DOCUMENT RESUME

ED 431 035 TM 029 869

AUTHOR	Perkhounkova, Yelena; Dunbar, Stephen B.
TITLE	Influences of Item Content and Format on the Dimensionality of Tests Combining Multiple-Choice and Open-Response Items: An Application of the Poly-DIMTEST Procedure.
PUB DATE	1999-04-00
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; Constructed Response; Estimation (Mathematics); Grade 7; Grade 8; *Junior High School Students; Junior High Schools; Language Arts; Mathematics Tests; Multiple Choice Tests; Test Construction; *Test Items
IDENTIFIERS	*Dimensionality (Tests); *DIMTEST (Computer Program); Open Ended Questions

ABSTRACT

The DIMTEST statistical procedure was used in a confirmatory manner to explore the dimensionality structures of three kinds of achievement tests: multiple-choice tests, constructed-response tests, and tests combining both formats. The DIMTEST procedure is based on estimating conditional covariances of the responses to the item pairs. The analysis was conducted on the results from the joint administration of the Constructed-Response Supplement to the Iowa Tests of Basic Skills and Form M of the Iowa Tests of Basic Skills, using samples of 952 seventh graders and 882 eighth graders for the language tests, and 889 seventh graders and 918 eighth graders for the mathematics tests. Three potential sources of dimensional distinctness among items in the tests, caused by item format, item content, and item location, were investigated. The results differed for language and mathematics achievement tests, in part because the tests were dissimilar with regard to content composition and the number of items in the whole tests as well as subtests that were analyzed. The Poly-DIMTEST analysis presented provides evidence of dimensional heterogeneity of tests intended to provide assessments that are balanced with respect to the content composition of a targeted construct. The analysis also suggests that combining items of different formats may introduce additional complexity into the dimensionality structure of the composite test. (Contains 1 figure, 8 tables, and 31 references.) (SLD)

Influences of Item Content and Format on the Dimensionality of Tests Combining Multiple-Choice and Open-Response Items: An Application of the Poly-DIMTEST Procedure

Yelena Perkhounkova and Stephen B. Dunbar

University of Iowa

Paper presented at the 1999 Annual Meeting of
the American Educational Research Association, Montreal, Canada

Influences of Item Content and Format on the Dimensionality of Tests Combining Multiple-Choice and Open-Response Items: An Application of the Poly-DIMTEST procedure

With the current growth of interest in open-response and other performance-type assessments in education, it has become more critical to understand similarities and dissimilarities of the results from tests of different types.

A great deal of research has been devoted to comparing scores from multiple-choice tests and open-response tests. Some researchers sought the answer to the question of whether the scores obtained on tests of different formats designed to measure the same construct could be considered as indicators of one construct or of different constructs (Ackerman & Smith, 1988; Breland & Gaynor, 1979; Bridgeman, 1992; Hoover & Bray, 1995; Ward, 1982). Others presumed that the changes in test format actually changed the measured construct (Frederiksen, 1984) and sought to reveal differences between constructs measured by the tests of different formats. For instance, Ackerman and Smith (1988) concluded that scores obtained from multiple-choice, open-response, and essay types of writing assessment provided different information, that is, the construct being measured was a function of the format of the test (also see Hogan & Mishler, 1980; Hoover & Bray, 1995). Thissen, Wainer, and Wang (1994) formulated the question as follows "Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests?" Their conclusion was that, although the free-response problems on both the Computer Science and Chemistry Advanced Placement tests assessed something different than the multiple-choice sections of those tests assessed, they, first, predominantly measured the same thing, and second, they were rather poor at measuring something different. Some studies found no format differences (Bennett, Rock, & Wang, 1991; Bridgeman, 1992; Bridgeman & Rock, 1993; Lukhele, Thissen & Wainer, 1994; van den Bergh, 1990; Traub & Fisher, 1977; Ward, 1982). Luckhele, Thissen, and Wainer (1994) reached the discouraging conclusion that "constructed-response items provide less information in more time at greater cost than do multiple-choice items." Results vary greatly across content areas, degree of structure in the format, and purposes of assessment.

Open-response items and multiple-choice items are often combined in one test in an attempt to cover a broader range of assessed skills while maintaining acceptable level of reliability. Test developers face the need to aggregate the scores on such assessments so as to obtain a meaningful summary score for each examinee. Aggregating test scores often raises questions about the dimensionality of the composite in part because unidimensionality is one of the most important prerequisites for using traditional IRT models to scale tests (Wilson & Wang, 1995).

