Validating Standards-Referenced Science Assessments

CSE Technical Report No. 529

Bokhee Yoon
New Standards
Office of the President, University of California

Michael J. Young
New Standards
CRESST/Learning Research and Development Center,
University of Pittsburgh

October 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences
Lauren B. Resnick and Michael J. Young, Project Directors, CRESST / University of Pittsburgh

Copyright © 2000 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR / Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions and policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

VALIDATING STANDARDS-REFERENCED SCIENCE ASSESSMENTS

Bokhee Yoon[1]
New Standards
Office of the President, University of California

Michael J. Young
New Standards
CRESST/Learning Research and Development Center, University of Pittsburgh

Abstract

As standards with accompanying assessments are being proposed and developed in various states and large districts as instruments for raising academic achievement, the validity of the standards-referenced assessments in shaping educational reform demands attention. In this paper, we examine the construct validity of the New Standards middle school Science Reference Examination focusing on evidence related to the internal and external structure of the assessment, the reliability of the assessment scores, and generalizability of the assessment results. The data were taken from the field test of spring 1998. Results related to the internal structure of the assessment suggest that although the assessment tasks measured a single common factor, this did not detract from the usefulness of scientific thinking or science concept subscores for instructional purposes. With respect to the external structure of the assessment, moderate correlations between the New Standards total scores, and the Stanford Achievement Test (9th edition) and the Otis-Lennon School Aptitude Test (7th edition) scores provided evidence that the scores from these assessments rank student performance in similar ways. However, these correlations do not indicate that the assessments are measuring the same construct. For evidence for the reliability of the assessment scores and decisions based on them, the results of the generalizability studies imply that reader variance could be made negligible by training readers with well-defined scoring rubrics. The high rates of decision consistency and accuracy at different total score cutpoints provide evidence that the New Standards Science Reference Examination could be used reliably to classify student performance on the basis of a total test score. For subscores, providing one cutpoint with a reference point to meet the standards would be instructionally informative.

Standards with accompanying assessments are being proposed as instruments for raising academic achievement (e.g., Resnick & Resnick, 1991; Shepard, 1995). The argument, broadly stated, is that if teachers and students know clearly what kinds of

[1] Authors are listed in alphabetical order.

learning are expected, they can direct their teaching and learning energies in a targeted way to meeting standards that will matter in their lives. The National Science Foundation systemic initiatives and other state-level, district-level, and university partnership efforts at reforming science education want and need assessments of student achievement that reflect these common goals for science education for science systemic reform to be truly viable.

The New Standards Science Reference Examinations are currently being developed to be systematically referenced to the New Standards *Performance Standards* for science (New Standards, 1997). The *Performance Standards,* which are based on the emergent national consensus content standards in science, offer a succinct and manageable set of specifications of the knowledge and skills that schools and students should be held responsible for. To the statements of the content ("what students should know and be able to do") derived from national content standards in science, New Standards has added examples of student work that indicate the kinds of evidence one might look for to see if a student has "met the standard."

This new use of assessments as a legitimate target of instruction and learning and the promise of positive consequences of standards-referenced assessments in shaping educational reform demand attention both to the validity of the assessments and their relation to the instructional program and to the ways in which the assessments are used to create opportunities for teacher and student learning. This comprehensive view of validity integrates the traditional considerations of content and criterion with the importance of consequences into a construct framework for examining score meaning and use (Messick, 1995).

Multiple sets of criteria have been suggested as frameworks for gathering validity evidence. Messick (1995) has proposed that general validity standards for aspects of construct validity should address content, substantive, structural, external, generalizability, and consequential aspects of construct validity. Focusing on criteria tailored to the use of performance assessments, Linn, Baker, and Dunbar (1991) proposed content quality, content coverage, cognitive complexity, meaningfulness, cost and efficiency, transfer and generalizability, fairness, and consequences, while Nitko (1996) has suggested that validity evidence should be gathered to address the content, substantive, internal structure, external structure, reliability, generalizability, consequential, and practical aspects of construct validity. It is important to note that in addition to the traditional aspects of validity, these

frameworks emphasize the consequential or value implications of score interpretation and use, as well as practical aspects such as the instructional features of the assessment.

The distinguishing features of these frameworks are:

- The importance of going beyond content representativeness, and examining the quality, relevance, and the types of thinking skills and processes required by assessment content (substantive and cognitive complexity);
- The gathering of evidence to examine the relationship among assessment tasks and assessment parts (internal structure) and the relationship of assessment scores to other factors (external structure);
- A view that integrates the accuracy and consistency of the assessment scores over time, assessors, and content domain (test reliability) with the generalizability of the assessment results over different types of people, or under different conditions (generalizability);
- An emphasis on the consequential or value implications of score interpretation and use, as well as the practical aspects such as the cost, practicality, and instructional features of the assessment.

