements in these three rows follow the pattern described in Equation 7b.

Since the logits have the linear form of Equation 8, the expression for the PC statistic in Equation 4 is a logistic regression equation without an intercept term in which the criterion variable is the PC statistic π_{jk}, the predictor variables are the dummy variables d_i, and the regression weights correspond to the unknown item parameters b_i. One of the increasingly popular logistic regression algorithms can be used to obtain maximum likelihood estimates of the regression coefficients. These coefficient estimates constitute

maximum likelihood estimates of item difficulty parameters for items 2,..., J. Once difficulty estimates for items 2,..., J have been obtained, the parameter for item 1 can be estimated according to Equation 5. A chi-square measure of fit can be used to test the hypothesis that the PC statistics fit the model in Equation 4.

Multiple Population Rasch Model. The Rasch model can be extended to more than one population by rewriting Equation 1 in terms of probabilities and item parameters specific to each group. Let g (g = 1,...,G) be a subscript designating a group. In terms of group specific item parameters, b_{jg}, the probabilities of correct and incorrect responses to item j from a member of group g with ability θ_v equal

$$\pi_g(a_{vj} = 1) = \exp(\theta_v - b_{jg})/[1 + \exp(\theta_v - b_{jg})] \qquad (9)$$
$$\pi_g(a_{vj} = 0) = 1/[1 + \exp(\theta_v - b_{jg})].$$

Using the reasoning leading from Equation 1 to Equation 4, we arrive at the conclusion that the PC statistic for item pair (j,k) in group g, π_{jkg}, is a logistic function of group specific item parameters:

$$\pi_{jkg} = \exp(b_{kg} - b_{jg})/[1 + \exp(b_{kg} - b_{jg})]. \qquad (10)$$

Obviously, for the logits in group g,

$$L_{jkg} = b_{kg} - b_{jg}. \qquad (11)$$

To uniquely determine the item parameter scale, assume that the sum of item difficulties equals 0 in each group so that for every group, the difficulty of item 1 can be expressed in terms of difficulties for the remaining items.

$$b_{1g} = -\Sigma_{i\neq 1} b_{ig} \qquad (12)$$

We will now develop a form for the logits which expresses them as a linear function of known elements in a design matrix and unknown regression coefficients. At the level of a global hypothesis test, this design matrix leads to a test of the hypothesis that PC statistics satisfy the BTL form with equal item difficulties in all groups. That is, if we let $\underline{b}_g$ be the vector of J item difficulties for group g, then the hypothesis states that there exists a vector of item difficulty values $\underline{b}$ such that $\underline{b}_g = \underline{b}$ for all g. At the level of a hypothesis about a specific item, the particular design matrix described here leads to a test of the following hypothesis. For group g, item j and a designated target group (say group g = 1), the null hypothesis tested is that $b_{jg} = b_{j1}$. That is, item j has the same difficulty in group g as it does in the target group.

According to Equation 6, the logits in the target group g = 1 can be written as

$$L_{jk1} = -2b_{j1} - \Sigma_{i\neq 1,j} b_{i1} \text{ when } k = 1 \qquad (13a)$$

and

$$L_{jk1} = b_{k1} - b_{j1} \text{ when } k \neq 1. \qquad (13b)$$

Let $b^*_{jg} = b_{jg} - b_{j1}$ be the difference between the scale value of item j in group g and the target group. Then for any group other than the target group, g = 2,...,G, Equation 6 and the definition of b^*_{jg} suggest that the logit for group G can be expressed as

$$L_{jkg} = -2b_{j1} - \Sigma_{i\neq 1,j} b_{i1} - 2b^*_{jg} - \Sigma_{i\neq 1,j} b^*_{ig} \text{ for } k=1 \qquad (14a)$$

$$= b_{k1} - b_{j1} + b^*_{kg} - b^*_{jg} \text{ for } k \neq 1. \quad (14b)$$

Because by definition $b^*_{jg} = b_{jg} - b_{j1}$, Equation 14a reduces to $L_{jkg} = -2b_{jg} - \Sigma_{i \neq 1,j} b_{ig}$ when $k = 1$ and Equation 14b reduces to $L_{jkg} = b_{kg} - b_{jg}$ for $k \neq 1$.

Equations 13 and 14 can be rewritten as a linear function of unknown coefficients, b_{jg} and b^*_{jg}, and known elements in a design matrix. The design matrix will have $GJ(J - 1)/2$ rows, where G is the number of groups. That is, it will have one element for each probability p_{jkg}. Further, it will have $G(J - 1)$ columns, one for each item 2,...,J in each of the G groups. Let d_{ih} be the indicator variable for item i ($i = 2,...,J$) in group h ($h = 1,...,G$). Hereafter, the subscript g will be used to designate the group associated with L_{jkg}, and the subscript h will be used to designate a group in various summations running over groups. Table 3 will be used to illustrate a design matrix for three groups and four items.