Lord and Novick (1968) defined dimensionality as the total number of abilities required to satisfy the assumption of local independence. A set of items is called locally independent if, for fixed values of the latent traits, the item responses are statistically independent. For a weaker form of local independence to hold (McDonald, 1981), it is sufficient that for fixed values of the traits the item responses are uncorrelated. Stout (1987) introduced the concept of essential dimensionality as the number of abilities required to satisfy the assumption of essential independence, a weak form of local independence. According to Stout (1990), a set of items is called essentially independent with respect to the latent variable vector $\boldsymbol{\theta}$ if, for a given subset of item responses, the average absolute conditional covariances of responses to item pairs approach zero as the length of the subset increases.

To assess essential unidimensionality of a test, Stout (1987) devised a statistical index and formulated a nonparametric statistical procedure, called DIMTEST, based on estimating conditional covariances of the item pairs, where the conditioning variable is an appropriately selected subscore. The idea that underlies the procedure is that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal ability. If the test is unidimensional, then the number-correct score can be substituted for θ when computing the estimates of the conditional item-pair covariances because number-correct is a consistent estimator of the expected number-correct which is a monotone transformation of the θ scale for the unidimensional model (Stout, 1990). If the test is not unidimensional, number-correct can still be used as the conditioning variable because it can be informally considered to be a consistent estimator of $\Theta_{\underline{TT}}$, the unidimensional latent variable "best measured" by the total test number correct score. According to Stout et al. (1996), $\Theta_{\underline{TT}}$ should be viewed as a direction or axis embedded in the multidimensional coordinate system $\boldsymbol{\Theta}$. Similarly, $\Theta_{\underline{C}}$ denotes the unidimensional latent variable best measured by number-correct on the item subset $\underline{C}$. The concept of "best measured" latent variable is described in Zhang and Stout (1996b).

Geometric Representation

According to the geometric representation of multidimensional latent variable models developed by Reckase and McKinley (1991), the vector representing an item lies on a line passing through the origin, where the origin is taken as the population multidimensional trait level mean. The direction of the vector is that in which the item has maximum discrimination and is referred to as the item's direction of best measurement. The relative length of the vector is a function of the magnitude of the item's discrimination. The location of the base is determined by the item's difficulty. Item subsets and the entire test can be represented in a similar manner.

Simple structure, as defined by Stout, et al. (1996), is said to exist for a $\underline{d}$-dimensional test if a $\underline{d}$-dimensional latent coordinate system $\boldsymbol{\Theta} = \{\Theta_{\iota} : 1 \leq \underline{i} \leq \underline{d}\}$ exists such that all items lie along the coordinate axes. Approximate simple structure exists for a test of dimension $\underline{d} \geq \underline{k}$ if a $\underline{k}$-dimensional latent coordinate system $\boldsymbol{\Theta}$ exists within the _d_-dimensional latent space such that all items lie in narrow sectors around the coordinate axes (Zhang & Stout, 1996a). These narrow sectors of items constitute the dominant dimensions of the approximate simple structure.

The graphical representation of a test demonstrating approximate two-dimensional simple structure is illustrated in Figure 1. Here, $\Theta_{\underline{C1}}$, $\Theta_{\underline{C2}}$, and $\Theta_{\underline{TT}}$ denote the unidimensional latent variables best measured by the two subtest scores and the total test score, respectively. For simplicity, all items in the figure are of average difficulty, and items in the same cluster have similar discriminations.

The covariance of two items, one from subtest 1 and another from subtest 2, conditional on $\Theta_{\underline{TT}}$ will be negative; the covariance of two items, both from either subtest 1 or subtest 2, conditional on $\Theta_{\underline{TT}}$ will be positive; and the covariance of two items, both from subtest 1, conditional on $\Theta_{\underline{C1}}$ will be close to 0. This reasoning was generalized to dimensions higher than two by Zhang and Stout (1996b). They suggested that the magnitude of the two-item conditional covariance is a function of the magnitudes of the items' discriminations and the angles between the item vectors and the vector for the conditioning variable, such as $\Theta_{\underline{TT}}$ or $\Theta_{\underline{C1}}$.

The DIMTEST procedure

The DIMTEST procedure is based on estimating conditional covariances of the responses to the item pairs. To perform testing associated with this procedure, Stout, Douglas, Junker, and Roussos (1993) developed the DIMTEST computer program. Li and Stout (1995) have modified the DIMTEST procedure, along with the corresponding program, to accommodate polytomous items. They named this version of the approach Poly-DIMTEST. DIMTEST and Poly-DIMTEST have demonstrated potential for detecting the lack of essential unidimensionality (Li & Stout, 1995; Nandakumar, 1993, 1994; Nandakumar & Stout, 1993; Nandakumar & Yu, 1995; Roussos, Stout, & Marden, 1993).

The following uses of DIMTEST have been proposed (Stout et al., 1996): to verify or refute unidimensionality, to assess whether a specified subset of items is dimensionally distinct from the reminder of the test, and to determine in some detail the dimensionality structure of the test.

