DOCUMENT RESUME

ED 480 871 TM 035 244

AUTHOR	Yen, Shu Jing; Bene, Nancy; Huynh, Huynh
TITLE	The Effect of Content Integration on the Construct Validity of Reading Performance Assessment.
PUB DATE	2000-04-00
NOTE	31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Integrated Activities; *Performance Based Assessment; *Reading Achievement; Reading Tests; Reliability; Science Tests; State Programs; *Test Content; *Testing Programs; Validity
IDENTIFIERS	Content Structure; *Maryland School Performance Assessment Program

ABSTRACT

Content integration in performance assessment involves mixing different areas of knowledge in one assessment. In this type of testing situation, assessment tasks are designed to measure the ability of students to solve problems by applying their knowledge and skills in multiple content areas. This study examined the effect of integrated science and reading items on the reliability and construct validity of reading assessment for the Maryland School Performance Assessment Program (MSPAP). Using the MSPAP 1998 reading data, this study demonstrated that the integrated science-reading items provide reliable and valid information about the student's reading abilities that are comparable to the nonintegrated items. Results suggest that the integrated science-reading items did not compromise the integrity of reading scales in question. There is evidence that the scales are essentially unidimensional. However, there are minor task-related factors that warrant further analysis on the latent structure of the reading scales. This finding is complemented by the high local item dependency among tasks. (Contains 4 tables and 23 references.) (Author/SLD)

The Effect of Content Integration on the Construct Validity of Reading Performance Assessment

Shu Jing Yen
Nancy Bené

Maryland State Department of Education

Huynh Huynh

University of South Carolina

Paper Presented at the 2000 Annual Meeting of the National Council on Measurement in Education

Abstract

Content integration in performance assessment involves mixing different areas of knowledge in one assessment. In this type of testing situation, assessment tasks are designed to measure the ability in which students may solve problems by applying their knowledge and skills in multiple content areas. This study examined the effect of integrated science and reading items on the reliability and construct validity of reading assessment for Maryland School Performance Assessment Program (MSPAP). Using the MSPAP 1998 reading data, this study demonstrated that the integrated science-reading items provide reliable and valid information about student's reading abilities that are comparable to the non-integrated items. The results suggested that the integrated science-reading items did not compromise the integrity of reading scales in question. There is evidence that the scales are essentially uni-dimensional. However, there are minor task-related factors that warrant further analysis on the latent structure of the reading scales. This finding is complimented by the high local item dependency among tasks.

Perspectives

In recent years several states have embraced content area integration as a preferred method for delivering curricula. This process involves the creation of instructional activities that combine skills and knowledge from multiple content areas such as reading and writing, reading and science, and mathematics and science. The integrated activities are expected to foster the development of general cognitive processes that may be useful in dealing with these problems (Herman, et al., 1992). Along with the use of integrated instructional activities, many states have relied on integrated tasks as part of their statewide assessment programs (MSDE, 1999; NYSED, 1998). Integrated tasks are also used in comparative studies of mathematics and science achievement across countries (Harmon et. al, 1997).

In performance assessment, content integration typically involves establishing a real-life context with a sense of purpose and intended audience (MSDE, 1999). The assessment tasks then require examinees to apply skills and knowledge from multiple content areas in order to solve real-life problems. (We will refer to these as integrated tasks in the remainder of this paper.) There are, of course, assessment tasks (non-integrated tasks) that are tied to only one content area.

For instructional purposes, test scores are needed for each separate content area (such as reading). To fulfill this need, item responses to integrated and non-integrated assessment tasks are often calibrated together in order to create the needed scale. There are a number of issues relating to the use of integrated and non-integrated assessment tasks (or items) in creating a scale for a given content area. One such issue relates to the nature of the construct assessed by the combined items. The strong person by task

interaction often found in performance assessment (Shavelson et al. 1997) may imply that tasks involving more than one content area may assess a construct that is different from the construct tapped by items that involve a single content area. The interaction may also imply that tasks that involve different combinations of content integration may assess different constructs. For example, reading tasks integrated with writing may measure different construct than reading tasks integrated with science.

Using data from 1994 Maryland Performance Assessment Program (MSPAP), Ercikan et al. (1997) examined the effect of mathematics and science integration on the validity and reliability of the separate mathematics and science scales. Their results indicate that the integrated mathematics and science items do not compromise the integrity of the separate scales. The contributions to test reliability for the integrated and non-integrated items are similar. The confirmatory factor analysis suggests that the integrated and non-integrated items measure the same construct.