For observations p_{jkg} such that $g = 1$ (for example, rows 1 - 6 in Table 3), predictors d_{i1} ($i = 2,...,J$) are defined as in Equations 7a and 7b, and all predictors with $h > 1$ equal 0. For observations L_{jkg} such that $g \neq 1$ (for example, rows 7 - 18 in Table 3), predictors d_{i1} ($i = 2,...,J$) are defined as in Equations 7a and 7b. Furthermore, for $h = g$, d_{ih}, is defined as in Equations 7a and 7b. When $h \neq 1$ and $h \neq g$, then $d_{ih} = 0$.

Having thus redefined the elements of the design matrix, Equations 13 and 14 can be combined in the expression

$$L_{jkg} = \Sigma_{i \neq 1} b_{i1} d_{i1} + \Sigma_{h \neq 1} \Sigma_{i \neq 1} b^*_{ih} d_{ih}. \quad (15)$$

Equation 15 expresses the logit as a linear equation with no intercept term. Using logistic regression, the unknown coefficients in the equation, b_{i1} and b^*_{ih}($i = 2,...J; h = 2,...,G$)can be estimated by regressing the sample estimates of probabilities p_{jkg} onto the elements of the design matrix.

Interpretation of the coefficient estimates in Equation 15 is different depending on whether the coefficient is associated with the target group. Specifically, the coefficient estimates associated with the target group are estimates of item parameters b_{i1} (i = 2,...,J). From Equation 12, the difficulty of item 1 can be estimated as $b_{11} = -\Sigma_{i \neq 1} b_{i1}$. Regression weights associated with predictors in other groups, g ≠ 1, will be estimates of the deviations $b^*_{jg} = b_{jg} - b_{j1}$. Consequently, the individual scale values for items 2 through J in non-target group g can be estimated as $b_{jg} = b^*_{jg} + b_{j1}$, and the scale value for item 1 in non-target group g can be estimated as $b_{1g} = -\Sigma_{i \neq 1} b_{ig}$.

For each logistic regression coefficient, there will be an asymptotic standard error and a t-statistic, the regression coefficient divided by its asymptotic standard error. For coefficients associated with group 1, the j^{th} t-statistic potentially provides a test of the null hypothesis $b_{j1} = 0$, the average item difficulty. In groups 2 - G, the j^{th} coefficient for group g is an estimate of the difference between the difficulty of item j in group g and group 1. Hence, the t-statistic provides a test of the null hypothesis $b_{jg} - b_{j1} = 0$. For each item, except item 1, logistic regression leads to a formal test of the hypothesis that the item's difficulty in group g is the same as its difficulty in the target group.

The submodel in which item difficulties are invariant across groups is one in which $b^*_{ih} = 0$ for all (i,h), and hence Equation 15 reduces to

$$L_{jkg} = \Sigma_{i \neq 1} b_{i1} d_{i1} \quad . \tag{16}$$

According to this submodel, the logit of every group can be expressed in terms of common item difficulty parameters represented by the variable b_{i1} (i = 2,...,J). This model can be fit to the PC statistics using standard logistic regression. The analysis will yield estimates of the common item parameters b_{i1} (i = 2,...,J), an asymptotic standard error for each parameter, and a chi-square measure of fit with (GJ - 2)(J - 1)/2 degrees of freedom. This

fit measure forms the basis for deciding whether the data can be reasonably fit with invariant item parameters across groups.

Simulation Studies

Study 1

This study used a 2 x 2 factorial design in which the factors varied were number of items (n = 6 or 11) and number of subjects (N = 100 or 500). There were 100 replications in each cell of the design. For each replication, two samples of ability parameters were randomly drawn from a standard normal distribution. The two samples within each replication simulate the situation in which there are two samples from different populations for which the Rasch model holds and item parameters are equal in the two groups.

For cells with six items, the item difficulties were fixed at -1.70, -1.02, -.34, .34, 1.02, and 1.70. For cells with eleven items, the difficulties were set at -1.70, -1.36, -1.02, -.68, -.34, 0.00, .34, .68, 1.02, 1.36, and 1.70. Responses were simulated according to Equation 1 using a program developed by Yoes (1987), and the PC statistics were computed in a nonindependent fashion. The logit option in the SPSS-X Probit program was used to fit the submodel and the full model to the PC statistics for each replication in each cell.

Table 4 summarizes the results of this study. The first two panels of the first two sections in Table 4 contain data on parameter estimates from both the submodel and the full model. Comparing the true item parameters to the mean estimates suggests that there is a slight bias, particularly with 100 subjects and estimates based on fitting the full model. The direction and magnitude of the bias is related to the sign and absolute value of the item parameter. However, comparing the observed standard deviations, s.d($\hat{b}$), of parameter estimates to the root-mean-square asympto. standard error estimates, RMS $\hat{\sigma}(\hat{b})$, shows

that the asymptotic estimates systematically underestimated the true standard deviations of difficulty and difference parameter estimates, and that the degree of underestimation seems to have increased with an increase from six to eleven items.

