ED 412 219 TM 027 409

AUTHOR	Yan, Jean W.
TITLE	Examining Local Item Dependence Effects in a Large-Scale Science Assessment by a Rasch Partial Credit Model.
PUB DATE	1997-03-00
NOTE	27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; Cluster Analysis; Goodness of Fit; *Item Response Theory; Monte Carlo Methods; Sampling; *Science Tests; *Test Items
IDENTIFIERS	Item Dependence; Large Scale Assessment; *Partial Credit Model; *Rasch Model; Testlets

ABSTRACT

Context-dependent items are traditionally analyzed independently, creating a situation in which the potential local item dependence effects among these items may cause a biased estimation of examinees' abilities. This study investigated the local item dependence effects on testlets in the tryout version of a statewide science assessment by a Rasch partial credit model. Cluster sampling combined with stratified sampling was used. Data were analyzed in five different configurations to study the relationships between context-dependent items at the individual item level and at the testlet level. It is shown that local dependence effects may be controlled and a better fit for testlet calibration can be obtained by employing the Rasch partial credit model for some, but not all testlets. (Contains 2 figures, 11 tables, and 35 references.) (Author/SLD)

EXAMINING LOCAL ITEM DEPENDENCE EFFECTS IN A LARGE-SCALE SCIENCE ASSESSMENT BY A RASCH PARTIAL CREDIT MODEL

Jean W. Yan
Michigan Department of Education

Abstract: Context-dependent items are traditionally analyzed independently, in which the potential local item dependence effects among these items may cause a biased estimation of examinees' abilities. The purpose of this study was to investigate the local item dependence effects on testlets in the tryout version of a statewide science assessment by a Rasch partial credit model. Cluster sampling combined with stratified sampling was used in the study. Data were analyzed in five different configurations to study the relationships between context-dependent items at the individual item level and at the testlet level. It is found that local dependence effects may be controlled and a better fit for testlet calibration can be obtained by employing the Rasch partial credit model for some, but not all testlets.

Key words: testlet, partial credit model, local item dependence, item response theory

Paper presented at the American Educational Research Association Annual Conference in Chicago, March, 1997. To request for paper, please contact Jean Yan at the Michigan Educational Assessment Program, Michigan Department of Education, P. O. Box 30008, Lansing, MI 48909 or send an e:mail to jyan@ed.mde.state.mi.us.

Introduction

Item response theory (IRT) assume that multiple-choice test items are independent when examinees' abilities are controlled, each item is analyzed independently and usually dichotomously (i.e., right or wrong). Consequently, the unit of analysis is the item itself. In many testing situations, however, such as a short story in a reading comprehension test, a table in a mathematics test, or an investigation in a science test, a context is established and examinees are often asked a series of questions related to that context. Wainer and Kiely (1987) called a set of these context-dependent items a "testlet" and defined it as: "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow (p.190)."

The immediate problem with conventional scoring methods under this circumstance is that the assumption of local item independence in IRT may be violated, because some items may be more strongly correlated within a context than between contexts. These correlations, which are context-specific rather than test-specific, may result in faulty measurement of the common factor between contexts and further affect the precision of person ability estimation. The information curve may be misled by the excess item correlation within a testlet since context-dependent items themselves may be statistically dependent. For test developers, the possible excess item correlation may produce inaccurate item calibration and the test reliability, which would further affect the test developers in their decision-making process on the fitness of the items. For test users, the potentially biased estimates of examinees may result potentially biased decisions on examinees' latent abilities on the tested subject. The effect of misclassification of an examinee could be very serious especially when a test is high stake. As for the examinees, they maybe benefit or get harmed due to the context-specific correlation, depending on whether or not these items correlate positively or negatively. An alternative to solve this problem is to analyze these items together as a unit by a Rasch partial credit model, in which the item dependence effect is built into the model.

During the last decade, there has been growing interest in treating a set of context-dependent items as the unit of analysis in educational measurement research. One main reason that test developers are using larger tasks as the fundamental units of tests and further shifting their focus to this field is that, besides the testlet characteristics such as capacity of providing more complex contexts and testing students' higher-order thinking, modern tests carry increasingly heavier responsibilities than before. A test result may now be used not only for achievement assessment, diagnosis, placement, or admission purposes, but also as an important reference to policy-making and education budgeting practices. The same amount of testing time and information are used to achieve more goals nowadays. Furthermore, researchers have experimentally projected that testlets as units of analysis can solve some of the measurement problems that could not be overcome by item-based analysis (Ebel, 1951; Wainer & Kiely, 1987; Rosenbaum, 1988; Thissen et al, 1988, 1989; Haladyna, 1992; Yen, 1984a, 1993).

The purpose of this study is to explore the local item dependence effects when context-dependent items in a large-scale science assessment were analyzed as individual items by the conventional Rasch dichotomous model and as testlets by a Rasch partial credit model.

Literature Review

Considerable studies and discussions about testlets have been contributed to applications of testlet concepts (Szeberënyi & Tigyi, 1987; Wainer et al, 1990, 1991, 1992), construction and development of testlets (Engelhart, 1942; Gerberich, 1956; Gronlund, 1965; Biggs & Collis, 1982; Mehrens & Lehmann, 1984; Collis et al, 1986; Haladyna, 1991), and measurement precision (Cureton, 1965; Cattell & Burdsal, 1975; Wainer et al, 1990; Sireci et al, 1991; Ercikan, 1993). Research on the local item dependence effect, however, did not receive much attention.

Rosenbaum (1988) compared item response distributions when it was conditional between but not within item "bundles" (testlets) with two sets of IRT assumptions, one set was traditional IRT and the other was less restrictive on local independence, allowing dependence among pairs of items that shared the same context. He proved a theorem that, at every level of the ability, the standard error of measurement (SEM) under a positively correlated bundle was at least as large as that from a conventional IRT model having the same item characteristic curves. He also found that positive dependence within bundles increased the SEM along the ability continuum. He suggested that, other things being equal, it is preferable not to use bundles of positively dependent items since it may cause a larger SEM.

Thissen, Steinberg, and Mooney (1989) used a multivariate logistic latent trait model (Bock, 1972) to examine the violation of the local independence assumption with computerized adaptive test data. They compared the results of a 4-testlet, 22-item test when the items were analyzed first as independent items and then as testlets. The results showed that, when testlet items were analyzed independently, the test information obtained was deceptively high. When those items were analyzed as testlets, the concurrent validity was slightly but significantly higher than that of the independently analyzed items. They concluded that the outcome of more information was "fooled" by the excess within testlet correlation among items and that the testlet scores appeared to be as least as valid as the individual item scores.

Yen (1993) used an IRT 3-parameter model and an IRT 2-parameter partial credit model to study multiple-choice tests of the *Comprehensive Test of Basic Skills, Fourth Edition* (CTBS/4; CTB Macmillan/McGraw-Hill, 1989) and the performance assessment data of a state education assessment program. Item information and discrimination estimates obtained by testlet scale and by item scale on reading and math tests were compared. It was found that testlet analysis did result in a larger SEM, but it could be seen as a reflection of reality. However, in many cases, there was not much difference in parameter estimates when items were scaled as testlets or as independent items.

Studies on the effect of loss of local independence so far mostly used IRT 2-parameter or 3-parameter polytomous models (e.g., the studies above). A hidden problem of using these models is that because they do not have sufficient statistics to separate the person parameter from the item parameter, they are sample dependent and results can be varied from sample to sample (Wright, 1992). Wilson (1988) used the family of Rasch models to study the local item dependence effect with an example of "superitems" (testlets) in the *Structure of the Learning Outcome Program.* The results showed that the rating scale model calibration provided no evidence of the violation of the local item dependence assumption. Dependencies between items were adequately summarized by item difficulties of the dichotomous model. On the other hand, the partial credit model calibration showed that one of the five testlets studied demonstrated a local item dependence effect. However, the sample size was very small in Wilson's study (1988). the data was collected from only 30 students of 9th and 10th grades, which is not comparable with a large scale assessment program.

Masters's (1982) partial credit model was originally developed to analyze multiple-category items and it has remained this way for most studies of this model. For multiple-choice item analysis, it has been used for foil analysis to gain more information. Other uses include theoretical exploration such as multi-dimensionality issues (De Ayala, 1991) and necessary and sufficient conditions to equate the estimates from dichotomous and partial credit models (Huynh, 1994). However, most comparisons were on the item level instead of on the testlet level. For a couple of studies on the testlet level (e.g., Wilson and Iventosch, 1988), either the items were performance-based or the research was experimental with small samples. Nevertheless, studies so far have found that the partial credit model added more detailed information to the item calibration and provided an opportunity to observe the local dependence between items within a testlet when those situations occurred.

From the literature on this topic, there has not been a study to examine the local dependence due to a testlet effect in a large-scale state assessment program using Masters's partial credit model. In addition, no study has explored the curriculum impact on different scales of the item analysis. Sometimes it is very possible that the context of constructing a testlet makes perfect sense in curriculum, and it may affect the analysis of scoring scales psychometrically. Other times, it does not have any impact at all. The results of this study provided evidence of a real life example in applying an alternative item analysis method to a large scale, high-stake assessment program.