The results obtained by Ercikan and her associates are based on science and mathematics. It is well known that these two content areas are quite similar to each other in terms of cognitive processes; so it may not be surprising that the integrated and non-integrated items measure similar constructs. It may be of considerable interest to know if construct similarity still hold in integrating assessment tasks from (seemingly unrelated) areas such as reading, science, mathematics and social studies.

Objectives

This study focuses on the effect of content integration to the construct validity of reading assessment that integrates with science. Data from the Maryland School Performance Assessment Program (MSPAP) are used to answer two fundamental questions:

(1) Do reading assessment tasks that integrated with science assess different constructs from non-integrated reading task?

(2) Is it meaningful to combine the scores on integrated and non-integrated items to yield a single reported total score?

The similarity of constructs measured by integrated and non-integrated items was examined by item-factor analysis and item-test correlation. The effect of calibrating integrated and nonintegrated together to yield a common scale was examined in terms of item fit and local item dependency. The validity implication of scaling and deriving scale scores including both item types on the same scale are discussed.

Data

This study uses reading assessment data for grade 8 from the 1998 MSPAP assessment. MSPAP is a statewide performance assessment for grade 3, 5 and 8 students. Three non-equivalent forms of the test are administered to randomly equivalent groups of students at each school, with six to eight tasks per form. Some tasks are single content area tasks (non-integrated tasks) and others are integrated (integrated tasks). There are two reading tasks in each of the two forms we used: one non-integrated reading task and one reading-science integrated task. A caveat must be made about these tasks. In two of

the tasks, one writing question was integrated at the end of the task. Strictly speaking, the effect of integrating writing within a reading task can also be considered. Since the only purpose of integrating writing items at end of some MSPAP tasks is to provide a rich context in which writing can be assessed, its potential effect to reading assessment is assumed to be minimum and therefore not being considered in this study.

A random sample of sample of 7502 students was extracted from the statewide data file for each form. Only the grade 8 data were chosen in this study because instructional content integration in the middle school (grade 8) is not as prevalent as in the primary school (grades 3 and 5). All MSPAP items are constructed response and are scored by Maryland teachers using activity- specific scoring tools. Item Response Theory (IRT) procedures were used to calibrate all reading items together in each of the three forms. The scaling procedures were implemented using the two-parameter partial credit model (Yen, 1993; Yen & Ferrara, 1997).

Method

Item Factor Analysis

The similarity of constructs measured by integrated and non-integrated items was first examined by item-factor analysis. Unidimensionality is an assumption of the item response theory (IRT) model used to calibrate MSPAP items. The unidimensionality assumption is seldom strictly met since there is always a possibility that other factors present that can affect test performance. The purpose of the item-factor analysis is to examine whether the test is 'essentially unidimensional' (Stout, 1990).

We consider data obtained from a random sample of 1,000 student each form, drawn from among 7,502 students. Sample size of 1,000 was chosen because this is adequate sample size for getting the asymptotic distribution free estimators used in this study (Bentler, 1995). Principal component analysis was first applied to examine the number of principal components that need to be retained based on the eigen-value-greater-than-one rule. Principal factor analysis was then conducted with the maximum correlation with any other items used as the communality estimates in the diagonal. The factors were rotated toward the hypothesis matrix by oblique promax solution.

To factor analyze mathematics and science integrated MSPAP tests, Ercikan et al. (1997) fit a factor model where the integrated and non-integrated items form a different but correlated factor. It is hypothesized that items that are associated with the non-integrated reading task form a different factor than those items associated with the integrated reading task. Such a model may be interpreted as the single-level realization of a hierarchical factor model, in which at a second order level, correlated first-order factors depend on a single underlying general factor (Bentler, 1995).

We also consider an alternative model: a bi-factor model that includes a general factor for all the items, plus two specific factors for each task (Schmid & Leiman, 1957). Unlike the hierarchical factor model, items are allowed to load directly on the general dimension. It is plausible for performance assessment tests where the primary dimension measures the targeted process skill and additional factors describe content area knowledge within tasks. In this model, item would be conditional independent between tasks, but conditionally dependent within task (Hozinger, K. J. & Swinefor, F. 1937).