In this paper, we examine the construct validity of the New Standards middle school Science Reference Examination focusing on evidence related to the internal and external structure of the assessment, the reliability of the assessment scores, and the generalizability of the assessment results. Evidence related to the assessment content coverage, the types of thinking skills required by the assessment, the inferential links from assessment tasks to the science standards, and the instructional features of the assessment are presented in New Standards (1998) and in Martin and Stage (1999).

Method

The New Standards Science Reference Examinations

The Science Reference Examinations are designed to be systematically referenced to the New Standards *Performance Standards* for science that were targeted for students at the end of fourth and eighth, and tenth grades. The seven standards are Physical Sciences Concepts, Life Sciences Concepts, Earth and Space Sciences Concepts, Scientific Connections and Applications, Scientific Thinking, Scientific Tools and Technologies, and Scientific Communication. Items have been developed to measure discipline-specific standards in Physical Science, Life Science, and Earth

and Space Science content areas. Items have also been cross-referenced to standards that measure a student's ability to construct knowledge and use it in scientific ways.

For purposes of reporting, examination items are classified into three scientific thinking clusters in Conceptual Understanding and Applications, Design and Acquisition, and Evidence and Analysis and into three science concepts in Physical Science, Life Science, and Earth and Space Science content areas.

The examinations were field tested to students in Grades 4, 8, and 10 in a number of jurisdictions throughout the nation in spring 1998. The examinations were administered over three class periods (e.g., three days). The first sitting consisted of a total of 24 multiple-choice and short and long constructed response tasks. Three forms (A, B, and C) were field tested for the first sitting. The second and third sittings required students to answer constructed response tasks based on the results of "hands-on," kit-based science investigations. A single form (A) of these kit-based performance tasks was field tested for the second and third sittings.

In this study, we focused on the Form A data[2] in middle school only. Table 1 shows the configuration of the Form A Science Field Test.

Table 1

Science Middle School Field Test Exam Configuration (Form A)

	Number of items					Max. score points
	Total	MC	CR	EV	Perf.	
Science concepts						
Physical Science	6	3	3	0	0	12
Life Science	25	3	3	4	15	59
Earth and Space Science	11	2	2	4	3	25
Scientific thinking						
Concepts and Application	17	6	5	1	5	36
Design and Acquisition	12	1	2	5	4	25
Analysis and Evidence	13	1	1	2	9	32
Total	42	8	8	8	18	93

Note. MC refers to multiple-choice items; CR refers to short constructed response items; EV refers to constructed response items that measure evidence pieces; Perf. refers to long performance tasks.

[2] Form A is one of two forms that New Standards will be developing at each level of the examination.

Data and Analysis

The construct validity of the Science Reference Examinations was examined using Form A in middle school. In addition to the Science Reference Examination, the Science portion of the Stanford Achievement Test, 9th Edition, (Stanford 9) and the Otis-Lennon School Abilities Test, 7th Edition, (OLSAT/7) were also administered in the field test of spring 1998.

The sample size used for the analyses of middle school Form A was 450. After cleaning and merging the data based on the retrieved documents, the number of students who responded to at least one item in a sitting (a class period) for Form A in middle school was 747; however, we selected only students who responded to at least 80% of questions during each of the sittings due to a large portion of blank responses on questions. The large non-response rate in constructed response items or performance tasks is not uncommon (Jakwerth, Stancavage, & Reed, 1998).

The following sets of analyses were performed to gather evidence of the internal and external structure of the assessment, the reliability of the assessment scores, and the generalizability of the assessment results.

Internal structure. The correlations between assessment tasks and scores of the Science Reference Examination were examined. The underlying dimensional structure of the assessment tasks was examined using a confirmatory factor analysis (CFA) for the total test, and second-order confirmatory factor analysis with clusters in Concept and Application, Design and Acquisition, and Evidence and Analysis or with content areas in Physical Science, Life Science, and Earth and Space Science as first-order factors and the total test as a second-order factor. Confirmatory factor analyses were performed using the software *M-Plus,* which allows for CFA with categorical variables (Muthén & Muthén, 1998).

External structure. The relationship of students' Science Reference Examination scores to their scores on the Stanford 9 Science and OLSAT/7 tests was examined. Convergent and discriminant evidence was also examined using multitrait-multimethod comparisons.

Reliability and generalizability. The accuracy and consistency of the assessment scores over assessors, clusters, content areas, and item types were examined using the framework of generalizability theory (Brennan, 1983; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). The consistency and

accuracy of classification decisions based on total score cutpoints were examined using the method outlined by Livingston and Lewis (1995).