The mean observed fit measures, shown at the right in Table 4, were smaller than the degrees-of-freedom, the expected value of the appropriate theoretical chi-square distribution. Except for the full model, the observed standard deviation of the fit measure was larger than $(2df)^{1/2}$, the standard deviation of the appropriate theoretical chi-square distribution. For both the submodel and the full model, less than 5% of the observed sample fit measures exceeded the critical value (alpha equal .05) of the corresponding theoretical chi-square distribution. This would seem to suggest that comparing the fit measure to the theoretical chi-square critical value for alpha equals .05 yields a very conservative test of the submodel and full model null hypotheses; that is, a true rate of rejection less than alpha. As a result, an observed fit measure which exceeds the critical value provides evidence disconfirming the model. However, an observed fit measure less than the critical value is not strongly supportive of the model.

The observed mean fit-difference statistic, on the other hand, appeared to be extremely large compared to the expected value for the theoretical chi-square distribution, leading to a liberal test of the invariance hypothesis; that is, a true rejection rate larger than the nominal alpha. The dependencies among the PC statistics appear to have caused the full model to fit "too well," compared to the fit for the submodel; therefore, a fit-difference above the critical value is not a valid indicator of parameter variation across groups when results are based on dependent PC statistics.

The last two sections of Table 4 show data on the estimated differences in item difficulties across groups as obtained from fitting the full model. Comparing the mean observed difference estimates [labeled Mean ($\bar{b}$) in Table 4] to the true differences of zero

suggests little, if any, systematic bias in estimates of item parameter differences. However, the root-mean-square asymptotic standard error estimates [labeled RMS $\hat{\sigma}(\hat{b})$] were smaller than the standard deviation s.d. ($\hat{b}$) of difference parameter estimates, indicating that the asymptotic estimates underestimated the variation in item parameter differences across replications. As with the item parameter estimates, the degree of underestimation increased with the increase from six to eleven items.

The t-statistic, the difference estimate divided by the asymptotic standard error, has too large a standard deviation. Hence, the proportion of t-statistics which exceeded the critical value in a Student's t-distribution was larger than alpha. As a means of testing the hypothesis that the difficulty of item j in group g equals that in the target group, the t-statistic provides a liberal test: that is, the true rejection rate under the null hypothesis is greater than alpha. It is our conclusion that the number of items with t-statistics above the critical value must be viewed as an upper bound on the number of items for which the difference is truly significantly different from zero at the chosen significance level.

Study 2

In Study 2, the PC statistics were computed in an independent fashion. Otherwise, Study 2 is exactly like Study 1. At present, the analysis has been completed only for the two cells involving 500 subjects, so only data from this condition are reported here.

The bottom panel of each section in Table 4 summarizes the results of this study. The first two sections give data on parameter estimates from the submodel and the full model. Comparing the true item parameters to the mean estimates suggests that there is little, if any, bias in the estimates of item difficulty parameters and the difference parameters, even with only six items. The observed standard deviation of the parameter estimates was larger in Study 2 than in Study 1, indicating a smaller error of estimation when PC statistics are computed nonindependently.

Comparing the observed standard deviations of parameter estimates in Study 2 to the root-mean-square asymptotic standard error estimates shows that, unlike in Study 1, the root-mean-square asymptotic standard error closely approximated the actual standard deviation of the parameter estimates.

The mean observed fit measure for the submodel, the mean observed fit measure for the full model, and the mean difference between these two fit measures, shown at the right in Table 4, are all slightly smaller than their degrees of freedom, the mean of the corresponding chi-square distribution. Furthermore, the standard deviations are all smaller than those of corresponding chi-square distributions. For a .05 level of significance, 5% or less of the observed submodel fit measures, full model fit measures, and fit difference statistics exceeded the chi-square critical value. These results suggest that comparing any of these three fit measures to its corresponding chi-square distribution yields a conservative test; that is, a true rate of rejection less than alpha when the null hypothesis is true. As a result, an observed fit measure which exceeds the critical value provides disconfirmation of the corresponding hypothesis. However, an observed fit measure less than the critical value is not strongly supportive of the model.

The bottom two sections of Table 4 show data on the estimated differences in item difficulties across groups as obtained from fitting the full model to independent PC statistics. Comparing the mean observed difference estimates to the true differences of zero suggests little, if any, systematic bias in estimates of item parameter differences. The root-mean-square asymptotic standard error again closely approximated the actual standard deviation of the parameter estimates. The t-ratio, the difference estimate divided by its asymptotic standard error, has slightly too small a standard deviation. Hence, the proportion of t-statistics which exceeded the critical value in Student's t-distribution at the .05 level was generally smaller than .05. As a means of testing the hypothesis that the

difficulty of item j in group g equals that in the target group, the t-statistic provides a conservative test: that is, the true rejection rate under the null hypothesis appears to be slightly less than alpha. It is our conclusion that the number of items with t-statistics above the critical value must be viewed as a lower bound on the number of items for which the difference is truly significantly different from zero at the chosen alpha level.

Example

The data for this example come from a fall and spring administration of a sixth grade mathematics test developed by a suburban school district in the upper midwest. Our two groups are the fall (N = 178) and spring test takers (N = 153). Table 5 shows results from a logistic regression analysis of the independent PC statistics. While there were 50 items on the test, we have presented a detailed analysis for only 10 of the items to keep the illustration small.