Methodology

The Testing Instrument

The newly-developed *Michigan High School Proficiency Test in Science* assesses Michigan's 11th graders in five scientific dimensions: Using Life Scientific Knowledge, Using Physical Scientific Knowledge, Using Earth Scientific Knowledge, Constructing Scientific Knowledge, and Reflecting Scientific Knowledge. A test was composed of 30 independent multiple-choice items, 4 testlets, each consisting of 4 context-dependent multiple-choice items and one constructed response question, one scientific investigation and one text criticism, each with two constructed response questions. A sample testlet is attached in Appendix A.

Data

The tryout data were used in this study. There were 10 tryout forms (Forms 20-29) in total and no items in common between forms. The forms were organized into four triplets and two quadruplets. The following table displays form compositions in the data collection design:

Table 1. Tryout Form Compositions.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Form 20	Form 22	Form 24	Form 26	Form 29	Form 23
Form 21	Form 23	Form 25	Form 27	Form 20	Form 25
Form 22	Form 24	Form 26	Form 28	Form 21	Form 27
			Form 29		Form 28

The forms within each group were spiraled (e.g., in group 1, forms were ordered repeatedly in Forms 20, 21, and 22 fashion) and were administered to students within a classroom. By doing so, no two forms were the same for the students sitting next to each other. Each tryout school received only one group of forms. Students taking different forms were considered to form randomly equivalent groups. In addition, each form was administered to two different groups of students. In other words, there were forms in common between groups. This design allowed the equating of forms by the assumption of randomly equivalent groups. (An alternative design of spiraling all forms within schools was not used due to security concerns.)

Cluster sampling in combination with stratified sampling was used in the tryout. A stratum was defined by the population size of the area where the school located. The number of schools participating in the science tryout was randomly sampled from each stratum roughly proportional to the population by the stratum school weight. When a school was chosen into the sample, all the 11th graders within that school were included. There were 10,074 students in total from 72 Michigan public high schools who eventually took the science tryout.

The Calibration Models

Both the traditional dichotomous rating scale and the partial credit scale in IRT Rasch models were used.

The dichotomous model (Rasch, 1960) is the simplest form in the family of Rasch models. It is used to estimate person and item parameters when items are scored dichotomously. For a dichotomously scored item *i*, the model specifies the probability of a correct response to the item as an exponential function of the difference between person ability β_n and item difficulty δ_{ij}:

$$\phi_{nij} = \frac{\pi_{ni(j=1)}}{\pi_{ni(j=0)}+\pi_{ni(j=1)}} = \frac{\exp(\beta_n-\delta_{ij})}{1+\exp(\beta_n-\delta_{ij})}, \text{where} \tag{1}$$

ϕ_{nij} is the person *n*'s probability of scoring 1 rather than 0 on item *i*,

β_n is the ability of person *n*, n=1, 2, ..., N,

δ_{ij} is the difficulty of item *i*, i=1,2, ..., L,

$\pi_{ni(j=1)} = \phi_{ni(j=1)}$ is the person *n*'s probability of scoring 1 on item *i*, and

$\pi_{ni(j=0)} = 1 - \phi_{ni(j=1)}$ is person *n*'s probability of answering item *i* incorrectly.

j=0 or 1 is the score of item *i*.

Parameters to be estimated in this model are person ability (β_n) and item difficulty (δ_{ij}).

Number of parameters=N+L-1, in which N is the number of students and L is the number of items. For example, for a 16-item test, the total number of parameters = N+16-1=N+15.

According to Masters (1982), the model can separate the person parameter, β_n, from the estimation equation for the items so as to make it possible to estimate item parameters sample free in the calibration. Consequently, the item and person parameters can be estimated on the basis of the existence of sufficient statistics. That is, the model establishes the parameter separability by conditioning the person parameters out of the calibration procedures entirely. Specifically, a test score of an examinee contains all the information for estimating a student's ability, and the item difficulties can be estimated from a simple count of persons completing each level or "step" (if partial credit model) of an item. The model is used in this study whenever items are analyzed independently.

The partial credit model (Masters, 1982) is an extension of the dichotomous model in that it provides a direct expression of the probability of an examinee with ability β_n responding at a particular performance level (e.g.,.*j*=1, 2, ..., m). For items with more than two performance levels, additional probability expressions are needed to describe the probability of getting score 2 rather than 1, 3 rather than 2, and so on, in terms of item step difficulty parameters $\delta_{i2}, \delta_{i3}, \ldots, \delta_{im}$. With the Rasch partial credit model, the probability of person *n* scoring *x* or completing any number of steps on item *i* is,

$$\pi_{nix} = \frac{\exp\sum_{j=0}^{x}(\beta_n-\delta_{ij})}{\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\beta_n-\delta_{ij})}, \; x=0,\ 1,\ 2,\ \ldots,\ m_i. \tag{2}$$

In eq. (2),

β_n is the person latent ability,

δ_{ij} is the difficulty parameter for the *j*th step in item *i*,

x=0, 1, 2, ..., m_i is the score that person *n* achieves in item *i*,

i=1, 2, ..., L is the item,

j=1, 2, ..., m is the step in item *i*, and n=1, 2, ..., N is the person.

The observation *x* in eq. (2) is the count of the completed steps for item *i*. The numerator contains only the difficulties of these *x* completed steps $\delta_{i1}, \delta_{i2}, \ldots, \delta_{ix}$. The denominator is the sum of all m_i+1 possible numerators (Wright and Masters, 1982). In other words, the formula is the ratio of *x*-step difficulties over the total possible *m*-step difficulties.

Parameters to be estimated in this model are person ability (β_n), step difficulty (δ_{ij}), and testlet difficulty, which is the average of all possible step measures for that testlet.

The number of parameters equals N+M+L-1, in which N is the person parameter, M= $\sum_{i=1}^{L} m_i$, the total number of steps in all the testlets, and L is the number of testlets.

For a 4-testlet test with each testlet having 4 items (i.e., steps), the number of parameters equals N+4x4+4-1 = N+16+4-1=N+19.

Although the partial credit model requires that the steps within an item be completed in sequence, the steps need not be equally difficult nor be ordered by step difficulties. If an item has only two performance levels (i.e., 0, 1), then the partial credit model reduces to the dichotomous model.

In the present study, the items within a testlet become "steps" and each "step" is scored 0 or 1. A testlet replaces the position of an item. The order of the items is the number of steps to be completed by an examinee.

Data Analysis

Data from all ten tryout forms were analyzed in five different configurations: (1) Context-dependent items that formed the testlets were scored as individual items; (2) An *original testlet* (i.e., the testlet by design in the test development) was scored as a whole unit; (3) Additional sixteen independent multiple-choice items in the same tryout form with the original testlets were randomly selected and scored as they were; (4) The same 16 independent items were randomly formed into 4 *random testlets* and were scored as testlets; and (5) The original testlets were reconfigurated into 4 new *reformed testlets*, with one item from each of the original testlets. The purpose of analyses (4) and (5) was to conduct a sort of concurrent validity analysis for the testlet effects. The constructed-response questions within a testlet were excluded from the analysis in this study to avoid errors from other factors such as interrater reliability in hand-scoring. All items were scored dichotomously. A testlet score was the sum of the context-dependent item scores within that testlet.

Person Ability Measure

In the unconditional maximum likelihood estimation procedure, the likelihood of the data matrix $((\mathbf{x}_{ni}))$ is the continued product of the unconditional probabilities, π_{nix}, over *n* and *i*,

$$\Lambda = \prod_{n}^{N} \prod_{i}^{L} \pi_{nix} = \frac{\exp \sum_{n}^{N} \sum_{i}^{L} \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\prod_{n}^{N} \prod_{i}^{L} \left[\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij}) \right]} . \tag{3}$$

In eq. (3),

π_{nix} is the probability of person *n* answering *x* steps in item *i* correctly,

x_{ni} is the observed score for person *n* on item *i*,

β_n is the person latent ability,

δ_{ij} is the step difficulty for item *i*, and

i=1, 2, ..., L is the number of items,

j=1, 2, ,.., m_i is the item step, and

n=1, 2, ..., N is the person.