Results and Discussion

Internal Structure

Correlations. In addition to a total score, New Standards is planning to report content-related scores for Physical Science (PS), Life Science (LS), and Earth and Space Science (ES), as well as cluster scores for Concept and Application (CA), Design and Acquisition (DA), and Evidence and Analysis (EA). New Standards content specialists assigned items to a cluster and a content exclusively. Once items were assembled as a form, it was important to examine how performance on a test item was related to performance on the total test score or subscores, and whether the tasks comprising the reported scores could be empirically identified as measuring a common factor or multiple factors. If the subscores measure something different from each other, then the average item correlations among the items in the same cluster (or content) would be higher than the item correlations with other clusters or content.

The inter-item correlations were summarized by content and cluster (see Table 2). Since the item responses were either binary or categorical, polychoric correlation coefficients were computed (Bollen, 1989). The average item correlations within a cluster (e.g., CA) were similar across clusters even though the average correlation among CA items (r = .19) is slightly lower than the average correlations among DA (r = .22) or EA items (r = .25). The average correlations between clusters were also similar with those within a cluster. For example, the average item correlation for CA items was .19, and the average item correlations of CA items with DA items or EA items were .21 and .22 respectively. The pattern of item correlations in content was similar to the pattern of correlations for cluster.

Table 3 presents the average correlations between a task (or an item) and its composite scores (i.e., total, cluster, and content scores). Note that the average correlations above and below diagonals are not symmetric. This is because the correlations were computed between item scores and composite scores. The correlations on the diagonal in Table 3 were expected to be higher than the correlations on the off-diagonal. However, the correlations both within and between clusters or contents were similar. The average item correlations ranged from .31 to

Table 2

Means of Item Correlations for Cluster and Content

	Cluster				Content		
	CA	DA	EA		PS	LS	ES
CA	.19			PS	.17		
DA	.21	.22		LA	.20	.22	
EA	.22	.24	.25	ES	.20	.23	.23

Note. CA = Conceptual Understanding and Application; DA = Design and Acquisition; EA = Evidence and Analysis; PS = Physical Science; LS = Life Science; ES = Earth and Space Science.

Table 3

Means of Item-Cluster, Item-Content, and Item-Total Scores Correlations

Item set	Cluster scores			Item set	Content scores		
	CA	DA	EA		PS	LS	ES
CA	.35	.31	.33	PS	.28	.32	.31
DA	.34	.34	.34	LA	.27	.37	.33
EA	.36	.33	.36	ES	.31	.37	.36

Note. Mean correlations of all items with a total score was .38. CA = Conceptual Understanding and Application; DA = Design and Acquisition; EA = Evidence and Analysis; PS = Physical Science; LS = Life Science; ES = Earth and Space Science.

.36 across clusters and from .27 to .37 across contents. The average item correlation with a total score (i.e., point-biserials) was .38, which indicated that all items were related moderately with the New Standards total test score. These results show that the test items correlated not only with their own construct but also with other constructs at the same level, suggesting that the items measure the same overall construct.

The relationships among content, cluster, and item type scores are summarized in Table 4. There were four different item types: multiple-choice items (MC), short constructed response items (CR), evidence tasks (EV), and performance tasks. Performance tasks were divided into P1 (second sitting or second testing period) and P2 (third setting or third testing period). The correlations among composite scores ranged from .62 to .76 in content scores, from .69 to .75 in cluster scores, and from .26 to .63 in item type composite scores. Interestingly, the correlations among

Table 4

Correlations among Content, Cluster, and Item Type Scores

	Content			Cluster			Item type			
	PS	LS	ES	CA	DA	EA	MC	EV	CR	P1
PS										
LS	.64									
ES	.62	.76								
CA	.80	.87	.81							
DA	.63	.84	.79	.70						
EA	.61	.88	.80	.75	.69					
MC	.56	.43	.43	.56	.35	.40				
EV	.51	.74	.74	.67	.83	.62	.26			
CR	.81	.81	.77	.88	.75	.74	.36	.63		
P1	.49	.84	.62	.73	.60	.81	.35	.49	.59	
P2	.53	.77	.77	.67	.73	.81	.30	.52	.61	.58

Note. PS = Physical Science; LS = Life Science; ES = Earth and Space Science; CA = Conceptual Understanding and Application; DA = Design and Acquisition; EA = Evidence and Analysis; MC = multiple-choice items; CR = short constructed response items; EV refers to constructed response items that measure evidence pieces; P1 = Performance task, second testing period; P2 = Performance task, third testing period.

open-ended questions (i.e., CR, EV, P1, and P2) were higher than the correlations with the MC composite score. For example, the correlations among open-ended questions ranged from .49 to .63, whereas the correlations between the MC score and open-ended questions ranged from .26 to .36.