Confirmatory factor analysis was conducted using EQS (Bentler, 1998). For ordered categorical data that are not themselves normally distributed, but may be assumed to reflect underlying variables that are normally distributed, it is recommended that polychoric correlation is used to represent the population correlation matrix (Gorsuch, 1983). However, estimation of polychoric correlation estimation requires a large sample and its accuracy may be problematic for small sample size. Since the categorical data methods require the cross-tabulation tables of the categorical variables. These tables will be sparse if there are too few subjects in some of the cells, then the computational procedures can break down (Bentler, 1995). Therefore, we chose to use the variance and covariance matrix for the estimation after confirming that the multivariate normality assumption seems to hold. The standardized skewness and kurtosis were examined for univariate normality. Mardia's coefficient provided in EQS measures of the degree to which the assumption of multivariate normality has been violated. Its normalized estimate is distributed as unit normal variate. Large value indicates positive kurtosis and large negative value indicate significant negative kurtosis (Bentler, 1995).

We consider the fully iterated arbitrary Generalized Least Square estimator provided by EQS, which is Browne's asymptotic distribution free (ADF) estimator (Browne, 1982, 1984). The Generalized Least Square method has been shown to be robust to the violation of normal theory. The ADF estimation procedure does not assume multivariate normality of the response data. Rather, the assumption of multivariate normality is replaced by Elliptical theory which allows symmetrical distributed variables

to have either heavier or lighter tails than their normal theory counterparts, but the kurtosis of the variables are assumed to be equal (Browne, 1982, 1984; Bentler, 1995).

In conducting item factor analysis with sizable number of items, most patterns were realized only once, and expected frequencies are near zero. In this case, the usual chi-square approximation for the distribution of multinomial goodness of fit statistics is inaccurate. Haberman (1977) has shown, however, that the difference in fit statistics for alternative models is distributed in large samples as chi-square, with the degrees of freedom equals to the difference in numbers of degrees of freedom of the alternative models, even when the frequency table is sparse. These differences will be used in the comparisons of alternative models.

Item-test correlation

The item-test correlations are indicators of the strength of the relationship between what is being measured by the item and the overall construct measured by the set of items in the test. If the integrated items assess a different construct than the non-integrated items, they would be expected to correlate lower with the total test score than the non-integrated items.

Item Fit and Local Item Independence

The effects of content integration were also examined in terms of item fit and measures of local dependency. If the integrated items assess a different construct than non-integrated items, these items may be expected to fit the model poorly. Two different types of model fit analyses were used. The first model fit measures the fit of item

responses to the model. It was evaluated by a generalization of the Q_1 statistics to a 2-parameter partial credit model (Yen, 1981). The second type of fit analysis examined pair-wise local item dependence among the items by a generalization of Q_3 statistic to a 2-parameter partial credit model (Yen, 1993).

The Yen's Q_1 statistic compares the observed and predicted trace lines. To calculate this statistic, a trait estimate is obtained for each student using the student's responses to all the items in the scale. Using the trait estimate and the item parameters estimated from a two-parameter partial credit model, the student's expected performance on each item are estimated. The deviation between the observed and predicted performance is calculated and is initially referenced to a chi-square distribution. A mathematical conversion was done to reference the chi-square statistic to a Z-score distribution for ease of interpretation (Yen, 1993).

A well-known problem of using chi-square statistic as measure of model fit is, for large sample sizes, small deviations from model predictions can yield large chi-square value with no practical importance. Rules of thumb have been developed to yield cut scores of Z value for flagging items with are practical significant. For the sample size of about 7500, items with Z greater than 19 were flagged as poor fit (1992, MSDE).

Local item dependency (LID) violates a fundamental assumption of IRT. Writers such as Sireci, Thissen, and Wainer (1991) have found existence of local item dependency in reading comprehension. They found evidence of LID among reading comprehension items linked to the same reading passage.

If item factor analysis suggests that there were some residual correlations left unexplained when fitting a one-factor model, these items may define the second or higher

factor. They are said to be locally dependent in the IRT terminology. In this situation, fitting a IRT model that assumes unidimensionality may be not appropriate. Therefore, it is necessary to examine the extent and the nature of the LID. It helps address the question in terms of whether we can differentiate properly the examinee's performance on both integrated and non-integrated items in a reliable way. It may also have implications about the validity of generalizing scores across integrated and non-integrated tasks (Yen, 1993).

The statistical procedure in identifying pairs of locally dependent items is based on a method proposed by Yen (1993). This method has been shown to yield similar results to other statistical methods for detecting LID (Ferrara, Huynh, & Michaels, 1996). To implement this procedure, the deviation between observed and predicted item performance for each examinee was obtained. A trait estimate is obtained for each student using the student's response to all items in the scale. Using the trait estimate and item parameters, the student's expected performance on each item is determined. The deviation between the student's observed and expected item performance is calculated. The measure of LID between two given items is the correlation of these deviation taken over students.