The logarithm of eq. (3) is

$$\lambda = \log \Lambda = \sum_{n}^{N} \sum_{i}^{L} x_{ni} \beta_n - \sum_{n}^{N} \sum_{i}^{L} \sum_{j=1}^{x_{ni}} \delta_{ij} - \sum_{n}^{N} \sum_{i}^{L} \log\left[\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij}) \right], \tag{4}$$

in which $\sum_{j=0}^{x_{ni}} \delta_{ij} = \sum_{j=1}^{x_{ni}} \delta_{ij}$ because $\delta_{i0} \equiv 0$. Taking the first derivative of eq. (4) with respect of β_n, one gets

$$\frac{\partial\lambda}{\partial\beta_n} = r_n - \sum_i^L \sum_{k=1}^{mi} k\pi_{nik}, \qquad i=1, L \tag{5}$$

where $r_n = \sum_i^L x_{ni}$ is the test score for person n,

π_{nik} is the probability of person n completing k steps in testlet i,

k=1, 2, ..., m_i is the number of steps (i.e., items here) in testlet i,

$\sum_{k=1}^{mi} k\pi_{nik}$ is the number of steps person n is expected to complete in testlet i, and

$\sum_i^L \sum_{k=1}^{mi} k\pi_{nik}$ is the number of steps person n is expected to complete on the L-testlet test, or the expected score of r_n, the test score for person n. Symbolically,

$$E(r_n)= \sum_i^L \sum_{k=1}^{mi} k\pi_{nik}. \tag{6}$$

Setting eq. (5) to 0, and solving for β_n, we will get an estimate of person ability, b_r.

The standard error of the estimate can be calculated by

$$SE(b_r)=[\sum_i^L (\sum_{k=1}^{mi} k^2 P_{rik} - (\sum_{k=1}^{mi} k P_{rik})^2)]^{-1/2}, \tag{7}$$

where P_{rik} is the estimated probability of a person with a score of r responding in step k to testlet i of the last iteration.

Testlet Measure

Taking the first derivative of eq. (4) above with respect to δ_{ij}, one gets

$$\frac{\partial\lambda}{\partial\delta_{ij}} = -S_{ij} + \sum_n^N \sum_{k=j}^{mi} \pi_{nik}, \text{ n=1, N } ; \text{ j=1, ..., k, ..., } m_i, \tag{8}$$

where $S_{ij}= \sum_{n=1}^N \sum_{j=1}^{x_{ni}} \delta_{ij}$ is the number of persons completing step j in testlet i. $\sum_{k=j}^{mi} \pi_{nik}$ is the probability of person n completing at least j steps in testlet i, and

$\sum_n^N \sum_{k=j}^{mi} \pi_{nik}$ is the number of persons expected to complete at least j steps in testlet i. In other words, it is the expected value of S_{ij}. Symbolically, the expected value for step difficulty (d_{ij}) in testlet i is

$$E(d_{ij})= \sum_n^N \sum_{k=j}^{mi} \pi_{nik}. \tag{9}$$

Setting eq. (8) to 0, and solving for δ_{ij}, we will get the estimate of testlet step parameter, d_{ij}.

The standard error of d_{ij} is

$$SE(d_{ij})=[\sum_r^{M-1} N_r(\sum_{k=j}^{mi} P_{rik} - (\sum_{k=j}^{mi} P_{rik})^2)]^{-1/2} \tag{10}$$

where N_r is the number of persons with score r, $M= \sum_{i=1}^L m_i$.

For the simplicity of this study, the testlets do not take response patterns into consideration, and students' raw scores on the items within a testlet are summed up to a single number-right score.

Local Item Dependence Measure

To assess the local dependence effect, dichotomously scored items were first calibrated with the Rasch dichotomous model as individual items and then by the partial credit model as testlets. The difficulties obtained from both calibrations were compared for their calibration errors and item/testlet fits. The item fit statistics were calculated as follows (Wright & Masters, 1982):

observed response: x_{ni},

$$\text{expected value of } x_{ni}\text{: } E_{ni} = \sum_{k=0}^{mi} k\pi_{nik}, \tag{11}$$

$$\text{where } \pi_{nik} = \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij}) / \Psi_{ni}, \tag{12}$$

and $\Psi_{ni} = \sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})$, (13)

variance of x_{ni}: $W_{ni} = \sum_{k=0}^{mi} (k - E_{ni})^2 \pi_{nik}$, (14)

kurtosis of x_{ni}: $C_{ni} = \sum_{k=0}^{mi} (k - E_{ni})^4 \pi_{nik}$, (15)

score residual: $y_{ni} = x_{ni} - E_{ni}$, (16)

standardized residual: $z_{ni} = y_{ni} / W_{ni}^{1/2}$, (17)

standardized residual squared: z_{ni}^2, (18)

score residual squared: $y_{ni}^2 = W_{ni} z_{ni}^2$, (19)

unweighted mean square: $u_i = \sum_{n=1}^{N} z_{ni}^2 / N$, the outfit statistics, where N is the number of persons in the sample,

(20)

weighted mean square: $v_i = \sum_{n}^{N} W_{ni} z_{ni}^2 / \sum_{n}^{N} W_{ni} = \sum_{n}^{N} y_{ni}^2 / \sum_{n}^{N} W_{ni}$, is the infit statistic, (21)

and finally,

standardized weighted mean square: $t_i = (v_i^{1/3} - 1)(3 / q_i) + (q_i / 3)$, (22)

where v_i is the weighted mean square, q_i is the standard deviation of the weighted mean square, and t_i is the standardized weighted mean square for testlet *i*. In the formula, q_i is

$$q_i = [\sum_{n}^{N} (C_{ni} - W_{ni}^2) / (\sum_{n}^{N} W_{ni})]^{1/2}. \quad (23)$$

Similarly, the person fit statistic can be obtained in this manner also. In the formula, the person fit statistic is

$$t_n = (v_n^{1/3} - 1)(3 / q_n) + (q_n / 3), \quad (24)$$

where v_n is weighted mean square, q_n is the standard deviation of the weighted mean square, and t_n is the standardized weighted mean square for person *n*.

The information-weighted fit statistic (v_i) obtained from the computer program BIGSTEPS (Linacre, 1995, ver. 2.6) has an expected value of 1. Values substantially less than 1 indicate dependence in the data; values substantially greater than 1 indicate noise.

Results and Discussions

Phi coefficients were calculated for all the original, the random, and the reformed testlets for all tryout forms. The mean coefficients for each testlet for overall forms are listed in Table 2 (see Appendix B).

Insert Table 2 Here

As is shown in Table 2 out of the 40 original testlets, only one testlet (Testlet 3, Form 21) had an average ϕ coefficient above .30, which is relatively high for item correlation. Five testlets had mean coefficients between .20 and .30, more than half of the testlets (23) obtained moderate mean coefficients between .10 and .20, and the remaining 11 testlets had mean coefficients less than .10. For random testlets, twenty-three of them had mean ϕ coefficients less than .10, seventeen had mean coefficients between .10 and .20, but no testlets had mean coefficients greater than .20. For the reformed testlets, only Testlets 4 in Forms 24 and 27 had mean ϕ coefficients above .20 (ϕ=.2429 and .2496 respectively). Half of them (20) were between .10 and .20, and the remaining eighteen were under .10. The summary is in Table 3 below.

Marginal mean coefficients for all forms by testlet (column means in Table 2) and for all testlets by form (row means in Table 2) were calculated also. For each marginal value, mean coefficients for the original testlets were higher than either random testlets or reformed testlets, except Form 24, where the reformed testlet mean was slightly, but not significantly, higher than the original testlet mean. Between the reformed and random testlet means, coefficient values varied irregularly. In some cases, random testlets had higher mean coefficients. Other times, vise versa. This outcome is not surprising, however, because the contents of the reformed testlets are not related to the same context any more, and they are almost equivalent to the random testlets in the sense of testlet construction. Overall, the results strongly suggest that context-dependent items did have higher correlations within-context than across-context or independent items did, which implied that local dependence may exist in some of the original testlets.

Table 3. Summary of Mean Item Correlations for the Testlets

ϕ Coef.	Original Testlets	Random Testlets	Reformed Testlets
> .30	1	0	0
.21 - .30	5	0	2
.11 - .20	23	17	20
.00 - .10	11	23	18
Total	40	40	40

Testlet Measure Results

One rationale for using testlets as unit of analysis is to determine whether the calibration errors are smaller when treating the context-dependent items in a testlet as a whole than trearing them individually (i.e., ignoring the context effect), as well as determining whether such scaling produces better fits of testlet and/or person estimates.

The *User's Guide to BIGSTEPS* (Linacre & Wright, 1995) states that "INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's ability." And "MNSQ is the mean-square infit statistic with expectation 1. Values substantially below 1 indicate dependence in your data; values substantially above 1 indicate noise" (p. 82).

In the same manual, it is explained that, when values of infit mean square (MNSQ) statistic are, say, less than .8 or the standardized MNSQ is less than -2 SDs, it means there are redundant items and the test developers need to investigate the items to see if the test has similar items, one item answers another, or an item correlates with other variables, that is, local dependence effects. When the infit MNSQ is larger than, say, 1.2, or its standardized MNSQ is greater than +2 SDs, it may mean different things, such as biased items, qualitatively different items, or curriculum interaction. In these cases, one needs to investigate areas related to the problems (Linacre & Wright, 1995, p. 95).