Confirmatory factor analysis. The dimensionality of test items was further examined by performing a set of confirmatory factor analyses (CFAs) using the software *M-Plus*, which allows for CFA with categorical variables (Muthén & Muthén, 1998), since all item response categories in the New Standards Science Reference Examination are either binary or polytomous. Two sets of analyses were performed, and the results are displayed in Table 5.

First, two one-factor CFAs were performed using all the items (full) and after deleting four items (i.e., fixing parameters to zero: reduced) in the science examination. The four fixed item parameters were items with low point-biserials in item analyses and insignificant loadings in CFAs. The second set of analyses used second-order factor analyses where the set of all items was used as an underlying construct (second-order factor) and the three clusters (or contents) were modeled as the first-order factors. These analyses were performed to examine whether the test

Table 5

Confirmatory Factor Analysis and Second-Order Factor Analysis

			Second-order CFA			
	One-factor CFA		Cluster		Content	
	Full	Reduced	Full	Reduced	Full	Reduced
Chi-square	484.098	662.172	459.157	623.211	480.16	657.00
DF	172	214	172	214	172	214
P-value	.000	.000	.000	.000	.000	.000
Factor loading						
F1 (CA / PS)			.968	.968	.911	.921
F2 (DA / LS)			.879	.880	.920	.931
F3 (EA / ES)			.947	.945	.930	.931

Note. CFA = Confirmatory factor analysis; CA = Conceptual Understanding and Application; PS = Physical Science; DA = Design and Acquisition; LS = Life Science; EA = Evidence and Analysis; ES = Earth and Space Science.

items measured three distinct factors in addition to measuring one common factor. The second-order CFAs were also performed on the data including all items and after fixing four item parameters to zero. In these models measurement errors of observed variables (items) were not allowed to correlate with each other.

Although the one-factor CFA models did not fit the data ($\chi^2 = 484.098$, $df = 172$, $p < .000$ with full; $\chi^2 = 662.172$, $df = 214$, $p < .000$ with reduced), the difference of the chi-squares between two models was significant ($\chi^2 = 178.07$, $df = 42$, $p < .0001$) indicating that the reduced model significantly improved on the full model. In addition, all factor loadings, which provide the direct effects of the factor on the observed variables, were significant.

Similar results were shown in the second-order CFA models. Although the models didn't fit the data, the model was improved when the insignificant parameters were fixed to zero in both content ($\chi^2 = 176.838$, $df = 42$, $p < .0001$) and cluster ($\chi^2 = 164.054$, $df = 42$, $p < .0001$). All factor loadings were significant here as well. In the second-order CFA models, the loadings of the second-order factor to the first-order factors were very high (ranging from .88 to .97 across clusters; from .92 to .93 across contents). The results seemed to indicate that the test items measure one common factor, rather than three factors. However, the dependencies among items (e.g., items in a performance task) need to be further explored.

External Structure

Relationship among New Standards Science, Stanford 9, and OLSAT/7 scores. The external structure of the middle school Science Reference Examination was studied by examining the relationship among New Standards Science Reference Examination, Stanford 9, and OLSAT/7 scores. The Science Reference Examination was designed to assess conceptual understanding, scientific inquiry, and problem solving, and included two class periods of hands-on performance tasks in addition to multiple-choice items and short constructed response tasks. The Science subtest of the Stanford 9 consists of 40 multiple-choice items that assess general science achievement, and the OLSAT/7 consists of 72 items and assesses verbal and nonverbal aptitude.

Given the designs of these tests, it was hypothesized that the correlation between the Science Reference Examination and Stanford 9 would be moderate and higher than the correlation of the Reference Examination with OLSAT/7, but lower than the correlation between Stanford 9 and OLSAT/7. That is, the correlation between two tests measuring science would be higher than the correlation between a science test and an aptitude test, but lower than the correlation between the two multiple-choice tests. Table 6 presents the relationship among Science Reference Examination (NS Science), Stanford 9, and OLSAT/7 scores.

As expected, the correlation between the NS Science and Stanford 9 scores (r = .63) was lower than the correlation between Stanford 9 and OLSAT/7 scores (r = .74), but similar to the correlation between NS Science and OLSAT/7 scores (r = .60). The results indicate that these three scores seem to rank students similarly to a certain extent. However, the moderate correlations do not necessarily indicate that these exams measure the same construct.