The goal of this research was to explore the dimensionality structure of tests in two achievement areas, language arts and mathematics, by means of the Poly-DIMTEST procedure. Three potential sources of dimensional distinctness among items in the tests were considered: that caused by item format, that caused by item content, and that caused by item location in the test (e.g., within item cluster associated with a common stimulus, such as passage or problem setting).

Method

Participants

The analysis was conducted on the results from the joint administration of the Constructed-Response Supplement to Iowa Tests of Basic Skills and Form M of the Iowa Tests of Basic Skills to a national sample of students in grades seven and eight. The number of students who took both, constructed-response and multiple-choice, language tests was 952 at grade 7 and 882 at grade 8. The number of students who took both mathematics tests was 889 at grade 7 and 918 at grade 8.

Instruments

The Iowa Test of Basic Skills (ITBS) is a battery of achievement tests in several subject areas (Hoover, Hieronymus, Frisbie, & Dunbar, 1996). The tests are composed of multiple-choice (MC) items that have four or five options. The primary purpose of the battery is to provide information that can be used for improving instruction. The following tests were of interest for this research:

1. The Integrated Writing Skills Test (IWST). This test assesses students' skills in using the conventions of standard written English by measuring their ability to apply accepted standards for spelling, capitalization, punctuation, and usage in writing. The number of items on the IWST is 55 at grade 7 and 57 at grade 8. The administration time is 40 minutes at both grade levels. It should be noted that the IWST is not the language assessment used in the regular achievement battery, but rather a special version designed to measure a variety of language skills in a single, integrated subtest (Hoover, Hieronymus, Frisbie, & Dunbar, 1996).

2. The mathematics tests: Math Concepts and Estimation, and Math Problem Solving and Data Interpretation. The Math Concepts part of the first test assesses a number of important skills including

but not limited to the understanding of number systems, arithmetic and geometric operations, measurement, fractions, probability and statistics, and equations. The Math Problem Solving part of the second test requires solving of mathematical problems or applying problem-solving strategies. The Data Interpretation part measures various skills in data interpretation. The total number of items on both mathematics tests is 87 at grade 7 and 92 at grade 8. The administration time for the combined math tests is 60 minutes at both grade levels; however, the two mathematics subtests described here are administered in separately timed testing sessions.

The Constructed-Response Supplement (CRS) to the ITBS, as used in this study, includes tests in two content areas: language, and mathematics. The tests are designed to assess achievement of many of the content objectives that are measured by the multiple-choice items of the ITBS with additional emphasis on those areas of learning where assessment may be enhanced by employing an open-ended format.

The constructed-response (CR) language supplement assesses students' ability to develop and organize their ideas and to express them according to the conventions of standard written English. At all grade levels, the language tests include three parts: editing, revising, and generating. Part 1 (editing) contains three short stories, reports, or letters. The task is to identify the stories' parts that need to be edited and to make changes in spelling, capitalization, punctuation, and the use of words or phrases. Part 2 (revising) asks students to revise a story by completing sentences, changing them to express the idea more clearly, correcting grammatical mistakes, and completing the story. In part 3 (generating) the examinee is given a specified topic. Three separate tasks are required: defining the subject of the story, writing at least three ideas for the story, and writing a complete sentence that could be included in the story.

The constructed-response language test includes 26 items (52 total score points) at grade 7 and 30 items (60 total score points) at grade 8. Depending on the complexity of the items, responses are scored on a 0-1, a 0-1-2, or a 0-1-2-3 scale. General guidelines for scoring and a scoring key are used to assign points to student responses. For instance, the general guidelines specify that the items in the editing part are worth two points: one point for identifying the error and one point for the correction. In the revising and generating parts, the general guidelines instruct scorers to accept reasonable answers and to ignore errors in mechanics that are unrelated to the specific skill measured by the item.

The mathematics constructed-response supplement was designed to assess mathematical problem solving, data interpretation, conceptual understanding, and estimation skills with open-ended exercises, presented either discretely or in clusters related to a common data source, that allow students to analyze and solve problems and to describe their thoughts and results using words, diagrams, graphs, symbols, calculations, and equations. It provides students with opportunities to employ a variety of solution strategies to yield valid results; to explain their reasoning, show their work, and justify their conclusions; to generate responses to problems that have more than one answer; and to establish and evaluate connections among mathematical concepts and procedures.