Column 2 of Table 5 shows the logistic regression estimates of the item parameters for the submodel. Parameters for items 2 - 10 in the target group (fall testing) are the logistic regression coefficients. The parameter for item 1 is the negative sum of item parameters (or regression coefficients) for items 2 - 10. The goodness-of-fit statistic for the submodel was 90.24 with 80 degrees-of-freedom ($p > .05$). (There are 80 degrees of freedom for this analysis, rather than 81, because it was run using the logit option in the SPSS-X Probit program, which always estimates an intercept constant rather than forcing the regression through the origin, and hence which leaves one less degree of freedom.) Columns 5 - 8 of Table 5 contain results from the logistic regression based on the full, multiple group model, for which the goodness-of-fit statistic was 63.33 with 71 degrees of freedom ($p > 05$). The difference in the fit measures for the full and submodels is 26.91 with 9 degrees of freedom ($p < .05$). The bottom panel of Table 5 shows results for the spring testing. Column

5, labeled regression coefficients, contains estimates of the difference between the items' difficulties in the fall and spring testings. The t-statistic associated with each item 2 - 10 provides a test of the null hypothesis that the difference between the item's parameter estimate for fall and spring testings equals 0. Using a .05 level of significance, the hypothesis would be rejected for three items.

Overall, the fit measures indicate that the PC statistics can be approximated reasonably well when item parameters are constrained equal for the fall and spring testings, but there is a substantial improvement in fit when this constraint is removed. The significant misfit seems to center on items 3, 4, and 7. After inspecting an earlier analysis of these data, Wright (personal communication) has suggested that the regression of the parameter estimates in the fall testing onto those for the spring testing is well fit by a straight line, but possibly with slope unequal to 1.00.

Discussion

The following conclusions are heavily based on limited simulation results involving only two test lengths, two sample sizes (one sample size in Study 2), and only one distribution of item and person parameters in each condition. In each condition, the null hypothesis of equal item parameters across populations was true. Nevertheless, the results from these limited conditions are suggestive of the following conclusions.

The results in Study 1 suggest that, despite dependencies among the PC statistics, the logistic regression approach appears to yield only slightly biased estimates of item parameters. The bias appears to decrease as the number of items and/or the number of subjects increase. While our focus is not on parameter estimation, a word on this subject seems in order. Van den Wollenberg et al. (1988) have criticized unconditional maximum likelihood (UML) estimates of item parameters as biased, a bias which cannot be removed

by a correction factor. Those authors argue that the computationally slow, conditional maximum likelihood (CML) estimation procedure and the minimum chi-square methods yield unbiased estimates.

In response to van den Wollenberg et al. (1988), Wright (1988) argued that the bias in UML estimates can be removed by a correction factor, except for short tests. He further argued that the pairwise estimation procedure is a better solution if more precision is desired for short tests. Our results in Study 1 tend to support Wright's argument. A comparison of our results with the most comparable results from van den Wollenberg et al. (See their Table 1.) shows that their ratio of estimated to true item parameters from CML is virtually the same as that found in Study 1 for logistic regression and 500 subjects.

When nonindependent PC statistics were used, the logistic regression asymptotic standard errors tended to underestimate the true standard deviation of parameters across replications. When independent PC statistics were used, the asymptotic standard errors far more accurately reflected the variation of parameter estimates across replications.

The hypothesis testing statistics, the t-ratio and the fit measure, deviated systematically from their theoretical counterparts, Student's t and the chi-square distribution. Hence, one cannot make precise statements about the probability of Type I errors simply by referring to either Student's t or the chi-square distribution. On the other hand, the deviations from the theoretical distributions were systematic in ways which suggest that the theoretical distributions might provide useful benchmarks for interpreting the t-ratio and the fit measure.

Consider first the t-ratio. When estimated from nonindependent PC statistics, the observed t-ratio consistently displayed a larger standard deviation than that of Student's t. Consequently, for alpha equal .05, more than 5% of the observed t-ratios exceeded the critical value in Student's distribution. This suggests that the number of items for which

the observed t-ratio exceeds the critical valu- in Student's distribution might be interpreted as an upper bound on the number of items for which the difference in item parameters is truly significantly different from zero at the chosen level of significance.

When independent PC statistics were used, the number of items exceeding the critical value behaved more like a lower bound on the number of items for which the difference in item parameters is truly significantly different from zero at the chosen significance level. The t-ratios displayed slightly smaller standard deviations than that of Student's distribution. For alpha equal .05, less that 5% of the observed t-ratios exceeded the critical value. For the conditions investigated in this study, it would seem that any t-ratio exceeding the critical value could be considered significant at the alpha level.

Turning now to the submodel fit measure for nonindependent PC statistics, that measure had a much smaller mean and a larger standard deviation than the corresponding theoretical chi-square distribution. Consequently, using a .05 level of significance, the observed fit measure exceeded the critical value for less than 5% of the replications. It would seem that an observed fit measure less than the critical value would support the submodel only weakly, but a fit measure greater than the critical value would provide evidence against the submodel.