By eq. (21) the infit MNSQ is the sum of squares of the difference between the observed score and the expected score divided by the sum of variances on item i over N persons. With the Rasch partial credit model, the smaller the discrepancy between the observed score and expected score, the larger the variance of x_{ni}. In the infit MNSQ formula, this means smaller residuals ($y_{ni} = x_{ni} - E_{ni}$). In other words, the formula will have a smaller numerator and a bigger denominator. As a result, v_i will be less than 1 when the numerator is smaller than the denominator.

Usually we expect an orderly pattern of responses. In other words, we want to see that the observed value is close to the expected value. However, when responses to an item are excessively orderly, that is, the observed scores are almost identical or identical to the expected scores, we may begin to suspect potential local dependence effects (Wright & Masters, 1982, p. 104). This would happen when problems like those mentioned earlier occur. An example of possible dependence is presented later in this section.

Table 4 (see Appendix B) displays the results of Forms 22, 23, and 27 as examples of testlet fit statistics for the original testlets and item fit statistics for the context-dependent items that configure these testlets.

Insert Table 4 Here

For all forms, seventeen out of 40 original testlets have 1 to 4 misfit items within a context when they were analyzed individually, but when they were analyzed as testlets, good testlet fit statistics were obtained. Considering Original Testlet 3 in Form 22 and Original Testlet 4 in Form 27 for example, when the items in those testlets were analyzed as individual items, all of the context-dependent items had misfit values beyond ±2 SDs (all 4 items have the "*" sign in col. 6). However, the items produced a proper testlet fit when they were analyzed as testlets (infit=1.03 for Testlet 3 in Form 22 and infit=.95 for Testlet 4 in Form 27). In addition, the standard errors of the estimates for the original testlets were uniformly .04, while the standard errors for the context-dependent items were larger, between .07 and .09 logit. These results meant that, for those context-dependent items, the testlet-based analyses were more appropriate statistically than the item-based analyses to examine students' abilities in the areas of interest.

Additional 20 original testlets, each also had 1 to 4 misfit context-dependent items when they were analyzed individually. However, when they were analyzed as testlets, the testlet calibrations were misfit (indicated by "*" sign in the table). Thirteen of these testlets had infit values substantially less than 1 (i.e., infit MNSQ < -2 SDs), implying that there may be local dependence effects in both the items of those testlets or the testlets themselves. This finding was a little surprising because these testlets were supposed to be independent to each other by design or by model control. It seemed that there may be some factors other than local dependence affecting the item and testlet calibration. The remaining 7 of these 20 testlets had infit values substantially greater than 1 (i.e., infit MNSQ > +2 SDs). For instance, Original Testlet 3 in Form 23 had misfit values for all its context-dependent items and the resulting infit MNSQ (1.22) for the testlet showed noise in the data this time. This implied students may have unexpected performance away from their expected scores. This outcome also suggested that test developers need to look at the testlet construction, content or quality of the items.

By the definition of fit statistics, Testlet 3 in Form 23 demonstrated one extreme (i.e., v_i greater than 1). The testlet was an earth science problem which required students to know the relationships between the ocean, coastal plateau, and mountain range. It was a relatively difficult testlet (difficulty measure=.98 logit). If a student were not clear about their relationships, the person would have a small probability of answering an item correctly. The items themselves were well written, with no signs of bias or trick, but for the two more difficult items (item #46's b=1.39 and item #48's b=1.19 logits), the percentages of students choosing a wrong option were larger than the percents of students choosing the right one (see Table 5 for detailed percentages). For item #46, the correct answer was option A. The percentage of students choosing A was 28% only, compared with 35% who chose the wrong option, D. The situation was similar for item #48. The percentage of students choosing the right answer, C, was 31%, while the percent choosing the wrong answer, D, was 35%. In addition, the average correlation among all 4 items was very small (r=.0656).

The results of large infit MNSQs (values substantially above 1.0) indicated large discrepancies between the observed scores and expected scores, implying students did not perform at their ability levels. These large discrepancies are considered "noise" in the item analysis. Usually one would suspect the item quality in this kind of situation. In this case, however, one may have to examine if there is an interaction of science dimensions within the testlet to seek possible reasons for poor performance. Nevertheless, "noise" in the item analysis does not have any relationship to local dependence. It is presented here to demonstrate another side of the infit statistic (i.e., values greater than 1.0). It also shows that large discrepancies between observed scores and expected scores do happen even though items are from the same context.

Table 5. Students Responses to Testlet 3, Form 23.

Item #	Option A	Option B	Option C	Option D
45	9.6%	13.6%	55.0%√	18.6%
46	27.9%√	22.5%	11.3%	35.0%
47	10.4%	15.5%	34.2%	36.6%√
48	9.8%	20.8%	30.8%√	35.4%

√ indicates the correct answer.

Testlet 4 in Form 23 provides an example of possible dependence. The testlet presented a diagram of the movement of carbon in the atmosphere and on the surface of Earth, and asked students to answer 4 questions based on the diagram. It was a relatively easy testlet (difficulty measure=-.82 logit) and most students chose the right answers of the items (see Table 6 for detail percentages). Looking at the item statistics, it seemed that distractors for three of the four items were not very effective because they attracted few students. By examining the item contents closely, we can see that if a student can answer item #52 (a concept item) correctly, he or she can answer the items #50, #51 and #53 fairly easily. Consequently, the observed and expected score differences would be very small.

As described in this section, small residuals imply possible local dependence. The average item correlation of this testlet (r=.2750) helped support the suspicion. This correlation was very high in this test, compared with the grand average correlation (r=.1429). When a situation like this is true, the infit statistic, v_i, will be very small (because the residual, y_{ni}, will be very small). For this testlet in particular, the infit MNSQ was .76, which indicated that possible local dependence may exist among the items and possibly with other items in the test also.

Table 6. Students Responses to Testlet 4, Form 23.

Item #	Option A	Option B	Option C	Option D
50	6.5%	79.5%√	7.0%	3.1%
51	11.1%	11.0%	28.0%	47.3%√
52	10.8%	66.0%√	8.6%	11.1%
53	7.3%	79.9%√	4.5%	4.7%

Across the forms, there were only 2 original testlets (Testlet 1 in Form 21 and Testlet 2 in Form 23) where the fit statistics were within the normal range regardless of which scoring model was used. Therefore, it would not matter if items in these testlets were analyzed independently or as testlets.

The strangest case was Testlet 3 in Form 26. All its 4 items were perfectly fit when analyzed individually, but the testlet fit was not acceptable (infit MNSQ=.88, less than -2 SDs). The reason of this outcome was unknown to the author. The only inference that can be made was that these items many be truly independent and should be analyzed independently, even though they were from the same context. More research is needed for this outcome. A summary of fit/misfit original testlets is in Table 7 below.

Table 7. Summary of Fit/Misfit Statistics for Context-Dependent Items

Number of Orig. Testlet (n=40)	Analyzed as Indiv. Items	Analyzed as Testlets
17	1~4 misfit items in each testlet	proper fit for all testlets
20	1~4 misfit items in each testlet	misfit for all testlets
2	proper fit for all items	proper fit for all testlets
1	proper fit for all items	misfit for the testlet

An analysis was also run for the random testlets and the independent items that form the random testlets. The results were similar to those of the original testlets. Out of 40 random testlets, 15 of them had from 1 to 4 misfit items when these items were analyzed as individual items, but they obtained very proper fit statistics when they were analyzed as testlets. Another 54 items that were distributed in 23 random testlets obtained misfit results no matter which model was used. Out of these 23 misfit testlets, 16 showed local dependence and 7 indicated noise in their data. Two random testlets (Testlet 4 in Form 21 and Testlet 3 in Form 29) obtained misfits when they were analyzed as testlets but had a very good fit for each item when they were analyzed as independent items. In addition, there was no random testlet that showed a proper fit for both scoring models, which ideally should be the case for these developer-designed independent items.

The outcome of misfit items converting into proper fit testlets that are related to no specific contexts is interesting, at the same time a little bit disturbing too. Theoretically, the developer-designed independent items should behave as statistically independent. However, the results of these 15 misfit-items-to-fit-testlets here showed that they were actually better off when they were analyzed as testlets. One needs to see if there is local dependence effects in these items or the results are just from random errors. The results for the random testlet analyses indicated that these labeled "independent" items may not be really statistically independent, even though they were designed to be so. Some items may be related to each other or to a common factor statistically, and more study is needed.

One difference between the random testlets and the original testlets in fit statistic analyses was that the range of the independent item standard errors (.07-.14) was larger than those of the context-dependent items in the original testlets (.07-.09). This suggested that student performance varied more for these independent items than for those context-dependent items, which further suggested that the context may have impact on student ability estimation and testlet calibration. A summary of fit/misfit statistics is shown in Table 8.