The constructed-response mathematics test includes 17 items (24 total score points) at grade 7 and 13 items (24 total score points) at grade 8. Responses to each exercise are scored on a 0-1 or a 0-1-2 scale. The basic elements of the scoring materials are the same ones used for the language test. The general guidelines describe the characteristics of responses at each score level. More specifically, a 2-point response demonstrates a complete and correct understanding of all mathematical concepts and processes embodied in the task, a 1-point response demonstrates a partial understanding of one or more of the mathematical concepts and processes embodied in the task, and a 0-point response demonstrates no understanding. For each exercise, the detailed scoring key specifically describes the kinds of answers and work that would earn full or partial credit. The administration time for each of the constructed-response tests is 30 minutes.

For the language tests, the correlations between multiple-choice and constructed-response scores were .715 at grade 7 and .730 at grade 8. After adjustment for unreliability, disattenuated correlations were .899 at grade 7 and .901 at grade 8. For mathematics tests, the correlations were .768 and .803 and disattenuated correlations were .963 and .971, for grades 7 and 8, respectively.

Procedure

Poly-DIMTEST Statistical Procedure. Suppose J examinees take an $\underline{N}$ item test. Each examinee generates a vector of items responses, $(\underline{U_{1j}},...,\underline{U_{Nj}})$. Each item response is scored from 0 to $\underline{r_i}$, where $\underline{r_i}$ is the maximum possible score for the $\underline{i}$th item. It is assumed that the item category characteristic curve for the fully correct response to the item $\underline{i}$, i.e. $\underline{U_{ij}} = \underline{r_i}$, is monotonically increasing in ability $\boldsymbol{\theta}$. It is also assumed that, conditional on fixed ability values, examinee responses to different items are statistically independent.

A single application of either the DIMTEST procedure or Poly-DIMTEST procedure evaluates the conditional covariance relationship between two subsets of test items. The first subset, called the assessment subtest (AT1), comprises items whose dimensionality is compared to the remaining test items. The second subset, called the partitioning subtest (PT), is used to partition the examinees into groups based on their scores on the PT.

The hypothesis tested is

H_0 : AT1 $\cup$ PT satisfies $\underline{d} = 1$

And the alternative hypothesis is

H_1 : AT1 is "dimensionally distinct" from PT.

Like DIMTEST, the Poly-DIMTEST procedure involves four main steps.

First, from an N-item test, a subset of $\underline{M}$ items (AT1) is selected based on either preliminary exploratory data analysis or expert opinion. These $\underline{M}$ items are presumed dimensionally homogeneous and distinct from the rest of the items, or in geometric terms, they lie in a narrow sector of the latent space described previously.

Second, from the remaining items, another subset of $\underline{M}$ items (AT2) is chosen so that the distribution of their difficulty levels is as similar as possible to that of the first subset. In addition, it is desirable that AT2 has the same direction of best measurement as the remaining items (PT). AT2 is used to eliminate statistical bias possible when the total test is shorter than 80 items. This bias is caused by conditioning on the number correct score instead of on the expected number correct score when unidimensionality holds.

Third, the examinees are partitioned into $\underline{K}$ subgroups on the basis of their total scores on the remaining (after the two subsets are removed) items (PT). Each subgroup $\underline{k}$ contains $\underline{J_k}$ examinees.

And finally, the DIMTEST statistic is computed as follows (Li & Stout, 1995).

Within each subgroup $\underline{k}$, two variance estimates $\hat{\sigma}_{\underline{k}}^2$ and $\hat{\sigma}_{\underline{u},\underline{k}}^2$ and the standard error of estimate $\underline{S}_{\underline{k}}$ are calculated using AT1's item responses: the usual variance estimate given by

$$\hat{\sigma}_{k}^{2} = \sum_{j=1}^{J_k} \frac{(Y_{j}^{(k)} - \overline{Y}^{(k)})^{2}}{J_k}$$

and the "unidimensional" variance estimate given by

$$\hat{\sigma}_{u,k}^{2} = \frac{1}{M^{2}} \sum_{i=1}^{M} \left[\frac{\sum_{j=1}^{J_k} U_{ijk}^{2}}{J_k r_i^{2}} - \left(\frac{\sum_{j=1}^{J_k} U_{ijk}}{J_k r_i} \right)^{2} \right],$$

where U_{ijk} denotes the ith item response by jth examinee in the kth subgroup; $Y_j^{(k)}$ denotes the proportion correct on the AT1 subtest obtained by jth examinee in the kth subgroup, and $\overline{Y}^{(k)}$ denotes the average proportion correct on the AT1 subtest for the examinees in the kth subgroup.

Next, the difference between these two variance estimates is standardized within each subgroup k and added across all K subgroups to obtain the statistic T_L:

$$T_L = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \left(\frac{\hat{\sigma}_{k}^{2} - \hat{\sigma}_{u,k}^{2}}{S_k} \right).$$

The statistic T_B is computed in the same way using item responses to the subtest AT2.