For independent PC statistics, the submodel fit measures, the full model fit measures, and the fit difference measures had means and standard deviations less than the corresponding theoretical chi-square distribution. This suggests that at any alpha level, less than alpha of the observations would exceed the critical value. Consequently, an observed fit measure (or fit difference measure) which exceeds the theoretical critical value should lead to rejection of the model with true significance level alpha (or less).

Particularly when estimating the PC statistics in an independent fashion, sample proportions of zero and one occurred frequently. At first, we treated these as missing data.

However, when we made the following substitutions, the t-ratio and the fit measure more nearly followed the desired theoretical distributions. Let n_{jk} be the number of people who correctly answered exactly one item of pair (j,k). When the sample value p_{jk} equalled 1.00, we substituted $p_{jk} = (n_{jk} - .5)/n_{jk}$. When p_{jk} equalled 0.00, we substituted $p_{jk} = .5/n_{jk}$. The reported results for Study 2 are based on these substitutions.

In this paper, we focused on the hypothesis of equal item parameters across groups. The general logistic regression approach is applicable to a wide class of models for PC statistics. Particularly, let **D** be a design matrix with GI(I - 1)/2 rows and G(I - 1) columns, let **b** be a vector containing the G(I - 1) item difficulties from the full model, and let **b*** a row vector of item difficulties from a submodel. The logistic regression approach can be used to fit any submodel in which the item difficulties of the submodel have the form **b*** = **Db**. Indeed, Anderson and Davison (1991) discuss the fitting of an even wider class of models using an extension of logistic regression. Thus, logistic regression of PC statistics, or its extensions, are potentially useful for fitting a very broad class of hypotheses about item difficulties in the Rasch model.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 2, 449-460.

Anderson, R. & Davison, M. L. (1991). Dimensionality in the Rasch model. Paper presented to the American Educational Research Association, Chicago, IL, April 3.

Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika, 39, 324-345.

Chen, T.H. (in progress). Parameter invariance in the Rasch model. Unpublished Doctoral Dissertation, University of Minnesota, Minneapolis, MN.

Chen, T.H. & Davison, M.L. (1991). Parameter invariance in the Rasch model. Proceedings of the International Educational Statistics and Measurement Symposium: Recent Developments on Multivariate Analysis and Item Response Theory and their Applications. Tainan, Taiwan, April 20, 1991.

Choppin, B.H. (1968). An item bank using sample-free calibration. Nature, 219, 870-872.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston MA: Klewer-Nijhoff.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. New York: Erlbaum Associates.

Luce, R.D. (1959). Individual Choice Behavior. New York: Wiley.

McGuire, D., & Davison, M. L. (in press). Testing group differences in paired comparisons data. Psychological Bulletin.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

van den Wollenberg, A.L., Wierda, F.W., & Jansen, P.G.W. (1988). Consistency of Rasch model parameter estimation: A simulation study. Applied Psychological Measurement, 12, 307 - 318.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, van den Wollenberg, and Wierda. Applied Psychological Measurement, 12, 315 - 318.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Yoes, M. (1987). Response Data Simulation Program. Unpublished Computer Program in PASCAL, University of Minnesota, Minneapolis, MN 55455.onsidered signif

Table 1

Paired Comparisons Statistics for 10 Math Items, Fall and Spring Testing

Fall

Item	1	2	3	4	5	6	7	8	9
2	.039								
3	.364	.918							
4	.029	.469	.022						
5	.042	.515	.036	.537					
6	.119	.839	.244	.904	.865				
7	.184	.876	.341	.913	.876	.618			
8	.750	1.000	.864	1.000	.990	.955	.897		
9	.089	.784	.152	.877	.813	.397	.317	.034	
10	.052	.756	.115	.844	.803	.333	.273	.016	.415

Spring

Item	1	2	3	4	5	6	7	8	9
2	.065								
3	.174	.820							
4	.078	.585	.182						
5	.058	.554	.143	.468					
6	.273	.864	.586	.808	.857				
7	.087	.627	.205	.536	.564	.150			
8	.600	.966	.870	.958	.961	.875	.976		
9	.071	.733	.389	.655	.750	.294	.638	.067	
10	.118	.780	.353	.674	.721	.273	.633	.034	.484

Table 2

Dummy Coding for J = 4 Items in One Group

		Dummy Coding d_i		
i	k	d_2	d_3	d_4
2	1	-2	-1	-1
3	1	-1	-2	-1
4	1	-1	-1	-2
3	2	1	-1	0
4	2	1	0	-1
4	3	0	1	-1