Table 8. Summary of Fit/Misfit Statistics for Independent Items

Number of Random Testlet (n=40)	Analyzed as Indiv. Items	Analyzed as Testlets
15	1~4 misfit items in each testlet	proper fit for all testlets
23	1~4 misfit items in each testlet	misfit for all testlets
2	proper fit for all items	proper fit for all testlets

It may be summarized that for the context-dependent items, mixed results have been obtained. More than 40% (17) of the original testlets demonstrated a better fit when they were analyzed as testlets. Half (20) of the original testlets had misfits by both models. Thirteen of these 20 testlets indicated possible local dependence, which suggests that further investigation of individual items in these testlets is needed regarding their contents, item construction, or item quality. Only 5% (2) of them obtained good fit as individual items and as testlets. For the independent items, the testlet fit statistics were not the same as the items fit statistics. Sixty items in 15 random testlets had obtained a better fit when they were analyzed as (hypothetical) testlets. Another 34% of the items (54) show misfit with these items being analyzed as testlets and as items. The results were contradictory to the test development in that these items did not contain local independence with them. It was suspected that there may be an implicit factor affecting item calibration.

Verification of Local Dependence Effects

One way to verify whether the context-dependent items demonstrate dependence to each other when they are analyzed individually is to first check the variance homoscedasticity of the item fit statistics and then conduct a one-way ANOVA to compare the means of the fit statistics regressed on testlets.

The fit statistic discussed in the last section is a weighted mean square with degrees of freedom by the number of students responding to an item minus 1. In this study, the degrees of freedom were relatively large for all forms since the test was large-scale. Consequently, the null hypothesis of local item independence within an original testlet would be easily rejected even though the dependence effect was very small. An alternative was to conduct a one-way ANOVA to verify whether the item fit statistics obtained by the Rasch partial credit model truly indicate local dependence between context-dependent items within a testlet.

In this ANOVA, the natural log of the infit statistic was the outcome variable and the testlet is the classification variable. If the confidence interval (CI) of its estimate includes 0 (because the expected value of infit is 1, so ln(E(infit) should be 0), it can be inferred that there is not enough evidence to show that items within a testlet are dependent.

Under normality and random sampling assumptions, the test statistic for a population variance equal to a predetermined value is

$$\frac{vs^2}{\sigma^2} \sim \chi^2_v, \qquad (25)$$

where v, equal to n-1, is the degree of freedom of the chi-square distribution, n is the number of examinees responding to the item, and s^2 is some mean square, equal to $\frac{ss}{v}$, ss is sum of squares. (In this study, s^2 is the weighted mean square of a context-dependent item.) Thus,

$$E(s^2) = \sigma^2, \text{ and} \qquad (26)$$

$$\text{var}(s^2) = \frac{2\sigma^4}{v}. \qquad (27)$$

Further, if we take the natural log of s^2, we get

$E[\ln(s^2)] \approx \ln(\sigma^2)$, and (28)

$$\text{var}[\ln(s^2)] \approx \frac{2}{v}. \qquad (29)$$

Consequently, because the term σ^2 was "logged out," if the degrees of freedom (*df*) for all the context-dependent items are the same, then the comparison between the infit statistics will not be biased. Otherwise, some adjustment may be needed. Table 9 lists discrepancies of *df*'s for all original testlets.

Table 9. Discrepancies For Testlets in the Tryout Forms

Form	Testlet 1	Testlet 2	Testlet 3	Testlet 4
20	4	3	4	5
21	1	2	3	9
22	3	2	2	6
23	5	2	2	4
24	7	3	3	3
25	4	3	6	6
26	3	3	4	7
27	8	1	1	7
28	2	2	2	14
29	1	3	1	4

Values in Table 9 show that the majority of discrepancies between *df*'s from the highest to the lowest within a testlet are between 1 to 4 out of about 1,000 students. Two testlets (Testlet 4 in Forms 21 and 28) have somewhat larger differences in *df*'s, 9 for Form 21 and 14 for Form 28, respectively.

We may assume that the small differences in *df* within a testlet are negligible because the infit statistic is a weighted mean square (i.e., variance is considered) and the sample size is large (1000 or so). A one-way ANOVA was conducted then for each form. The results were shown in Table 10 (see Appendix B). The graph of confidence intervals (CI) was displayed in Figure 1 (see Appendix B).

Insert Table 10, Figures 1-2 here

As stated earlier, the expected value for the infit statistic is 1 and its natural log is 0. It can be seen from Figure 1 that 35 out of 40 testlet statistics have included 0 in their CIs across the forms. Two testlets (Testlet 3 in Form 23 and Testlet 4 in Form 25) had values above 0 (indicating noise) and three testlets (Testlet 4 in Form 23, Testlet 3 in Form 27, and Testlet 1 in Form 28) had values below 0 point (indicating local dependence). The omnibus F statistics in Table 10 helped support the evidence. For all ten forms, 7 of them had nonsignificant F tests, indicating all testlets may include 0 and their infit statistics were within the normal range. Forms 21, 23, and 28 have significant F tests, implying that some of their testlets may have misfit statistics. The large SDs for some testlets in the table also showed that these testlets would have a wide confident interval. Figure 1 explains the outcome graphically.

Figure 2 shows the point estimates of ln(infit MNSQ) for all testlets. The majority (31) of estimates fall between -.05 and +.05, very close to 0, which provided the evidence to support that the testlet-based analysis produced appropriate fit statistics for the majority (30) of the original testlets in this test when a CI was built for each testlet.

Mean Person Ability Measures Results

It is acknowledged that the main purpose of any data analysis method in education is to try to measure person abilities as precisely as possible. Table 11 in Appendix B presents results for mean person ability measures for different data configurations. In the table, the first column is the data configuration. The second column is the mean of the estimated person measures for the examinees in different data configurations in each tryout form. The estimates are in logits. For most forms, the original testlets had slightly lower mean person measures than the context-dependent items did, except Form 26. In addition, their values varied between -.50 and .50 logit values, right around the middle point of 0 on the ability continuum. Only the independent-item data configuration for Forms 24 and 26 and the random testlets in Forms 24, 26 and 27 had mean measures greater than .50 logit value. Most of the time, these measures did not differ much for most forms no matter how the context-dependent items were analyzed, individually or as testlets.

Column 3 is infit mean-square (MNSQ) for the mean person measure. It is the average of the infit mean-squares associated with responses of the sample and it has an expected value of 1.0. Values in Column 3 show that regardless of types of data configuration, no infit MNSQ statistic had a value substantially below 1.0. The lowest value is .92, and the highest is 1.0, which indicated that in average there was not enough evidence to prove unexpected behavior affecting responses to items or testlets near students ability levels.

Outfit in Column 4 is an outlier-sensitive fit statistic. Its MNSQ is the mean-square outfit statistic with an expectation of 1.0. As with the infit statistic, values substantially less than 1.0 indicate dependency, while values substantially greater than 1.0 indicate the presence of unexpected outliers. In this sample, the outfit MNSQ statistics ranged from .94 to 1.10, which indicated that the data fitted the model relatively well.

Insert Table 11 Here

One phrase to be explained here is "data fit the model." Usually in statistical analyses, researchers test whether a model fits data because the model is designed to imitate data, so it has to be faithful to the data as much as possible. Otherwise, another model is used.

The Rasch model used here, however, is not designed to fit any data. Instead it is developed to define measurement. As Wright (1992) pointed out: "The Rasch model is a statement, a specification of the requirements of measurement -- the kind of statement that appears in Edward Thorndike's work, in Thurstone's work, in Guttman's work (p. 197)." Therefore, ".... The Rasch model is theory centered: data must fit, else get better data (p. 200)." As a result, the phrase "data fit the model" is used in this study.

In summary, for the context-dependent items, there was no significant difference in person fit statistics when the items were analyzed individually or as testlets. For the independent items, the person fit statistics stayed the same regardless of which model was used. For the reformed testlets, even though the testlets were not context-specific, they nevertheless still produced proper person fits as those of the original testlets did.

Conclusions

Based on the results of this study, the following conclusions are made:

1. Context-dependent items correlated more closely within-context than across-context for most original testlets, which provided primitive evidence that local item dependence may exist within a context.
2. Where there was a local item dependence effect emerged in the context-dependent items, the IRT assumption of local independence may be violated for some context-dependent items. Under this circumstance, it would be theoretically preferable to use the Rasch partial credit model. Evidence in this study showed that such a local dependence effect may be controlled and a better fit for testlet calibration can be obtained for some, but not all, original testlets by employing the model.
3. Caution must be exercised in any revision of the misfit testlets. Often only one or two misfit items causes misfit of the whole testlet. When the problematic item(s) are not highly correlated to other items in the context, the test developers only have to eliminate or revise the misfit item(s) instead of discarding the whole testlet.

This conclusion may be more meaningful to test developers than to curriculum specialists or teachers. Very often during the testlet development an item is found to be problematic in measurement or for other concerns such as ethnic or gender bias. As a result, the whole testlet is discarded because of the underlying assumption that a testlet is considered as a complete piece and all of its parts are clustered together closely and should not be separated. If one part goes wrong, the whole work is terminated. The results from this study imply that when context-dependent items are not highly correlated with each other, deleting the problematic item may not affect the remaining part of the testlet significantly. Therefore, one can still keep the technically sound items, and revise or eliminate the bad item, or, replace it with a new item. It is not necessary to discard the whole testlet or make any changes in other testlets either.