In order to have a context in which to evaluate the size of Q_3 statistic, it should be noted that in calculating this statistic, item score is included in both the estimating of expected performance and the observed performance (Yen, 1993). Therefore there is a known negative bias in Yen's Q_3 statistic. When the local independence is true, the expected value is approximately $-1/(n-1)$. In this case, the expected value of Q_3 statistic is

-0.08 for a 13-item test. We used $|Q_3|>.20$, as a criterion for flagging significant LID for further examination.

Analysis and Results

Item Factor Analysis

The principal component analysis showed that the two correlation matrixes are both positive-definite. Two components were retained based on the eigen-value-greater-than-one rule for both forms. For form A, the two large eigen values are 4.52 and 1.51. The variance explained by the first and second factor was 38% and 13%, respectively. Together thy account for 51% of the standardized variance. For item factor analysis, this amount is considered acceptable because items by themselves are not very reliable. For form B, the two large eigen values are 5.35 and 1.58. The variance explained by the first and second factor was 41% and 12 %, with the total of 53%. For both forms, there seems to be a dominating factor underlying the data.

Principal factor analysis was then conducted with the maximum correlation with any other items used as the communality estimates in the diagonal. The first two factors were rotated toward the hypothesis matrix by oblique promax solution. The correlation between the two factors is 0.55 for form A and 0.62 for form B. For both forms, the factor pattern show that all items related to the non-integrated task have high and positive loadings on the one factor while the science/reading integrated items have high and positive loadings on another factor.

The plot of factor pattern, based on the oblique rotation, revealed the reference axis goes through two clusters of items for both forms. Since the Harris-Kaiser rotation

did not improve the alignment, therefore, it was concluded that this is the best rotation attainable. Inspection of the plot reveals that each factor is dominated by items from a particular task. In other words, all the non-integrated reading items are well aligned on one principal axis; and all the reading and science integrated items aligned on the other axis.

Confirmatory factor analysis was conducted using EQS (Bentler, 1998). Due to the categorical nature of some of the items examined, both the univariate and multivariate normality assumption were carefully examined. For form A, no items exhibit extreme skewness or excessive kurtosis. The normalized Mardia's coefficient estimate (.46) was not significantly different from zero. The assumption of multivariate normality appears not have been violated. For form B, no items exhibit extreme skewness or excessive kurtosis either. However, the normalized Mardia's coefficient estimate (1.37) indicate that the multivariate kurtosis may be excessive, but not statistically significant.

The likelihood ratio tests of significance for the one factor, hierarchical, and bi-factor model as reported in Table 1 are all significant. The difference in fit statistics for alternative models is distributed in large samples as chi-square, with the degrees of freedom equals to the difference in numbers of degrees of freedom of the alternative models (Haberman, 1977). These differences will be used in the comparisons of alternative models. The one-factor model has the worst fit among the three models. The hierarchical model provided a significant improvement in fit over the one-factor model ($G^2 = 188$ for form A, $G^2 = 191$ for form B, $df = 1$, $p < .00001$),. The bi-factor model provided a significant improvement in fit over the hierarchical model ($G^2 = 56$ for Form A, $G^2 = 164$ for form B, $df = 14$, $p < .00001$), suggesting that a general reading dimension

plus two task specific dimension are required to fully describe the data. The Chi-square to the degrees of freedom ratio (G^2/ *df*) and Comparative Fit Index (CFIT) also suggests that the bi-factor model is the best fitting model among the three compared. The AIC index, however, suggests that for form A, the bi-factor model may be overfitting and the hierarchical model is a better fitting model. This could due to AIC index's tendency to reward more parsimonious models.

Table 1

Confirmatory Factor Analysis

	Model	G^2	*df*	G^2/ *df*	CFIT	AIC
Form A						
	One Factor	334.56	65	5.15	.50	13.16
	Hierarchical	146.44	64	2.29	.88	-349.63
	Bi-Factor	89.95	50	1.80	.94	-271.11
Form B						
	One Factor	503.51	65	7.75	.60	-160.62
	Hierarchical	312.12	64	4.88	.72	-178.46
	Bi-Factor	148.03	50	2.96	.89	-206.46

There is significant residual variation from a one-factor model. For form A, average absolute standardized residuals, from the one-factor model, is 0.08. However, most of the standardized residuals are relatively small. Seventy-two percent of the standardized residuals with absolute value less than .10. The largest standardized residual (.26) is accounted for by the residual correlation between item 6 and 7. The implication is when a data are calibration using IRT models that assume unidimensionality, items with high residuals may be locally dependent on other items, to

be discussed later. With the bi-factor model, the average absolute standardized residuals reduced to .03 and none of the standardized residuals are greater than .10.