The Stout's statistic T is given by

$$T = \frac{\left[T_L - \left(V_1 / V_2 \right)^{1/2} T_B \right]}{\left(1 + V_1 / V_2 \right)^{1/2}}.$$

The ratio $\underline{V}_1 / \underline{V}_2$ is an empirical weight used to adjust for the difference in maximum total scores between the AT1 subtest and the AT2 subtest. $\underline{V}_1$ is the variance of the set of integers from 0 to the maximum total score on the AT1 subtest, and $\underline{V}_2$ is the variance of the set of integers from 0 to the maximum total score on the AT2 subset.

Stout's statistic, $\underline{T}$, has been shown to be asymptotically normally distributed with mean 0 and variance 1 as the test length increases when essential unidimensionality holds (Li & Stout, 1995; Stout, 1987). The hypothesis of unidimensionality is rejected when $\underline{T} > \underline{Z}_\alpha$, where $\underline{Z}_\alpha$ is the upper $100(1 - \alpha)$ percentile of the standard normal distribution, and α is the desired level of significance.

For reasonable control of errors in hypothesis testing, Stout recommends selecting AT1 so that it includes no more than 1/3 of the total test items, or ideally, 1/4. It is also recommended that the length of the total test be no less than 20.

Analyses. Poly-DIMTEST was used to conduct a series of statistical tests of a priori hypotheses suggested by content analyses of the tests in question. Data from two grades, 7 and 8, were analyzed to compare the results across grades. The initial analysis strategy was the same for both language and mathematics tests at both grades. To investigate how item content classification and item location within a stimulus-based cluster influence the tests' dimensionality, two types of analysis were planned, content-based and stimulus-based. Item subsets, to be used as input AT1s to the Poly-DIMTEST program, were defined based on either the items' content classification or their relation to a common stimulus. The number of items in content categories of tests used in this analysis is shown in Tables 1 and 2.

The language and mathematics MC tests and language CR tests had enough items to use DIMTEST to perform relatively comprehensive dimensionality assessments of these tests, whereas the smaller number of items in the mathematics CR tests (18 at grade 7 and 13 at grade 8) precluded using DIMTEST to explore the dimensionality of these tests in isolation.

In the next stage of the analysis, the MC and CR tests in the same content area were combined, so as to form one combined language test and one combined mathematics test at each grade. To analyze these combined tests, content-based subsets of items were defined to use as AT1s for testing with the Poly-DIMTEST procedure. Also for each combined test, the total set of constructed-response items was used as AT1 to test the hypothesis that the directions of best measurement for the MC language test and the CR language test coincide.

However, for both language and mathematics content areas, the original plan of analysis required modification once initial DIMTEST findings became available. The analysis strategy for individual tests--depending on their content composition and format--evolved from the above plan in such a way that features of individual item clusters or entire tests as well as features of the DIMTEST method could be revealed. The details of these adjustments to the analysis are described in the results section.

Results

Language

The content similarity of the MC and CR language tests allowed comparison of the effects of item content on the tests' dimensionality across test formats as well as grades. At both grades, MC language tests and CR language tests included items that were designed to measure spelling, capitalization, punctuation, language usage, and written expression skills. The Poly-DIMTEST procedure was applied to test whether the subsets of items measuring separate content skills were dimensionally similar to the remaining test items. The results are reported in Tables 3 and 4 for MC language tests and CR language tests, respectively. The first column of the tables lists AT1 subsets. The second and fifth columns display the numbers of items in the corresponding AT1s for grades 7 and 8, respectively. The remaining cells contain the results from the Poly-DIMTEST program, namely the values of the T statistic and its observed significance level.

Because the MC written expression category included more items than was allowed by the DIMTEST algorithm, 12 items at grade 7 and 15 items at grade 8 were excluded from the analysis when testing the hypothesis that MC expression items were dimensionally indistinct from the remaining MC language items.

Although the results were similar across grades, they were strikingly disparate for different test formats. The values of _T_ were rarely significant (at $\alpha = .05$) when Poly-DIMTEST was applied to content-based item subsets of the MC language tests, but _T_ was almost always significant for content-based item subsets of the CR language tests. For the six content categories of the MC tests (spelling, capitalization, punctuation, usage, expression, and multiple skills), the average _T_ was 0.552 at grade 7 and 1.133 at grade 8; whereas for the five content categories of the CR tests (spelling, capitalization, punctuation, usage, and expression), the average _T_ was 3.821 at grade 7 and 3.207 at grade 8.

The length of the MC expression subset (23 at grade 7 and 26 at grade 8) was sufficient to compare the other MC content-based item subsets to it. The results of testing the hypotheses H_o : Spelling (Capitalization, Punctuation, Usage, or Multiple Skills) $\cup$ Expression satisfies _d_ = 1 are reported in Table 5. For both grades, all _T_s were far from reaching significance. In essence, this suggested that not one of the five content subsets was dimensionally distinct from the expression subset.