4. It seemed that an implicit factor other than the local item dependence affects the misfit original testlets. Even when the Rasch partial credit model was applied, unacceptable fit statistics were obtained.
5. Local item dependence effects may even exist in some developer-designed independent items in this study. However, they may be caused by random errors.
6. There was no significant different between the Rasch partial credit model and the Rasch dichotomous model in average person ability measures. Competitive estimates were obtained by both models.

Implications

What do the results and the conclusions based on those results in this study mean to the educational testing community? Primarily, they mean that test developers have to decide which scoring scale they should use for the context-dependent items in the item calibration. If the within-context item correlation is higher than the across-context correlation and a proper testlet fit is obtained, the Rasch partial credit model is theoretically preferable because the model provides a complete picture of the testlet and its characteristics that cannot be achieved by the dichotomous model. The partial credit model also may control the possible local item dependence effects for at least some of the context-specific testlets and provide more precise estimates of item parameters. However, there will be a lot of obstacles in the practical implementation if more than one scale is used for the same item format. For example, one important and practical factor is the cost of data analysis. Even though some evidence of local dependence has been shown here, it is almost impossible to score those items as testlets with the Rasch partial credit model and other items with the dichotomous model for such a large-scale statewide assessment, because the cost will be increased dramatically.

For test users such as curriculum specialists and teachers, more precise testlet calibrations mean better test items and further a better test design. One of the main reasons for constructing testlets in an achievement test is to assess a student's knowledge of a content area in a more comprehensive context that simulates the real-world situations. If the Rasch partial credit model can help control the potential local item dependence effects, educators can adopt the testlet format as a teaching tool and an alternative assessment in their curriculum development and classroom teaching, which further enhances student learning. However, this approach may also cause a lot of confusion and tension in the education community and to the public, especially parents and school boards, simply because of the complexity of the scoring process and educational measurement theory behind the process. The public relationship and political effects are not negligible factors too.

Limitations

Every study has its limitations. The major limitation of this study may be the quality of the data. Since the data were from a tryout administration, there were no previous item statistics available. Therefore, there was no reference of item quality, testlet formation or other related information.

Another limitation was the nature of the testlet formation. Because the original testlets here were designed to assess students' multiple traits (e.g., using life science, reflecting scientific knowledge), their items were not linked to a common factor. Therefore, it is unlikely that student abilities would be affected by a single context. If these testlets had been developed as unidimensional instead of multi-dimensional, the results may have been quite different.

Student motivation may also contribute to thei performance. Because it was a tryout and not an operational administration, the results did not have any impact on student records, and therefore, it did not matter if they performed seriously or not. Consequently, student attitudes may confound the results of the study.

Furthermore, for the simplicity of the study, neither the response patterns of the testlets nor the constructed-response questions were considered in the research design. Whether this would affect the results is not known.

Recommendations for Further Research

This study demonstrated a technique for analyzing testlets with potential local item dependence. Although the models functioned consistently, the lack-of-quality data left some uncertainties on the inconsistent final results. There is a need to use full operational data to conduct the study again to verify the outcomes.

Testlets in this study were multi-dimensional. It is necessary to use the models in this study to investigate the local item dependence with unidimensional testlets. It is anticipated that dimensionality of a testlet has an impact on the validity of the results.

As mentioned above, only the multiple-choice items within the testlets were used in the analysis. To fully investigate the local item dependence effects, full testlets, that is, multiple-choice items and constructed-response items, should be used in future studies.

In this study, only the fit statistic generated from the BIGSTEPS was used. Other statistics such as Q2 and Q3 (Yen, 1984a) were not considered in the analyses. In addition, R. Smith (April, 1996, personal contact) proposed a "between-fit" statistic contrary to Linacre and Wright's (1995) infit and outfit statistics. It will be helpful to the item/testlet analysis field to compare the efficiency of these and other currently available fit statistics.

In addition, individual person ability estimation by the Rasch partial credit model needs to be explored in comparison with that by the dichotomous scoring model.

Appendix A: A Sample Testlet in MHSPT Science

The MHSPT tryout test consisted of 30 independent multiple-choice items, 1 scientific investigation, 1 text criticism, and 4 context-dependent testlets with each having 4 context-dependent multiple-choice items and one constructed-response question. There were ten tryout forms with no item overlapping in any of two forms. However, each form was administered to two different schools for the equating purpose. A sample testlet is as follows.

Below is a data table which shows the melting and boiling points of common substances. Study the table. Then do Number 1 through 5.

Substance	Melting Point (°C)	Boiling Point (°C)
Water	0	100
Alcohol	-117	78
Nitrogen	-210	-196
Oxygen	-218	-183

1. Which substance should be a *liquid* at -90 degrees?

A water
B alcohol
C nitrogen
D oxygen

2. As each substance in the table is cooled down, the atoms and molecules undergo a

A physical changes as they move faster
B physical changes as they move slower
C chemical changes as they move faster
D chemical changes as they move slower

3. Because alcohol freezes and boils at lower temperatures than water, mixing alcohol and water could be a useful application for a

A better radiator coolant in cars during the summertime
B better windshield-washer fluid in cars during the wintertime
C clean and inexpensive alternative to gasoline
D clean and inexpensive alternative to engine lubricants

4. In order to change water from a solid to a liquid, energy must be

A removed
B added
C created
D destroyed

Appendix B: Tables

Table 2. Mean ϕ Coefficients for Items within Different Testlets by Form

Form	Type	Tlet.1	Tlet.2	Tlet.3	Tlet.4	Form
		Items	Items	Items	Items	mean
20	Original	.1203	.1169	.1731	.1180	.1321
	Random	.1761	.0790	.1298	.0725	.1144
	Reformed	.1799	.0681	.1258	.0644	.1096
21	Original	.1473	.0500	.3433	.1944	.1838
	Random	.0437	.0364	.0250	.0526	.0394
	Reformed	.1451	.1168	.1243	.0733	.1149
22	Original	.1838	.1292	.0747	.0934	.1203
	Random	.1287	.0300	.0492	.1484	.0891
	Reformed	.1378	.0673	.0500	.1890	.1110
23	Original	.1440	.1778	.0656	.2758	.1658
	Random	.0900	.0814	.1192	.0433	.0835
	Reformed	.1074	.0975	.1513	.0991	.1138
24	Original	.1126	.1376	.0620	.1714	.1209
	Random	.0866	.1175	.1170	.1474	.1171
	Reformed	.0431	.0835	.1247	.2429	.1236
25	Original	.1579	.1049	.1155	.0423	.1052
	Random	.1269	.0772	.1027	.0674	.0936
	Reformed	.0356	.1038	.0793	.0622	.0702
26	Original	.1421	.1243	.1824	.0980	.1367
	Random	.1455	.0972	.0941	.1209	.1144
	Reformed	.0758	.0422	.1516	.1809	.1126
27	Original	.1314	.0760	.2954	.1043	.1518
	Random	.0771	.0960	.1296	.0987	.1004
	Reformed	.0133	.0285	.1636	.2496	.1138
28	Original	.2306	.0318	.2487	.0970	.1520
	Random	.1510	.0585	.1215	.0730	.1010
	Reformed	.0366	.1230	.1044	.1275	.0979
29	Original	.2059	.1589	.0914	.1859	.1605
	Random	.1099	.0654	.0689	.1616	.1015
	Reformed	.1162	.1498	.1340	.0929	.1232
Mean	Original	.1576	.1107	.1652	.1381	.1429
By	Random	.1136	.0739	.0957	.0986	.0955
Testlet	Reformed	.0891	.0881	.1209	.1382	.1091

Table 4. Comparison of Original Testlets and Context-Dependent Items on Error and Fit by Form

Form 22

1	2	3	4	5	6
		Testlet	Context		Item
Orig.	SE of	Infit	Depend.	SE of	Infit
Testlet	Testlet	MNSQ	Item	Item	MNSQ
Testlet 1	.04	1.03			
			11	.07	1.03
			12	.07	.89*
			13	.08	.91*
			14	.07	1.05
Testlet 2	.04	.90*			
			28	.07	.97
			29	.07	1.02
			30	.07	1.00
			31	.07	.91*
Testlet 3	.04	1.03			
			45	.08	.83*
			46	.08	1.25*
			47	.07	1.16*
			48	.07	.90*
Testlet 4	.04	.91*			
			50	.07	.89*
			51	.07	.96
			52	.07	1.22*
			53	.07	.96

Table 4. (cont'd)