Table 3

Dummy Coding (Full Model) for J = 4 Items in G = 3 Groups

			Dummy Variable d_{ih}								
g	i	k	d_{21}	d_{31}	d_{41}	d_{22}	d_{32}	d_{42}	d_{23}	d_{33}	d_{43}
1	2	1	-2	-1	-1	0	0	0	0	0	0
1	3	1	-1	-2	-1	0	0	0	0	0	0
1	4	1	-1	-1	-2	0	0	0	0	0	0
1	3	2	1	-1	0	0	0	0	0	0	0
1	4	2	1	0	-1	0	0	0	0	0	0
1	4	3	0	1	-1	0	0	0	0	0	0
2	2	1	-2	-1	-1	-2	-1	-1	0	0	0
2	3	1	-1	-2	-1	-1	-2	-1	0	0	0
2	4	1	-1	-1	-2	-1	-1	-2	0	0	0
2	3	2	1	-1	0	1	-1	0	0	0	0
2	4	2	1	0	-1	1	0	-1	0	0	0
2	4	3	0	1	-1	0	1	-1	0	0	0
3	2	1	-2	-1	-1	0	0	0	-2	-1	-1
3	3	1	-1	-2	-1	0	0	0	-1	-2	-1
3	4	1	-1	-1	-2	0	0	0	-1	-1	-2
3	3	2	1	-1	0	0	0	0	1	-1	0
3	4	2	1	0	-1	0	0	0	1	0	-1
3	4	3	0	1	-1	0	0	0	0	1	-1

Table 4
Selected Simulation Results
Model Fitting Results
(n=6 Items)

Type of PC Statistic	Model	Item Parameter Estimates	True Item Parameter						X^2 Goodness of Fit Stat.*		
			-1.70	-1.02	-.34	.34	1.02	1.70	Mean	sd	%sig (α=.05)
Non-Independent (N=100)	Submodel								17.51	7.62	2
		Mean ($\hat{b}$)	-1.71	-1.03	-.35	.33	1.04	1.73			
		s.d. ($\hat{b}$)	.24	.20	.16	.19	.19	.24			
		RMS $\hat{\sigma}(\hat{b})$	.22	.15	.11	.11	.15	---			
	Full Model								5.76	1.84	0
	Group 1	Mean ($\hat{b}$)	-1.76	-1.05	-.37	.36	1.07	1.76			
		s.d. ($\hat{b}$)	.36	.27	.23	.27	.26	.31			
		RMS $\hat{\sigma}(\hat{b})$	.27	.20	.16	.16	.19	---			
	Group 2	Mean ($\hat{b}$)	-1.72	-1.04	-.33	.31	1.04	1.74			
		s.d. ($\hat{b}$)	.30	.28	.23	.25	.25	.31			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---			
	Difference								**11.75**	**7.25**	**45**
Non-Independent (N=500)	Submodel								16.67	7.79	3
		Mean ($\hat{b}$)	-1.71	-1.02	-.33	.35	1.01	1.71			
		s.d. ($\hat{b}$)	.11	.08	.08	.07	.07	.10			
		RMS $\hat{\sigma}(\hat{b})$	.10	.07	.05	.05	.07	---			
	Full Model								5.46	2.14	0
	Group 1	Mean ($\hat{b}$)	-1.73	-1.03	-.31	.34	1.02	1.71			
		s.d. ($\hat{b}$)	.15	.11	.10	.10	.12	.14			
		RMS $\hat{\sigma}(\hat{b})$	.12	.09	.07	.07	.09	---			
	Group 2	Mean ($\hat{b}$)	-1.71	-1.02	-.34	.36	1.01	1.71			
		s.d. ($\hat{b}$)	.16	.10	.11	.10	.09	.14			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---			
	Difference								**11.21**	**7.14**	**42**
Independent (N=500)	Submodel								22.73	6.31	1
		Mean ($\hat{b}$)	-1.72	-1.01	-.33	.35	1.00	1.71			
		s.d. ($\hat{b}$)	.21	.13	.11	.10	.13	.19			
		RMS $\hat{\sigma}(\hat{b})$	.19	.13	.10	.10	.13	---			
	Full Model								18.28	6.02	5
	Group 1	Mean ($\hat{b}$)	-1.75	-1.01	-.33	.33	1.02	1.73			
		s.d. ($\hat{b}$)	.24	.16	.15	.15	.17	.22			
		RMS $\hat{\sigma}(\hat{b})$	.23	.17	.14	.14	.17	---			
	Group 2	Mean ($\hat{b}$)	-1.72	-1.01	-.34	.36	1.00	1.71			
		s.d. ($\hat{b}$)	.25	.16	.14	.12	.16	.23			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---			
	Difference								**4.45**	**2.44**	**2**

* For Submodels, $df(X^2) = 24.SQRT(2df) = 6.93$; For Full Models, $df(X^2) = 19.SQRT(2df) = 6.16$
For Differences between the submodels and the full models, $df(X^2) = 5.SQRT(2df) = 3.16$

Table 4 (cont.)
Model Fitting Results
(n=11 Items)