Convergent and discriminant validity. The external structure of the NS Science Reference Examination was further examined using a multitrait-

Table 6

Relationships Among NS Science, Stanford 9, and OLSAT/7 Scores

	Stanford 9	OLSAT/7	*N*	Mean	*SD*
Stanford 9			372	26.10	7.22
OLSAT/7	.74		405	49.73	14.98
NS Science	.63	.60	450	53.06	12.33

Note. NS = New Standards.

multimethod matrix. Multitrait-multimethod validity is an aspect of construct validity that was developed by Campbell and Fiske (1959). This method is used when two or more traits are being measured by two or more methods. The multitrait-multimethod validity index provides reliability coefficients, convergent validity coefficients, and discriminant validity coefficients. Reliability coefficients are the estimated reliability of each trait. Convergent validity coefficients (monotrait-multimethod) are correlations between measures of the same construct using different measurement methods. Discriminant validity coefficients (heterotrait-heteromethod) are correlations between measures of different constructs using the same method of measurement or correlations between different constructs using different measurement methods. Ideally, discriminant validity coefficients should be substantially lower than reliability or convergent validity coefficients (Crocker & Algina, 1986). The convergent validity values for a trait should also exceed correlations for that trait in the same method, but different traits (heterotrait-monomethod).

For these data, subscores of the Stanford 9 Science subtest and the science concept subscores of the Science Reference Examination were used to measure student achievement in Physical Science (PS), Life Science (LS), and Earth and Space Science (ES). Table 7 presents the results.

Table 7

Convergent and Discriminant Validity: Multitrait-Multimethod Matrix

	NS science			Stanford 9			No. of items	Max. score
	PS	LS	ES	PS	LS	ES		
NS science								
PS	**.53**						6	12
LS	.64	**.83**					25	59
ES	.62	.76	**.72**				11	25
Stanford 9								
CA	.37	.53	.52	**.71**			14	14
DA	.40	.55	.53	.70	**.64**		14	14
EA	.40	.51	.52	.71	.67	**.67**	12	12

Note. Values on the diagonal are reliability coefficients; underlined values are convergent validity coefficients; off-diagonal values are discriminant validity coefficients. NS = New Standards. PS = Physical Science; LS = Life Science; = ES = Earth and Space Science.

All traits (PS, LS, and ES) in the same method (heterotrait-monomethod) correlated higher than those in different methods (monotrait-multimethod or heterotrait-heteromethod). For example, the correlations among traits in NS Science and Stanford 9 ranged from .62 to .76 and from .67 to .71, respectively, whereas both convergent and discriminate values ranged from .37 to .55.

Furthermore, the convergent values did not exceed the discriminant values (off-diagonal values). This lack of convergence between the Science Reference Examination and Stanford 9 can be explained as methods that are not measuring the same construct. That is, the differences between the two tests in terms of kinds of tasks are so great that the tests are measuring different things. The lack of convergence could be due also to the differences in the content validity of categorizing items in both tests into three content areas (PS, LS, and ES). Nevertheless, from the viewpoint of having alternative measures of the same content, high values for the validity coefficients were desirable. However, from the position of constructed response tasks measuring different aspects of content knowledge than those measured by multiple-choice items, high correlations between the two measures would have been problematic.

Generalizability

Two groups of generalizability studies were performed to examine the sources of variation accounting for students' Science Reference Examination scores. The first group of analyses employed a fully crossed person-by-item-by-reader *(p x i x r)* design. Samples of student papers (n_p = 50) were scored across all items (n_i = 42) by two readers (n_r = 2). These analyses were used to examine the main effects for readers, and the reader-by-person and reader-by-item interactions.

The second group of analyses used person crossed with item-nested-in-type *(p x i:t)* designs. Separate sets of analyses were performed by nesting items within (a) cluster type, (b) content area, and (c) item type. Nested designs were used since the items could be classified into one of several mutually exclusive cluster, content, or task "types." Since these categories represented the only levels of the type facets that will be used in the Science Reference Examination, type was treated as a fixed facet. As in Shavelson and Webb (1991), these data were analyzed first as *p x i: t, t random* designs. The variance components for the random part of the design, namely, *p*, *i*, and *pi,e,* were obtained, and the person term was modified by adding the person-by-type interaction term, which was divided by the number of levels of the fixed facet.

The separate *p x i d*esigns for each level of the fixed facet *t* were also calculated. The data for these analyses consisted of n_p = 450 students at middle school.

***p x i x r* design.** Table 8 shows the variance components and percentages of total variance explained by the variance components. The variance component for person was small and accounted for only 7% of the total variance, indicating that students did not systematically differ in their Science Reference Examination scores.

Even smaller were the variance components associated with the reader facet. The variance components for reader, reader-by-item, and reader-by-person were negative-valued and small in absolute value. Negative values for variance components occur due to sampling error or when the underlying measurement model has been misspecified. Following conventional practice, these negative variance components were set equal to zero (see Shavelson & Webb, 1991, pp. 36-38). The magnitudes of the reader-related variance components indicate that the readers were well calibrated and generally consistent in their scoring across persons and items.