Similar results were found for form B. Average absolute standardized residuals from the one-factor model, is very small: .05; and 67% percent of the standardized residuals with absolute value less than .10. The largest standardized residual (.26) is accounted for by the residual correlation between item 4 and 5. With the bi-factor model, average absolute standardized residuals reduced to .05 and 80% percent of the standardized residuals are less than .10. However, the standardized residual correlation for item 4 and 5 remains high (.20). Although fitting a more complicated model decreased the overall standardized residuals, there are still relatively large co-variation between these two items still not accounted for by the bi-factor model.

The results from the item factor analysis seems to indicate that the test is essentially unidimensional (Stout, 1990). The first latent root is 38% and 41% of the variance for form A and B respectively, while the rest explain very little of the variance. It is clear that the vast preponderance of variance and covariance is associated with the first factor.

The bi-factor model, although did not fit the data based on the fit statistic, does indicate that there is a general latent dimension and two minor task-related dimensions. The existence of a dominating factor is supported by the general factor. There are clearly task-related factors. However, whether the minor factors account for a lot of the observed covariance among the items requires further investigation. These two task-related factors appear to be content-based local dependence (Yen, 1993) among the items.

Item-test Correlation

Cronbach's coefficient alpha, proportion correct, item-test correlations, and Yen's Q_1 statistic are reported in Table 2 for both forms. Coefficient alpha suggests that the reliability of both forms is very good, given that there are only 13 items in each form. The proportion correct was calculated for each item by dividing student's average score on the item by the maximum possible score on the item. For form A, integrated items seem to be more difficult (average proportion correct=.40) than the single content area item (average proportion correct=.59). Interestingly, the item-test correlations suggest that the integrated items have a stronger relationship with test than the non-integrated items. For form B, the integrated items are, on average, more difficult than the non-integrated items; however, the item-test correlations for the non-integrated items seem to be higher than those of the integrated items. The item-total correlations are affected by the number of score levels, higher numbers of score levels tend to have higher item-test correlations. This could explain why different results were obtained for the two forms. However, the overall results suggest that the integrated items measure more or less the same construct as the non-integrated items.

Item Fit

Yen's Q_1 statistics are reported in the last column of Table 1. For both forms, all 7,502 student's responses were included in the computation. The fit statistics suggest that the integrated items fit the two-parameter partial credit model as well as the non-integrated items. Only item 13 in form B was flagged as a poor fit using the Z>19 criterion discussed earlier. Item 11 in form B also showed a high fit statistic, although it

was not flagged. Both items are associated with the non-integrated reading task in form B. Careful examination of the expected and observed tracelines seem warranted to further explored where the misfit occurred. Deviation between observed and expected frequencies for each decile indicate that the misfit is due to sparse data at low end of the ability distribution for item both items. A problem that is not unusual in fitting performance assessment items to IRT model.

Table 2

Proportion Correct, Item-Test Correlations, and Item Fit statistics[1]

	Item	Score Levels	Proportion Correct	Item-Test Correlation	Yen's Q_1
Form A (Cronbach's alpha=.86)					
Integrated					
	1	3	.74	.56	3.73
	2	3	.41	.63	10.60
	3	4	.47	.62	6.86
	4	4	.26	.66	10.34
	5	3	.36	.63	5.86
	6	4	.24	.65	5.68
	7	4	.34	.65	13.98
Non-integrated					
	8	3	.45	.54	2.11
	9	3	.52	.62	2.94
	10	3	.73	.54	9.88
	11	3	.62	.62	7.01
	12	3	.54	.59	2.21
	13	3	.65	.61	3.60
Form B (Cronbach's alpha=.88)					
Integrated					
	1	3	.77	0.61	5.80
	2	2	.72	0.50	7.71
	3	4	.59	0.62	16.85
	4	3	.45	0.66	9.31
	5	3	.35	0.63	4.47
	6	3	.55	0.67	3.96
	7	3	.33	0.56	6.71
Non-integrated					
	8	3	.69	0.65	9.03
	9	4	.56	0.72	9.03
	10	3	.51	0.63	6.04
	11	4	.42	0.69	18.60
	12	3	.63	0.69	10.62
	13	3	.61	0.67	24.56

[1] Sample size is 7,502 for both forms

Local Item Dependency

If items were locally dependent, then there would be nonzero residual correlations after accounting for the first factor. Based on the results from item factor analysis, it seems that there are some residuals correlations left unexplained under a one-factor model. Therefore, we should expect some items to exhibit local dependency when scaling the items using item response models that assume unidimensionality. The issue to explore here was whether the items showed LID due to communality of within task or across tasks.