Form 23

1	2	3	4	5	6
		Testlet	Context		Item
Orig.	SE of	Infit	Depend.	SE of	Infit
Testlet	Testlet	MNSQ	Item	Item	MNSQ
Testlet 1	.04	.96			
			11	.07	.97
			12	.07	1.11*
			13	.09	.92
			14	.07	.99
Testlet 2	.04	.98			
			28	.07	1.05
			29	.07	.96
			30	.07	.96
			31	.08	.97
Testlet 3	.04	1.22*			
			45	.07	1.13*
			46	.08	1.16*
			47	.07	1.09*
			48	.07	1.10*
Testlet 4	.04	.76*			
			50	.09	.88*
			51	.07	.88*
			52	.07	.86*
			53	.09	.86*

Table 4. (cont'd)

Form 27

1	2	3	4	5	6
Orig. Testlet	SE of Testlet	Testlet Infit MNSQ	Context Depend. Item	SE of Item	Item Infit MNSQ
Testlet 1	.04	1.00			
			11	.08	.96
			12	.08	1.16*
			13	.07	1.04
			14	.07	.93*
Testlet 2	.04	1.12*			
			28	.07	1.02
			29	.08	1.31*
			30	.09	1.04
			31	.07	.92*
Testlet 3	.04	.79*			
			45	.09	.83*
			46	.07	.90*
			47	.08	.92*
			48	.08	.86*
Testlet 4	.04	.95			
			50	.09	1.31*
			51	.07	.90*
			52	.08	.88*
			53	.08	.85*

Table 10. CIs for One-Way ANOVA for Context-dependent Items

Form 20

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	.0112	.0575	.0287	-.0803	TO	.1027
Testlet 2	4	.0150	.0775	.0387	-.1082	TO	.1383
Testlet 3	4	-.0388	.0907	.0454	-.1831	TO	.1055
Testlet 4	4	-.0122	.0761	.0381	-.1334	TO	.1089

F ratio = .4236 probability = .7396

Form 21

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0009	.0487	.0243	-.0783	TO	.0766
Testlet 2	4	.1195	.0857	.0429	-.0169	TO	.2558
Testlet 3	4	-.1209	.1073	.0537	-.2917	TO	.0499
Testlet 4	4	-.0175	.0801	.0400	-.1450	TO	.1099

F ratio = 5.6103 probability = .0122

Form 22

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0331	.0843	.0422	-.1673	TO	.1011
Testlet 2	4	-.0262	.0499	.0249	-.1056	TO	.0531
Testlet 3	4	.0200	.1967	.0983	-.2930	TO	.3329
Testlet 4	4	.0002	.1372	.0686	-.2181	TO	.2184

F ratio = .1431 probability = .9322

Form 23

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0049	.0791	.0396	-.1308	TO	.1210
Testlet 2	4	-.0158	.0434	.0217	-.0848	TO	.0532
Testlet 3	4	.1130	.0281	.0141	.0683	TO	.1578
Testlet 4	4	-.1393	.0133	.0066	-.1604	TO	-.1182

F ratio = 18.6909 probability = .0001

Form 24

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0352	.1366	.0683	-.2526	TO	.1823
Testlet 2	4	.0527	.0933	.0466	-.0957	TO	.2011
Testlet 3	4	-.0254	.1438	.0719	-.2541	TO	.2034
Testlet 4	4	-.0532	.0730	.0365	-.1694	TO	.0629

F ratio = .6559 probability = .5946

Form 25

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0446	.1002	.0501	-.2040	TO	.1147
Testlet 2	4	-.0245	.0686	.0343	-.1336	TO	.0846
Testlet 3	4	-.0232	.0346	.0173	-.0783	TO	.0319
Testlet 4	4	.0578	.0335	.0168	.0045	TO	.1112

F ratio = 1.9327 probability = .1782

Table 10. (Cont'd)

Form 26

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0267	.0609	.0305	-.1237	TO	.0702
Testlet 2	4	.0004	.1374	.0687	-.2182	TO	.2190
Testlet 3	4	-.0205	.0264	.0132	-.0624	TO	.0215
Testlet 4	4	.0180	.1527	.0764	-.2250	TO	.2611

F ratio = .1431 probability = .9321

Form 27

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	.0186	.0985	.0493	-.1382	TO	.1753
Testlet 2	4	.0614	.1491	.0746	-.1759	TO	.2987
Testlet 3	4	-.1315	.0461	.0231	-.2048	TO	-.0581
Testlet 4	4	-.0314	.2023	.1012	-.3534	TO	.2905

F ratio = 1.4697 probability = .2722

Form 28

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0782	.0257	.0129	-.1192	TO	-.0373
Testlet 2	4	.1046	.1313	.0657	-.1044	TO	.3136
Testlet 3	4	-.0822	.0647	.0323	-.1851	TO	.0207
Testlet 4	4	.0430	.0513	.0257	-.0386	TO	.1247

F ratio = 5.5278 probability = .0128

Form 29

Testlet	# of Items	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Testlet 1	4	-.0492	.0917	.0458	-.1951	TO	.0966
Testlet 2	4	-.0566	.0832	.0416	-.1890	TO	.0758
Testlet 3	4	.1026	.1032	.0516	-.0617	TO	.2669
Testlet 4	4	-.0435	.1203	.0601	-.2349	TO	.1478

F ratio = 2.3075 probability = .1284

Table 11. Summary of Measured (Non-Extreme) Persons Fit by Form

1	2	3	4
Item/testlet Composition	Mean Measure	Infit MNSQ	Outfit MNSQ
Form 20 (n=1030)			
16 context-dependent items	-.27	1.00	1.01
4 original testlets	-.28	.97	.97
4 reformed testlets	-.37	.96	.96
16 MC independent items	.18	1.00	1.02
4 random testlets	.12	.94	.93
Form 21 (n=1046)			
16 context-dependent items	.06	.99	1.03
4 original testlets	.04	.95	.97
4 reformed testlets	-.02	.93	.94
16 MC independent items	-.29	1.00	1.02
4 random testlets	-.38	.97	.97
Form 22 (n=1044)			
16 context-dependent items	-.03	1.00	1.01
4 original testlets	-.04	.96	.97
4 reformed testlets	-.11	.95	.96
16 MC independent items	-.26	.99	1.05
4 random testlets	-.30	.92	.96
Form 23 (n=1051)			
16 context-dependent items	.25	1.00	.99
4 original testlets	.20	.95	.96
4 reformed testlets	.22	.94	.95
16 MC independent items	.36	1.00	1.00
4 random testlets	.37	.95	.94
Form 24 (n=1024)			
16 context-dependent items	.10	.99	1.04
4 original testlets	.08	.94	.95
4 reformed testlets	.00	.93	.96
16 MC independent items	.57	.99	1.02
4 random testlets	.71	.92	.92

Table 11 (cont'd

1	2	3	4
Item/testlet Composition	Mean Measure	Infit MNSQ	Outfit MNSQ
Form 25 (n=1016)			
16 context-dependent items	.07	1.00	1.02
4 original testlets	-.03	.96	.98
4 reformed testlets	.06	.96	.98
16 MC independent items	.21	1.00	1.04
4 random testlets	.10	.96	.96
Form 26 (n=896)			
16 context-dependent items	.15	.99	1.05
4 original testlets	.16	.96	.97
4 reformed testlets	.05	.93	.96
16 MC independent items	.63	1.00	1.00
4 random testlets	.78	.95	.96
Form 27 (n=945)			
16 context-dependent items	.14	.98	1.10
4 original testlets	.03	.93	.95
4 reformed testlets	.05	.92	.93
16 MC independent items	.47	.99	.99
4 random testlets	.61	.94	.96
Form 28 (n=944)			
16 context-dependent items	-.22	1.00	1.01
4 original testlets	-.36	.95	.97
4 reformed testlets	-.33	.96	.96
16 MC independent items	.01	1.00	1.00
4 random testlets	-.09	.96	.97
Form 29 (n=947)			
16 context-dependent items	.47	1.00	.99
4 original testlets	.47	.97	.97
4 reformed testlets	.53	.94	.94
16 MC independent items	.09	.99	1.03
4 random testlets	.14	.92	.96