The largest sources of variability were seen in the variance components for item and person-by-item. The percentages of total variability for these components were 47% and 31% respectively. The magnitudes of these components suggest that

Table 8

Generalizability Studies for Person x Reader x Item Designs

	N	Estimate	%
Person *(p)*	50	.0491	7
Reader *(r)*	2	0	0
Item *(i)*	41	.3402	47
pi		.2233	31
pr		.0001	0
ri		0	0
pir, e		.1072	15
	Standard errors		Coefficients
n_r = 2			
Relative	.08		.88
Absolute	.12		.77
n_r = 1			
Relative	.09		.86
Absolute	.13		.76

overall student scores differed from one item to another, and that the relative standing of students differed from one item to another. That is, students who scored high on one item did not necessarily score high on another item. The large item and person-by-item variances also imply the importance of balancing items between forms for fairness in comparing scores between forms.

The residual variance components (*pir, e*) showed a substantial (although smaller) percentage (15%) of total variance due to the three-way interaction of persons, items, and readers and/or other unmeasured sources of variability. The standard errors and generalizability (Rho) and dependability (Phi) coefficients based on the variance components are also shown in Table 8. These variance components were used in two decision studies to examine the effect of using one or two readers to score all the items taken by each student, and include standard errors and coefficients for both relative and absolute decisions.

As expected, the standard errors for absolute decisions were larger than those for relative decisions, due to the presence of additional variance components in calculating the absolute standard errors, and led to the higher values of the generalizability coefficients for relative decisions over absolute decisions. Going from two readers (Rho = .88; Phi = .77) to a single reader (Rho = .86; Phi = .76) did not affect the generalizability very much: The coefficients did not exceed .02. Based on these data, using a single reader would be acceptable but the coefficient for absolute decisions is a concern.

***p x (i:t), t fixed* designs.** The three sets of generalizability studies were performed by nesting items within (a) cluster type, (b) content area, and (c) item type. In addition, person-by-item designs were performed for each level of the fixed facet to further examine the sources of variability across levels of the fixed facet. The results are shown in Tables 9, 10, and 11.

In the mixed designs, similar results were shown in the estimated variance components either averaging over across clusters or averaging over across contents (11% of the total variance for person variance component; 36% and 37% for item variance components; 52% and 54% for the residual components). The item variance component in the mixed design was large (37% percentage of total variance in cluster; 36% in content), but smaller than the item variance component in Table 8 (47% of the total variance). The variance components for persons accounted for the same proportion of variance in both cluster (11%) and content (11%). The residual

Table 9

Total and Cluster Scores: *Person x (Item:Cluster Type)* Design

	Mixed *p x (i:t)* design, *t* fixed[a]		Separate *p x i* designs					
			Concepts & Application		Design & Acquisition		Evidence & Analysis	
	Estimate	%[b]	Estimate	%	Estimate	%	Estimate	%
Persons (p)	.0771	11	.0797	9	.0781	11	.0789	15
Items (i)	.2700	37	.3954	46	.2786	38	.0952	28
pi, e	.3788	52	.3923	45	.3797	52	.3600	67
	SE	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*	Coef.
Relative	.10	.90	.15	.78	.18	.71	.17	.74
Absolute	.12	.83	.22	.63	.23	.59	.19	.69

Note. n_p = 450; n_i = 42 for a total, 17 for Concepts and Application, 12 for Design and Acquisition, and 13 for Evidence and Analysis.

[a] Averaging over levels of fixed facet (cluster).

[b] Percentage of total variance.

Table 10

Total and Content Scores: *Person x (Item:Content Type)* Design

	Mixed *p x (i:t)* design, *t* fixed[a]		Separate *p x i* designs					
			Physical Science		Life Science		Earth & Space Science	
	Estimate	%[b]	Estimate	%	Estimate	%	Estimate	%
Persons (p)	.0766	11	.0792	12	.0792	10	.0961	6
Items (i)	.2508	36	.1336	21	.2989	41	.1938	52
pi, e	.3779	54	.4246	67	.3564	49	.2259	42
	SE	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*	Coef.
Relative	.10	.89	.27	.53	.12	.83	.19	.72
Absolute	.12	.84	.31	.46	.16	.73	.23	.64

Note. n_p = 450; n_i = 42 for a total, 6 for Physical Science, 25 for Life Science, and 11 for Earth and Space Science.

[a] Averaging over levels of fixed facet (content).

[b] Percentage of total variance.

Table 11

Total and Item Type Scores: *Person x (Item:Item Type)* Design

	Mixed *p x (i:t)* design, *t* fixed[a]		Separate *p x i* designs							
			Multiple choice		Constructed response		Evidence tasks		Performance tasks	
	Estimate	%[b]	Estimate	%	Estimate	%	Estimate	%	Estimate	%
Persons (p)	.0761	14	.0190	7	.2559	27	.0864	16	.0843	15
Items (i)	.1184	22	.0454	18	.1313	14	.0549	10	.1692	30
pi, e	.3515	64	.1895	75	.5559	59	.3992	74	.3145	55
	SE	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*	Coef.
Relative	.09	.90	.15	.45	.26	.79	.22	.63	.13	.83
Absolute	.11	.87	.17	.39	.29	.75	.24	.60	.16	.76

Note. n_p = 450; n_i = 42 for a total, 8 for multiple-choice tasks , 8 for constructed response tasks, 8 for evidence tasks, and 18 for performance tasks.