To investigate this issue, the Q_3 statistic was separated into Within-Task and Across-Tasks. For a set of n item test, there are n(n-1)/2 pairs of Q_3 statistics. The Within-Task LID was further sorted by task, each has 7 and 6 items respectively. Therefore, there are 21 within integrated task Q_3 statistics and 15 within non-integrated task Q_3 statistics to examine. Across-Tasks LID involves examining the dependency between two sets of items from each task, resulting in 42 (7 x 6) Q_3 statistics. The frequencies of Q_3 values ware summarized in Table 3.

The criterion used for flagging significant LID is when $|Q_3|>.20$. The stem and plot frequency shows that there is only one item pair (item 8 and 9) in form A exhibits high and positive within task LID. These two items were organized in steps, knowing item 8 increases the chances of a student's chance of getting a high score on item 9. This seems to reflect what Yen (1993) refers to as Item-Chaining effect. Two item pairs (item 4 and 5, item12 and 13) had a high and positive within task LID in form B. They also reflect the Item-Chaining effect. Positive LID means that if a student performs higher

than expected on one item, he or she will probably perform higher than expected on another item, and vice versa.

Table 3

Frequency of the Within-Task and Across-Tasks Q_3 Statistic

Form A

	Within Task LID			Across-Tasks LID	
	Integrated		Non-Integrated		
-0.2		-0.2		-0.2	**17**
-0.1	1	-0.1		-0.1	25
-0.0	13	-0.0		-0.0	
0.0	4	0.0	8	0.0	
0.1	2	0.1	6	0.1	
0.2		0.2	**1**	0.2	
Sum	21		15		42

Form B

	Within Task LID			Across-Tasks LID	
	Integrated		Non-Integrated		
-0.3		-0.3		-0.3	**1**
-0.2		-0.2		-0.2	**16**
-0.1		-0.1		-0.1	25
-0.0		-0.0	8	-0.0	
0.0	13	0.0	5	0.0	
0.1	7	0.1	1	0.1	
0.2		0.2	**1**	0.2	
0.3		0.3		0.3	
0.4		0.4		0.4	
0.5	**1**	0.5		0.5	
Sum	21		15		42

Cross-Task LID occurred more frequently than within task LID and merits further investigation. This result seems to contradict the expectation that within passage LID would be high when items are tied to a particular passage (Hozinger, K. J. & Swinefor, F. 1937). Table 4 reports the seventeen item pairs being flagged as having large and negative Q_3 statistic for both forms. Negative LID means if a student performs lower than expected on one item, he or she will probably perform higher than expected on another item, and vice versa. The "*" indicate the LID was identified between the item pair. They seem to reflect what Yen (1993) refers to as the Content Related LID. For form A, the science and reading integrated task measures the extent to which the examinees can extract scientific information from the text (Reading for Information). The non-integrated reading task, in contrast, requires the examinees to identify and explain why one particular reading passage was better in helping students perform an investigation (Reading to Perform). For form B, the integrated task measures whether the students can follow directions in text to perform a scientific experiment (Reading to Perform). The non-integrated task, on the other hand, asks student to discuss a poem that they read on the test (Reading for Literacy Experience).

Table 4

Across Tasks LID: Item Pairs with maximum Q_3

Form A

		Integrated (Reading for Information)						
	Item	1	2	3	4	5	6	7
	8					*	*	
Non-integrated	9			*	*	*	*	
(Reading to Perform)	10							
	11			*	*	*	*	
	12				*	*	*	
	13			*	*	*	*	

Form B

		Integrated (Reading to Perform)						
	Item	1	2	3	4	5	6	7
	8				*	*		
No-integrated (Literary Experience)	9	*		*	*	*	*	*
	10				*	*		
	11				*	*		
	12				*	*		
	13				*	*	*	

Conclusions and Discussions

Performance assessment differs in many aspects from the traditional multiple choice items. There is a need to cover a broad range of content in an authentic, real-life context. Rather than thriving to create independent items, a deliberate effort made when constructing multiple choice items, performance assessment engages students in a series of items related to the same topic. To meet the goal of authenticity, it is necessary to create items that may be dependent. To meet the goal of mirroring classroom instruction, it is necessary to create items integrating different content areas in the assessment. Content integration, in particular, encourages students to involve a variety of skills either learned from their curricular or prior knowledge. Such performance assessment is likely to face greater psychometric challenge then the multiple-choice items. It is therefore essential to gauge the flexibility offered by performance assessment against any potential negative psychometric consequence.