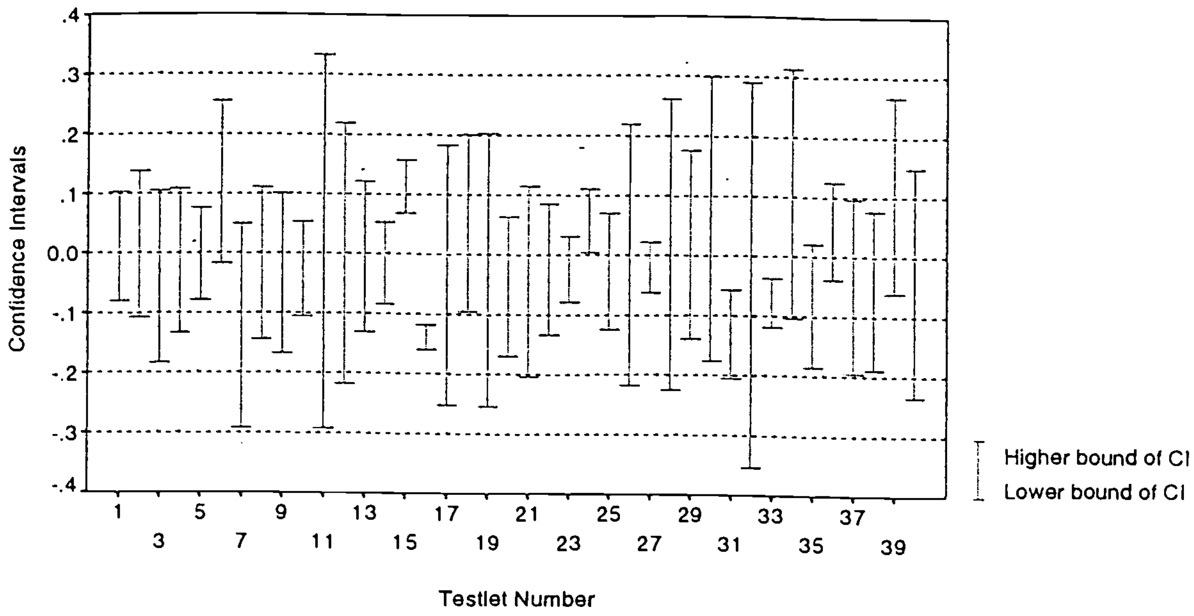

Fig. 1. CIs of ln(infit MNSQ) for Original Testlets

(Note: The first 4 testlets are from Form 20, and so on.)

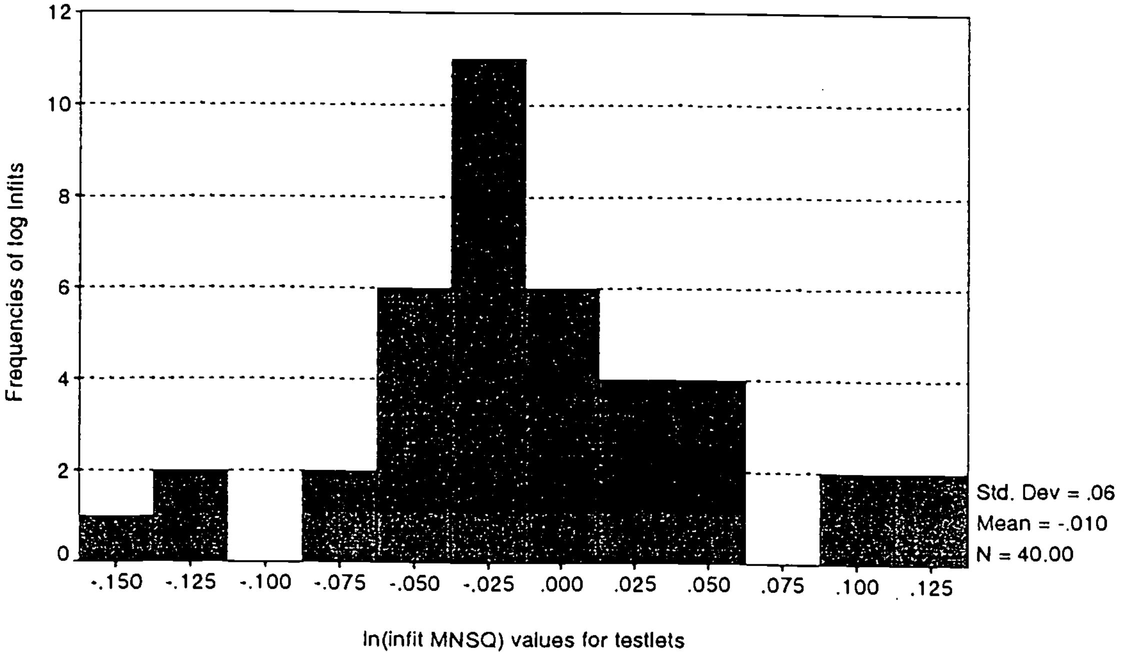

Fig. 2. Frequency Distribution of ln(infit MNSQ) for Original Testlets

REFERENCES

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

Biggs, J. B. & Collis, K. F. (1982). Evaluating the quality of learning: The SOLO taxonomy (structure of observed learning outcomes). New York: Academy Press.

Cattell, R., & Burdsal, C., Jr. (1975). The radial parcel double faCtoring design: A solution to the item-vs.-parcel controversy. Multivariate Behavioral Research, 10, 165-179.

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem solving ability. Journal of Research in Mathematics Education, 17, 206-211.

CTB Macmillan/McGraw-Hill. (1989). Comprehensive Tests of Basic Skills (4th ed., Technical Bulletin No. 1). Monterey, CA: Author.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 25(2), 327-346.

De Ayala, R. (1991, April). The Influence of Dimensionality on Estimation in the Partial Credit Model. Paper presented at the International Objective Measurement Workshop.

Ebel, R. L. (1951). Writing the test item. In E.F. Lindquist (Ed.), Educational Measurement (1st ed., pp. 185-249). Washington, DC: American Council on Education.

Engelhart, M. D. (1942). Unique types of achievement test exercises. Psychometrika, 7(2), 103-116.

Ercikan, K. (1993). Measurement Accuracy in Testlet Methodology. Paper presented at the National Council on Measurement in Education. Atlanta, GA.

Gerberich, J. R. (1956). Specimen Objective Test Items. New York: Longmans, Green and Co.

Gronlund, N. E. (1965). Measurement and Evaluation in Testing (5th ed.). New York: Macmillan.

Haladyna, T. M. (1991). Generic questioning strategies in the teaching of statistics. Educational Technology: Research and Development, 39(1), 73-82.

Haladyna, T. M. (1992). Context-dependent item sets. Educational Measurement: Issues and Problems, 11, 21-25.

Huynh, H. (1994). On equivalence between a partial credit items and a set of independent Rasch binary items. Psychometrika, 59(1) 111-119.

Linacre, J. M. (1995) BIGSTEPS, version 2.6 [computer software]. Chicago, IL:MESA Press.

Linacre, J. M. & Wright, B. D. (1995). BIGSTEPS, the User's Guide. Chicago,IL: MESA Press.

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

Mehrens, A. W., & Lehmann, I. J. (1984). Measurement and Evaluation in Education and Psychology (4th ed.). New York: Holt, Rinehart and Winston, Inc.

Rasch, G. (1980). Probabilitic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press. (Original work published by Copenhagan: Danmarks Paedogogiske Institut, 1960).

Rosenbaum, P. R. (1988). A note on item bundles. Psychometrika, 53, 349-359.

Sireci, S., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Szeberënyi, J. & Tigyi, A. (1987). The use of application test, a novel type of problem-solving exercise, as a tool of teaching and assessment of competence in medical biology. Medical Teacher, 9(1), 73-82.

Thissen, D., & Steinberg, L. (1988). Data analysis using Item Response Theory. Psychological Bulletin, 104, 385-395.

Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. Journal of Educational Measurement, 26, 247-260.

Wainer, H., & Kiely, G. (1987). Item clusters and computer adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of simulated hierarchical and linear testlets. Journal of Educational Measurement, 29, 243-251.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. Journal of Educational Measurement, 28, 311-324.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. Applied Psychological Measurement, 12(4), 353-364.

Wilson, M. & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. Applied Measurement in Education, I(4), 319-334.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? Invited debate at the AERA Annual Meeting. 1992.

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

Yen, W. M. (1984a). Effect of local item dependence on the fit and equating performance of the three parameter logistic model. Applied Psychological Measurement, 8(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for Managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

U.S. DEPARTMENT OF EDUCATION

EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

REPRODUCTION RELEASE

I. DOCUMENT IDENTIFICATION

Title: Examining Local Item Dependence Effect in a Large-Scale Science Assessment by a R. Partial Credit Model

Author(s): Jean W. Yan

Date: 5-22-97

II. REPRODUCTION RELEASE

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, or electronic/optical media, and are sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document. If reproduction release is granted, one of the following notices is affixed to the document.

Detach and complete this form and submit with your document.
This form may be copied as needed.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jean Yan

5-22-97

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

If permission is granted to reproduce the identified document, please CHECK ONE of the options below and sign the release on the other side.

☒ Permitting microfiche (4" x 6" film) paper copy, electronic, and optical media reproduction (Level 1)

OR

□ Permitting reproduction in other than paper copy (Level 2)

Documents will be processed as indicated, provided quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

OVER

Signature Required

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated on the other side. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Jean Yan

Printed Name: Jean Yan

Organization: Michigan Department of Education

Position: Educational Research Consultant

Address: P.O. Box 30008
Lansing, MI

Tel. No: 517-373-8393 Zip Code: 48909

III. DOCUMENT AVAILABILITY INFORMATION

(Non-ERIC Source)

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents which cannot be made available through EDRS).

Publisher/Distributor: ______________________

____

Address: ______________________

Price Per Copy: ______________________

Quantity Price: ______________________

IV. REFERRAL TO COPYRIGHT/ REPRODUCTION RIGHTS HOLDER

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

____

____

____