[a] Averaging over levels of fixed facet (item type).

[b] Percentage of total variance.

variance component (pi, e) accounted for most of the variation in the total science score (52% in cluster; 54% in content) and was substantially larger than that in the *p x i x r* design (15%). The coefficients for both relative and absolute decisions were higher compared to those in the *p x i x r* design since the variance component for person-by-item interaction was confounded in the residual term.

When *p x i* designs were performed for each cluster and each content separately, the proportions of item variance component varied, ranging from 18% to 46% across clusters, and from 21% to 52% across contents. Similarly, the coefficients for relative and absolute decisions also varied across content or cluster scores. Generalizability coefficients ranged from .71 to .78 for relative decisions and ranged from .59 to .69 for absolute decisions across clusters. Generalizability coefficients ranged from .53 to .83 for relative decisions and ranged from .46 to .73 for absolute decisions across contents. The differences in the number of items and the composition of item types in each cluster (e.g., CA) and content (e.g., PS) should be taken into account for interpreting the results.

In the mixed design of nesting items within item type, the variance components of the item were smaller (22% of the total variance) than those in the design of nesting items within content or cluster. Similar results were shown in *p x i* designs as well. This can be expected since scoring rubrics among items are more similar within a task format than within a cluster or a content.

In *p x i* designs for each task format, the variance component for persons in multiple-choice items was small (7% of the total variance) compared to those in constructed response items (27% of the total variance), evidence tasks (16% of the total variance), and performance tasks (15% of the total variance). Students' scores in multiple-choice items varied the least compared to other item-type scores, whereas the greatest variation across item difficulties was shown in performance tasks (30% of the total variance).

Accuracy and Consistency of Cutpoint Decisions

Sets of analyses were performed to estimate the accuracy and consistency of decisions based on different total score cutpoints. The *accuracy* of the decisions is the extent to which they would agree with the decisions that would be made if each student could somehow be tested with all possible forms of the examination. The *consistency* of the decisions is the extent to which they would agree with the

decisions that would have been made if the students had taken a different form of the Science Reference Examination, equal in difficulty and covering the same content as the form they actually took.

Correct classifications occur when the decision made on the basis of the all-forms-average (or true score) agrees with the decision made on the basis of the form actually taken. A *false positive* misclassification occurs when a student who is actually below the cutpoint on the basis of his or her all-forms-average is classified incorrectly as being above the cutpoint. Similarly, a *false negative* classification occurs when a student whose all-forms-average is above the cutpoint is classified as being below the cutpoint. Consistent classifications occur when the two forms agree on the classification of a student as either being above or below the cutpoint; inconsistent classifications occur when the decisions made by the forms differ.

Estimates of decision accuracy and consistency were made for cutpoints at the first quartile, the median, and the third quartile of the Science Reference Examination distributions for total, science concepts, and scientific thinking scores. That is, students were classified on the basis of their scores as being above or below one of these cutpoints. Then an analysis was performed to estimate the accuracy of that classification and the consistency with which it could be made. These analyses[3] made use of the techniques described in Livingston and Lewis (1995), as implemented by Young and Yoon (1998), and used as reliability estimates the generalizability coefficients for relative decisions from the *p x i:t* design in Tables 9 and 10. Table 12 presents the percents of consistent and accurate classifications, as well as the false positive and false negative rates.

The percentage of consistent classifications ranged from 86% to 89% for the total score and 69% to 86% for subscores across the cutpoints, while the percentage of accurate classifications ranged from 95% to 96% for the total score and 89% to 95% for subscores. The consistent classification rates tended to be lower at the median cutpoint than at the cutpoints at the quartiles, regardless of score. In other words, inconsistent classifications occurred more frequently at the median cutpoint when the agreements were based on two forms. False positive rates ranged from 4% to 6% for the total score and 4% to 13% for subscores across cutpoints, and false negative rates ranged from 3% to 5% for the total score and 3% to 12% for subscores.

[3] For an alternative approach see Rogosa (1999a, 1999b).