Each MC language test consisted of several stand-alone items and of six clusters of items united by a common passage. The CR language tests included three passage-based and two skill-based item clusters. Because one of the skill-based clusters included too few items (the three-item writing part) for the DIMTEST procedure to have sufficient power, it was excluded from the analysis. Values of _T_ were obtained by testing clusters other than the CR writing cluster against the remaining items in the test of which they were a part. The results (shown in Tables 3 and 4) were mixed for the language tests of both item formats. At both grades, T for half of the six MC passage-based clusters was significant. The significance of T for the last (sixth) cluster could be attributed to its location near the end of the test, although the test was not speeded and technically the items of this cluster were not last: the test ends with two stand-alone items. Also, the last passage and its items were identical for grades 7 and 8, thus confounding the effects of the passage's location and its contents. The results were similarly inconsistent for the CR language tests. At both grades, one of the three passage-based clusters produced a significant T. This cluster was located somewhere in the middle of the test rather than at its end. At both grades, T for the Revision part of the CR test was significant. The Revision part consisted of four items: one writing item and three written expression items.

The results from Poly-DIMTEST analysis of combined language tests are reported in Table 6. The hypothesis that the directions of best measurement for the MC language test and for the CR language test coincide was not rejected at either grade. Because the one to four ratio of the lengths of AT1 to the total test is believed to be providing the best results when testing with the DIMTEST procedure, the analysis was repeated with shortened CR tests. At this step, six CR items at grade 7 and ten CR items at grade 8 were excluded from the analysis. T was again far from significant at both grades.

The five-item CR expression subset was compared to the MC expression subset (23 items at grade 7 and 26 items at grade 8). At both grades, the resulting T was not significant (p=.784 at grade 7 and p=.231 at grade 8).

When content-based item subsets that combined items of both formats were compared to the items remaining in the corresponding combined language test, the values of T were significant for all five content categories that MC tests and CR tests had in common (i.e., spelling, capitalization, punctuation, usage, and expression). At both grades, the average T was 3.691. The Poly-DIMTEST procedure was reapplied to reduced expression subsets and resulted in lower, but still significant values of T.

A dilemma that we encountered while using DIMTEST concerned selection of AT2 by the Poly-DIMTEST program. To create a shortened CR language test at grade 8, we excluded ten items, so that the remaining 20 items were representative of the entire test. Then we performed a Poly-DIMTEST analysis using this short CR test as AT1 and the entire MC language test as AT2 and PT. As we reported previously, the resulting T was not significant (T = 0.317, p = .376). Next, attempting to improve content balance of the short CR test, we decided to exchange one of the items of the 20-item CR test for a CR item that had been previously excluded from the analysis. To our surprise, the value of T given this slightly modified AT1 was significant (T = 3.488, p = .000). To find an explanation for such a dramatic difference in $\underline{T}$ values, we examined intermediate results given by the program. The values of $\underline{T}_L$ for these two runs were similar ($\underline{T}_L$ = 10.922 for the first run and $\underline{T}_L$ = 10.844 for the second run). However, the values of $\underline{T}_B$ were 10.226 and 3.189 for the two runs, respectively. DIMTEST calculates $\underline{T}_L$ on the basis of responses to AT1, and $\underline{T}_B$ on the basis of responses to AT2. Next, we inspected the AT2s chosen by DIMTEST for these two runs and found that the change of a single item in AT1 effected the choice of the second AT2 so much that only half of its original items remained from the first run. Recall that AT2 is selected so that the distribution of item difficulties is as similar as possible to that of AT1. The item originally included in AT1 was easier than the new item. We found it remarkable that a change of one item in a 20-item test could effect the choice of AT2 and consequently the resulting value of $\underline{T}$ so much that the $\underline{T}$ became significant. Because the first choice of CR items was judged to be better balanced with respect to content/skills, we have decided to accept the first, insignificant value of $\underline{T}$ which, in addition, was consistent with other results. However, the demonstration of such extreme sensitivity of results to the selection of AT1 and AT2 remains an issue for routine application of this method for the evaluation of dimensionality.

Mathematics

At both grades, MC mathematics tests consisted of four separately timed subtests: Concepts, Estimation, Problem Solving (PS), and Data Interpretation (DI). Poly-DIMTEST was applied to compare each of these subtests to the items comprising the three remaining subtests. The results (shown in Table 7) somewhat disagreed at different grades. At grade 8, all *T*s were significant (average *T* = 3.348); whereas at grade 7, two values of *T* (for Concepts and Estimation subtests) could not be rejected at the .05 level. At this grade level, the average *T* was 2.397. Because Concepts subtests included more items than recommended for DIMTEST analyses, the procedure was reapplied for shortened Concepts subtests (20 items at grade 7 and 21 items at grade 8). This resulted in a significant *T* for the Concepts subtest at grade 7 (*T* = 1.911, *p* = .028).