The effect of content integration on the construct validity of reading assessment was examined in this study. Item Factor analysis and item-test correlation was conducted to examine the similarity of the construct measured by the integrated and non-integrated items. The results from the exploratory factor analysis indicates that the test is essentially unidimensional (Stout, 1990). Size of factors after the first, as opposed to their statistical significance, is one major indication. The vast preponderance of variance and covariance is associated with the first factor while the rest of factors explained very little of the total variance.

There are statistically significant factors for the task-related factors, orthogonal to the general factor, as indicated by the confirmatory factor analysis. However, there is

clear evidence that the integrated items measure predominantly the same construct as the non-integrated items. The analysis based on item-test correlations give additional support of this conclusion.

The bi-factor model indicates that there is a general latent dimension and two minor task-related dimensions. The existence of a dominating factor is supported by the general factor. There are clearly task-related factors. However, whether the minor factors account for much of the observed covariance among the items requires further investigation. Since the main purpose of this study is not to examine the latent structure of the reading items in question, further analysis on the relationship between the major reading factor and the minor task factors may shade some light on this issue.

Stout (1990) establishes that "essential unidimensionality" and "essential independence" is sufficient to justify unidimensional scoring. Essential independence differs from local independence in that small, specific factors may be present but as the number of items becomes large, the average residual covariance from a one-factor model approaches zero. We found that the average residual covariance from a one-factor model was close to 0, indicating that essential independence appears to hold for the current data sets.

Is it meaningful to combine scores from integrated and non-integrated items, to yield a single reported score on reading? The evidence is clear that the scores definitely may be combined especially when IRT weights are used. Our analysis showed that integrated items fit the 2-parameter partial credit model as well as the non-integrated models. The existence of the minor task related factors seem to cause some Cross-Task LID, however, the problem does not appear to have negative measurement consequences.

For example, item 6 and 7 in form A, although showed large standardized residual from the one-factor model, fit the 2-parameter partial credit model very well.

The Within-Task LID, although present, is no more problematic than assessment based on multiple-choice items (Yen, 1993). For the MSPAP reading assessment, efforts were made to write items which elicit student's responses that are as independent as possible within a task. It is evident that this effort results in minimum LID within tasks. Within-Task LIDs we found are all due to Item-Chaining. However, Item- Chaining is desirable in performance assessment because it models real life situations.

Most of the LID problems seem due to the communality across tasks. This finding is consistent with the findings from the item factor analysis. There seems to be task specific factors being measured, orthogonal to the general reading factor. In particular, 'Performing a Task' seems to involve specific cognitive skills that are above and beyond the general reading dimension measured by all the items. The cognitive demand for 'Reading for Information", similarly, involves some unique cognitive skills that are beyond the general reading dimensions. Since these additional factors may be desirable because they mirror classroom instructional practice, further studies in the cognitive skills that are required in answering the integrated versus non-integrated items may facilitate a deeper understanding about student's learning.

Limitations of the Study

It would be valuable to examine the performance of other LID measures other than Q_3 statistic. Measures that are less affected by the item scores would be more desirable. For a relatively short test such us ours, substantial negative Q_3 values are

expected due to part-whole contamination. For example, if a student score relatively higher on a non-integrated item than an integrated item, the student's true response will be the average of two item scores. Then the student's expected performance would be too low for the non-integrated item and too high for the integrated items. This may explain the large number of cross-task LID found in this study.

Although the results suggest Cross-Task LID did not result in poor fitting items when IRT calibration was applied, the impact of Cross-Task LID to test information and standard error of measurement (SEM) need further investigation. Thissen, Steinberg, and Mooney (1989) and Sirei, Thissen, and Wainer (1991) pointed out that an inappropriate assumption of local item independence produces overestimates of test information and reliability and underestimation of SEM. They recommended the use of testlets. One reason testlet were not currently used in MSPAP was that items can form a testlet only if they belong to the same passage or task. Plus when testlets are formed, the items contributing to the testlet no longer remain as separate items in the total test scores. In the assessment where reporting outcome scores are essential, the use of testlets diminishes the meaning of outcome score. Other innovative ways of removing Cross-Task LID should be further researched to provide an alternative to testlet.