Table 12

Consistency and Accuracy of Decisions Based on Total Score Cutpoints

Cutpoint location	% Consistent classifications	% Accurate classifications	% False positives	% False negatives
Total score				
Q1	89	96	4	4
Median	86	95	5	5
Q3	88	95	6	3
Physical Science				
Q1	74	89	7	12
Median	69	89	13	10
Q3	79	91	9	6
Life Science				
Q1	86	94	4	7
Median	82	94	7	6
Q3	86	95	5	5
Earth and Space Science				
Q1	80	92	7	8
Median	76	91	11	7
Q3	82	93	8	5
Concepts & Application				
Q1	83	93	5	7
Median	79	92	7	8
Q3	80	90	12	3
Design and Acquisition				
Q1	79	93	8	7
Median	76	91	11	6
Q3	80	90	12	3
Evidence & Analysis				
Q1	80	91	4	10
Median	77	92	8	8
Q3	80	92	8	6

Note. Q1 = quartile 1; Q3 = quartile 3.

The total score produced the highest consistency rates (≥ 86%) and accuracy rates (≥ 95%), whereas the score for Physical Science produced the lowest consistencies (69-74%) and accuracies (89-91%). These results clearly follow from the fact that the total score was based on 42 tasks yielding a reliability of .90, whereas the score for Physical Science was based on 6 tasks with a reliability of .53.

The consistency and accuracy of classifying cutpoints of the Science Reference Examination total score distributions indicate that the total score can be used to make decisions of high accuracy and consistency across a range of cutpoints. For subscores, providing one cutpoint with a reference point to meet the standards would be informative; however, providing multiple cutpoints for subscores would be problematic.

Summary

This paper has focused on validity evidence related to the internal and external structure, the reliability, and the generalizability of the New Standards Science Reference Examination. The evidence indicates the following:

- Evidence on internal structure of the New Standards Science Reference Examination suggests that middle grade Science Field Test items measure a single common factor, rather than three *statistically distinct* factors related to either the scientific thinking clusters or science concept areas. Subscores would provide meaningful interpretations for instructional purposes and would be *logically distinct* if items in a subscore satisfy the content and substantive aspect of validity.

- As predicted, the correlation between the New Standards Science Reference Examination and the Stanford 9 Science test measuring science was moderate and lower than the correlation between the Stanford 9 and the OLSAT/7 aptitude test. However, it was approximately the same as the correlation of the New Standards Science Reference Examination with the OLSAT/7. The moderate correlations of the New Standards total scores with the Stanford 9 and the OLSAT/7 scores imply that these scores rank student performance in similar ways. However, this doesn't mean that they measure the same thing. The relationships of these scores should be further examined with multiple groups with different instructional coverage and interpreted substantively with a caution. Furthermore, the lack of convergence between the Science Reference Examination and the Stanford 9 in measuring PS, LS, and ES did not support the possibility that they were two alternative methods measuring the same constructs.

- Consistent with other generalizability studies (Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Ruiz-Primo, & Wiley, 1999; Webb, Schlackman, & Sugrue, 1999), the person-by-item variance component accounted for the largest percentage of the total variability, indicating that student performance on the assessment will vary depending on the tasks administered. The negligible values of the variance components associated with rater, person-by-rater, and rater-by-item indicate the raters were well calibrated and generally consistent in their scoring across persons and items.

- The consistency (86%+) and accuracy (95%+) of classifying cutpoints of the Science Reference Examination total score distributions were reasonably high, indicating that the total score can be used to make decisions of high accuracy and consistency across a range of cutpoints. For subscores, providing one cutpoint with a reference point to meet the standards would be instructionally informative; however, providing multiple cutpoints for subscores would be problematic.

References

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 546-553.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Jakwerth, P. R., Stancavage, F. B., & Reed, E. D. (1998). *NVS NAEP validity studies: An investigation of why students do not respond to questions.* Palo Alto, CA: American Institutes for Research.

Lane, S., Liu, M. Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33,* 71-92.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-23.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Martin, M. B., & Stage, E. K. (1999, April). *Good science: Substantive issues in validating standards-referenced science assessments.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus The comprehensive modeling program for applied researchers: User's guide.* Los Angeles, CA: Muthén & Muthén.

New Standards. (1997). *Performance standards (*Vols. 1, 2, 3). Rochester, NY: National Center on Education and the Economy.

New Standards. (1998). *New Standards Science Reference Exam release package.* Oakland, CA: Regents of the University of California.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Resnick, L. B., & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Rogosa, D. (1999, July). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations* (CSE Tech. Rep. No. 509). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Rogosa, D. (1999, August). *Accuracy of year 1, year 2 comparisons using individual percentile rank scores: Classical test theory calculations* (CSE Tech. Rep. No. 510). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*, 61-71.

Shavelson, R. J., & Webb, N. W. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership, 52*(5), 38-43.

Webb, N., Schlackman, J., & Sugrue, B. (1999, April). *The dependability and interchangeability of assessment methods in science.* Paper presented at the 1999 NCME annual meeting, Montreal.

Young, M. J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (CSE Tech. Rep. No. 479). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).