PS subtests included four clusters of items at grade 7 and three clusters at grade 8, such that items in each cluster were based on a common problem setting. At either grade, DI subtests included four four-item clusters, such that items in each cluster were related to a common data source. To explore the hypothesis that each cluster of the PS and DI subtests was dimensionally similar to the other items in the total MC mathematics test, Poly-DIMTEST was run for each cluster, so that the items of the cluster were used as AT1 and the remaining mathematics items comprised AT2 and PT. The results are reported in Table 7. For three of the four problem solving clusters at grade 7, and for two of the three problem solving clusters at grade 8, the hypothesis of dimensional similarity between the cluster and the remainder of the MC mathematics test was rejected. Although the last PS cluster at grade 7 was identical to the first PS cluster at grade 8, the value of T associated with this cluster was significant at grade 7, but was not significant at grade 8. The results for DI clusters disagreed for different grades. At grade 7, none of the four clusters' Ts were significant, whereas at grade 8, three of the four clusters' Ts were significant at 0.05 level. It is worth noting that the last two DI clusters at grade 7 were identical to the first two DI clusters at grade 8.

We were surprised with the finding that whereas the T for the grade 7 DI subtest suggested that the entire collection of DI items was dimensionally distinct from the remaining mathematics items, individual DI clusters were not shown to be dimensionally distinct from the remaining items in the test. To investigate the reason for this disparity, we performed four additional DIMTEST analyses, so that each time one of the DI clusters was used as AT1 and the items comprising Concepts, Estimation, and Problem Solving subtests were used as AT2 and PT. The idea was to exclude the remaining three DI clusters from the DIMTEST tests of hypotheses that each DI cluster was dimensionally similar to the combined Concepts, Estimation, and PS test. However, only one of the four Ts increased enough to become significant. The values of T for DI clusters 1 through 4 were 0.738 ($p = .230$), 0.618 ($p = .268$), 0.395 ($p = .346$), and 1.701 ($p = .044$), respectively. Then, we compared each of the four DI clusters to the PS subtest only. Not one of the Ts was significant. And finally, we formed four four-item sets from the DI items, such that each set included one item from each of the four DI clusters, and compared these sets to the PS subset. Of the four Ts, only one was significant.

The next hypothesis tested was the hypothesis that the directions of best measurement for the MC mathematics test and for the CR mathematics test coincide. The resulting values of T were 3.039 (p = .001) at grade 7 and 3.434 (p = .000) at grade 8. Because there were 12 two-category items among 18 CR items at grade 7, these dichotomous CR items were compared to the MC mathematics items. The hypothesis of dimensional similarity between dichotomous CR items and MC items was rejected (T = 5.370, p = .000).

At both grades, CR mathematics items were classified according to the skills they were designed to measure into two broad categories: concepts/estimation and problem solving/data interpretation. Poly-DIMTEST was used to explore dimensional similarity between MC items and CR items for these two general content categories. The results are reported in Table 8. The hypothesis that the combined concepts/estimation CR subtest was dimensionally similar to combined concept/estimation MC test was rejected at both grades, and a similar hypothesis for problem solving/data interpretation combined content category was rejected at grade 8, but not at grade 7. The results of the analysis that included only dichotomous CR items at grade 7, were similar to the results for the entire collection of CR items.

Discussion

This study employed the DIMTEST statistical procedure in a confirmatory manner to explore dimensionality structures of three kinds of achievement tests: multiple-choice tests, constructed-response tests, and tests combining both item formats. Three potential sources of dimensional distinctness among items in the tests--caused by item format, item content, and item location--were investigated. The results differed for language and mathematics achievement areas, in part because the tests were dissimilar with regard to content composition and the number of items in the whole tests as well as subtests that were analyzed. To accommodate DIMTEST's requirements, the analysis strategy was tailored to each test. Although results generally agreed for grades 7 and 8, there were some exceptions to this rule for mathematics tests.

To facilitate interpretation of the results, an observation should be made regarding the terms "dimensionally distinct" and "dimensionally similar" that are commonly employed by researchers who use DIMTEST. Although we also use these terms to describe the results, we find them somewhat confusing. If the value of the _T_ statistic as given by the DIMTEST program turns out to be significant, the AT1 and PT tests are declared "dimensionally distinct." Often, however, an insignificant T may lead to a wrong conclusion that AT1 and PT are "dimensionally similar," or moreover, that the test combining AT1 and PT is essentially unidimensional.