Type of PC Statistic	Model	Item Parameter Estimates	True Item Parameter										X^2 Goodness of Fit Stat*		
			-1.70	-1.36	-1.02	-.68	-.34	.00	.34	.68	1.02	1.36	1.70	Mean	s.d.	%sig (α=.05)
Non-Independent (N=100)	Submodel													65.15	22.08	2
		Mean ($\hat{b}$)	-1.69	-1.44	-1.01	-.66	-.34	.00	.34	.70	1.00	1.40	1.70			
		s.d. ($\hat{b}$)	.23	.22	.18	.13	.16	.16	.17	.18	.17	.20	.20			
		RMS $\hat{\sigma}(\hat{b})$	.10	.12	.09	.10	.09	.09	.09	.09	.10	.09	---			
	Full Model													21.76	4.04	0
	Group 1	Mean ($\hat{b}$)	-1.72	-1.43	-1.01	-.66	-.36	-.01	.33	.71	1.02	1.42	1.72			
		s.d. ($\hat{b}$)	.33	.28	.24	.21	.22	.21	.23	.27	.27	.30	.28			
		RMS $\hat{\sigma}(\hat{b})$	.14	.15	.12	.13	.12	.12	.12	.12	.14	.13	---			
	Group 2	Mean ($\hat{b}$)	-1.70	-1.46	-1.04	-.67	-.32	.02	.35	.69	1.01	1.40	1.73			
		s.d. ($\hat{b}$)	.32	.30	.28	.22	.24	.26	.24	.25	.21	.27	.31			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---	---	---	---	---	---			
	Difference													**43.39**	**21.19**	**95**
Non-Independent (N=500)	Submodel													62.77	18.45	0
		Mean ($\hat{b}$)	-1.70	-1.36	-1.02	-.68	-.35	.01	.33	.68	1.02	1.36	1.71			
		s.d. ($\hat{b}$)	.09	.09	.07	.07	.08	.08	.07	.08	.08	.08	.10			
		RMS $\hat{\sigma}(\hat{b})$	.04	.05	.04	.05	.04	.04	.04	.04	.05	.04	---			
	Full Model													20.03	3.63	0
	Group 1	Mean ($\hat{b}$)	-1.70	-1.37	-1.03	-.68	-.35	.01	.34	.69	1.02	1.36	1.71			
		s.d. ($\hat{b}$)	.12	.11	.11	.10	.11	.11	.10	.10	.11	.11	.14			
		RMS $\hat{\sigma}(\hat{b})$	.06	.07	.05	.06	.05	.05	.05	.05	.06	.06	---			
	Group 2	Mean ($\hat{b}$)	-1.70	-1.35	-1.01	-.69	-.35	.01	.33	.68	1.02	1.36	1.71			
		s.d. ($\hat{b}$)	.13	.12	.09	.11	.10	.11	.12	.12	.12	.12	.14			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---	---	---	---	---	---			
	Difference													**42.74**	**17.99**	**96**
Independent (N=500)	Submodel													94.53	11.04	0
		Mean ($\hat{b}$)	-1.69	-1.37	-1.02	-.68	-.35	.00	.33	.69	1.04	1.35	1.69			
		s.d. ($\hat{b}$)	.11	.14	.09	.12	.10	.10	.11	.11	.12	.11	.13			
		RMS $\hat{\sigma}(\hat{b})$	.12	.14	.10	.12	.10	.11	.11	.10	.13	.11	---			
	Full Model													85.19	10.72	0
	Group 1	Mean ($\hat{b}$)	-1.71	-1.39	-1.04	-.69	-.35	.02	.33	.70	1.05	1.37	1.71			
		s.d. ($\hat{b}$)	.17	.17	.14	.14	.14	.13	.16	.14	.14	.15	.18			
		RMS $\hat{\sigma}(\hat{b})$	.17	.18	.15	.16	.14	.14	.15	.14	.16	.16	---			
	Group 2	Mean ($\hat{b}$)	-1.69	-1.36	-1.00	-.68	-.35	-.01	.34	.68	1.04	1.35	1.69			
		s.d. ($\hat{b}$)	.16	.18	.13	.16	.14	.13	.15	.14	.17	.15	.19			
		RMS $\hat{\sigma}(\hat{b})$	---	---	---	---	---	---	---	---	---	---	---			
	Difference													**9.35**	**4.22**	**3**

* For Submodels, $df(X^2) = 99.SQRT(2df) = 14.07$; For Full Models, $df(X^2) = 89.SQRT(2df) = 13.34$;
For Differences between the submodels and the full models, $df(X^2) = 10.SQRT(2df) = 4.47$

Table 4 (cont.)
Full Model Results on Estimated Parameter Differences
(n=6 Items)

		True Item Parameter						
		-1.70	-1.02	- .34	.34	1.02	1.70	
Type of PC Statistic	Stat.	True b=0.0	0.0	0.0	0.0	0.0	0.0	Mean
NonIndependent	mean($\hat{b}^*$)	.04	.01	.04	-.05	-.03	-.01	0.0
(N=100)	s.d.($\hat{b}^*$)	.46	.37	.32	.35	.35	.38	---
	Rms $\hat{\sigma}(\hat{b}^*)$	.27	.24	.22	.22	.24	---	---
	mean(t)	.13	.05	.18	-.24	- .13	---	0.0
	s.d.(t)	1.65	1.58	1.45	1.58	1.47	---	---
	% sig t (α =.05)	21	21	15	18	17	---	18.4
NonIndependent	mean($\hat{b}^*$)	.02	.00	- .03	.01	- .01	.00	0.0
(N=500)	s.d.($\hat{b}^*$)	.20	.15	.14	.13	.16	.20	---
	RmS $\hat{\sigma}(\hat{b}^*)$	.12	.10	.10	.10	.10	---	---
	mean(t)	.21	.02	- .31	.14	- .11	---	0.0
	s.d.(t)	1.67	1.41	1.45	1.30	1.51	---	---
	% sig. t (α=.05)	25	13	17	13	21	---	17.8
Independent	mean($\hat{b}^*$)	.03	- .01	- .01	.03	- .01	-.02	0.0
(N=500)	s.d.($\hat{b}^*$)	.24	.18	.18	.17	.20	.24	---
	RmS $\hat{\sigma}(\hat{b}^*)$	.24	.20	.19	.19	.20	---	---
	mean(t)	.13	- .05	- .08	.16	- .06	---	0.0
	s.d.(t)	.98	.90	.95	.91	1.01	---	---
	% sig. t (α =.05)	2	2	5	3	2	---	2.8