Content integration often involves items that measure a range of different content. It was meant to encourage students to involve a variety of skills they learned in other content area, from their curriculum, or their prior knowledge. However, one negative consequence is the potential LID when items measure unique content. When performance is differentially affected by exposure in the curriculum or prior knowledge,

items may show LID (Yen, 1993). It would be valuable to examine whether they may show differential item functioning as well.

These additional content-related factors, while desirable because they measure important dimensions in the context of integrated assessment, may represent some nuisance dimension that the test did not purport to measure. Shealy and Stout (1996) conceptualize that differential item functioning (DIF) is the multidimensional influences on item responses that render a test less valid for one group of examinees than for another. They utilize the terms, *target ability,* which is a latent trait that test is intended to measure, and *nuisance determinates*, as the unintended traits that influence some portion of the examinees' responses to test items. Therefore, it is essential to investigate DIF in an effort to determine the degree of influence of 'nuisance determinants' on responses to the test items. For example, it would be valuable to identify whether the nuisance determinates are related to Content-Based LID and whether the DIF can be expressed in one or more items whose responses depend on the contents.

References

Bentler, M. B. (1995). EQS Structual Equations Program Manual.

Bentler, M. B. (1998). EQS/PC. Version 3. Multivariate Software Inc.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.). *Topics in Applied multivariate analysis,* pp. 72-114. London: Cambridge University Press.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structure. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Ercikan, K., Schwarz, R., Weber, M., Michaels, H., & Farrara, S., (1997). *The Effect of Integrated Items on the Validity and Reliability of Tests: Science and Mathematics Integration in a Statewide Performance Assessment.* Paper presented at the annual National Council for Measurement in Education, Chicago.

Ferrara, S. , Huynh, H., & Michaels, H. (1996) Contextual explanations of local dependence. *Journal of Educational Measurement, 36(2)*, 119-140.

Gorsuch, R. L (1983). *Factor Analysis.* Lawrence Erlbaum Asso.

Haberman, J. S. (1977). Log-linear models and frequency tables with small expected cell counts. *Annuals of statistics, 5,* 1148-1169.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzales, E. J., & Orpwood, G., (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study.* International Association for the Evaluation of Educational Achievement (IEA).

Herman, J. L, Aschbacher, P. R., & Winter, L. (1992). A practical guide to alternative assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Hozinger, K. J. & Swinefor, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Maryland State Department of Education (MSDE), 1992. *Final Technical Report, Maryland Performance Assessment Program, 1991.* Maryland State Department of Education.

Maryland State Department of Education (MSDE), 1999. *Improving Classroom Assessment: A Toolkit for Professional Developers.*

New York State Department of Education (NYSED), 1998, *Revised Regents Comprehensive Examination in English: Test Sampler Draft, Spring, 1998.*

Schmit, J., &Leiman J. M. (1957). The development of hierarchical factor solution, *Psychometrika, 22*, 83-90.

Shavelson, R., Baxter, G., Pine, J., (1997). Performance Assessments in Science. *Applied Measurement in Education, 4(4)*, pp. 347-362.

Shealy, R., & Stout, W. (1996). An item response theory model for test bias. In Wainer, & P. Holland (Eds.), *Differential Item Functioning, theory and practice*. Hillsdale, NJ: Erlbaum.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1993). Scaling Performance Assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 20,* 187-213.

Yen, W. M. & Ferrara, S. (1997). The Maryland School Performance Assessment: Performance Assessment with psychometric quality suitable for high stake usage. *Educational and Psychological Measurement, 57(1),* 60-84.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

TM035244

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Effect of Content Integration on the Construct Validity of Reading Performance Assessment

Author(s): Shu Jing Yen, Nancy Bene, Hoi Hoang Hoang

Corporate Source: paper presented at the 2000 annual meeting of the National Council on Measurement in Education

Publication Date: April, 2000

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, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic 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.

Sign here, please

Signature: Shu Jing Yen	Printed Name/Position/Title: SHU JING YEN, Psychometrician
Organization/Address: Maryland State Dept. of Education	Telephone: 4107670042 / FAX: 4103330052
	E-Mail Address: sjingyen@msde.state.md.us / Date: Dec. 18, 2000

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM 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 that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

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

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)