Table 4 (cont.)
Full Model Results on Estimated Parameter Differences
(n=11 Items)

Type of PC Statistic	Stat.	True Item Parameter											
	True b =	-1.70 0	-1.36 0	-1.02 0	-.68 0	-.34 0	.00 0	.34 0	.68 0	1.02 0	1.36 0	1.70 0	Mean
Non-Independent (N=100)	Mean ($\hat{b}^*$)	.02	-.03	-.03	-.01	.04	.03	.02	-.02	-.01	-.01	.01	0.0
	s.d. ($\hat{b}^*$)	.45	.37	.38	.33	.31	.35	.32	.36	.35	.42	.43	
	RMS $\hat{\sigma}(\hat{b}^*)$	.20	.19	.17	.17	.16	.16	.16	.17	.17	.19	---	
	Mean (t)	.08	-.15	-.14	-.09	.24	.14	.15	-.12	-.05	-.05	---	0.0
	s.d. (t)	2.19	1.99	2.17	1.96	1.92	2.13	2.00	2.12	1.99	2.18	---	
	% sig t (α=.05)	37	29	36	28	30	31	34	34	31	34	---	32.4
Non-Independent (N=500)	Mean ($\hat{b}^*$)	.00	.02	.02	-.01	.00	.00	-.01	.00	-.01	-.01	-.01	0.0
	s.d. ($\hat{b}^*$)	.18	.16	.15	.15	.15	.14	.16	.15	.18	.15	.19	
	RMS $\hat{\sigma}(\hat{b}^*)$	.09	.08	.08	.07	.07	.07	.07	.07	.08	.08	---	
	Mean (t)	.00	.25	.20	-.06	-.01	.01	-.08	-.04	-.09	-.09	---	0.0
	s.d. (t)	2.08	1.94	1.91	2.09	2.03	2.00	2.21	2.07	2.29	1.81	---	
	% sig t (α=.05)	36	30	31	33	38	35	31	34	34	28	---	33.0
Independent (N=500)	Mean ($\hat{b}^*$)	.02	.03	.04	.01	-.01	-.02	.01	-.02	-.01	-.02	-.02	0.0
	s.d. ($\hat{b}^*$)	.24	.23	.20	.18	.19	.17	.21	.18	.21	.20	.25	
	RMS $\hat{\sigma}(\hat{b}^*)$	.24	.22	.21	.20	.19	.19	.19	.20	.21	.22	---	
	Mean (t)	.08	.14	.19	.03	-.03	-.12	.04	-.09	-.07	-.08	---	0.0
	s.d. (t)	1.00	1.02	.95	.91	.99	.90	1.10	.92	1.00	.90	--	
	% sig t (α=.05)	3	5	4	5	1	2	4	3	4	2	---	3.3

Table 5

Parameter Estimates for Ten Mathematics Items from Logistic Regression

	LR Submodel			LR Full Model			
	Item Parameter	S.E.	t	Regression Coefficient	S.E.	Item Parameter	t
Item 1	-1.594	--	--	--	--	-1.655	--
2	1.062	0.214	4.96**	1.037	0.256	1.037	4.06**
3	-0.879	0.228	-3.86**	-1.307	0.319	-1.307	-4.10**
4	1.353	0.199	6.79**	2.117	0.324	2.117	6.53**
5	1.160	0.178	6.53**	1.333	0.245	1.333	5.45**
6	-0.464	0.188	-2.47*	-0.280	0.248	-0.280	-1.13
7	0.148	0.182	0.81	-0.268	0.255	-0.268	-1.05
8	-2.048	0.310	-6.60**	-2.462	0.445	-2.462	-5.54**
9	0.386	0.217	1.78	0.422	0.262	0.422	1.61
10	0.876	0.245	3.58**	1.063	0.296	1.063	3.59**
Spring							
Item 1				--	--	-1.533	--
2				0.019	0.358	1.056	0.05
3				0.906	0.424	-0.401	2.14*
4				-1.513	0.410	0.604	- 3.69**
5				-0.374	0.364	0.959	- 1.03
6				-0.437	0.392	-0.717	- 1.11
7				0.926	0.360	0.658	2.57*
8				0.903	0.598	-1.559	1.51
9				-0.105	0.347	0.317	- 0.30
10				-0.447	0.382	0.616	-1.17

* p < .05
**p < .01