It is not always recognized that failure to reject a hypothesis of essential unidimensionality does not imply that the test is essentially unidimensional. This is also true when AT1 is selected based on expert opinion. For instance, when we used the CR language test for AT1 and the MC language test for AT2 and PT, the resulting T was not significant at both grades. Because other DIMTEST analyses provided evidence that, first, both CR and MC language tests were not essentially unidimensional, and second, their dimensionality structures differed, the conclusion that the CR language test is "dimensionally similar" to the MC language test would be obviously wrong. Sometimes, when there is no a priori knowledge of AT1's dimensional structure, the AT1 may be erroneously presumed dimensionally homogeneous, thus creating potential for misinterpretation of DIMTEST's results. Therefore, we prefer using the geometric terms "direction of best measurement" or equivalently, "best measured [by the test's, or subtest's, score] unidimensional latent variable."

In attempting to interpret the results, we have a reservation concerning the magnitude of values of the T statistic. Because Hattie, Krakowski, Rogers, and Swaminathan (1993) found that T was not monotonically related to the underlying dimensionality and recommended against using it as a general index of dimensionality, we attend to the significance of T rather than to its magnitude.

Language

A surprising finding for language tests is the apparent interaction of item content/skills and item format. The Poly-

DIMTEST analysis provides evidence that the CR language test and the MC language test are possibly assessing the same "best measured unidimensional latent variable," which, we have reasons to believe on the basis of the tests' content and design, is language achievement. However, the difference between the internal dimensionality structures of the CR language test and MC language test is remarkable. Although the MC test cannot be strictly judged essentially unidimensional because some of its passages introduce "nuisance" dimensions into the test, with regard to content/skills the MC language test can be considered essentially unidimensional. The weak evidence of dimensional distinctness introduced by content is inconsistent across the two grades included in this study. At the same time, the analysis unambiguously indicates that the CR language tests conform to approximate simple structure, that is, their items lie in sectors around the coordinate axes corresponding to the content/skills classification. Furthermore, skill-related heterogeneity of the CR test remains evident in the analysis combining the CR items and the MC items.

This finding, if confirmed by other analyses, is important for both test developers and test users. Because traditional item response theory models, commonly used for scaling tests, assume unidimensionality, many test developers focus on creating approximately unidimensional tests. Thus, they might find the particular MC format of the language test analyzed in this paper appealing. On the other hand, if the goal is to build a language test in such a way that the subskill scores have meaningful interpretations, then the format of choice should be the format of the CR language test analyzed in this research.

It must be recalled that the MC language test (IWST) analyzed in this paper departs substantially from the regular ITBS MC language tests. The IWST has fewer items, and their format is different. Unlike the regular language tests, individual content scores for the IWST are not reported.

Similar to the IWST test, the editing part of the CR test consists of several texts--stories, reports, and letters--that contain errors in spelling, punctuation, capitalization, and language usage. The difference between the tests of different formats is in the way examinees respond to the errors. In the MC format, the text's fragments that possibly include mistakes are identified for the student by underlining. For each underlined segment, four options are offered, the "no change" option and three ways to correct an error if one is present. Because there is no indication of the location of errors in the texts of the CR test, examinees have no clues that could help them to identify errors. In scoring, one point is given for locating an error, and another one for a proper correction.

According to the results of the analysis reported in this paper, the distinction between correct-the-error MC items and find-and-correct-the-error CR items has a substantial impact on the apparent dimensionality of the measure of language skills.

Mathematics

The Poly-DIMTEST analysis of the mathematics tests suggests that both the MC test and the test combining MC and CR items are not essentially unidimensional. Although results somewhat differ for two grades, they clearly link the tests' apparent multidimensionality to item format, content, and location. The entire CR test as well as its subset composed of dichotomous items appear to be dimensionally distinct from the MC mathematics test.

The conflicting results for DI item clusters at grade 7 may be explained by insufficient power when testing AT1s that include only four items. It could be that the hypothesis of dimensional similarity of the entire DI subtest to the remaining mathematics items was rejected because the DI subtest was sufficiently long (16 items) to provide adequate testing power. This outcome, however, makes DIMTEST's analyses of four-item clusters suspect.

Conclusion

The Poly-DIMTEST analysis presented in this paper provides evidence of dimensional heterogeneity of tests intended to provide assessments that are balanced with respect to the content composition of a targeted construct, such as language or mathematics achievement. Furthermore, the analysis suggests that combining items of different formats may introduce additional complexity into the dimensionality structure of the composite test. This outcome questions using unidimensional estimation procedures based on the traditional IRT models for scaling and equating achievement tests similar to the tests examined in this research. However, because of its inherent limitations, the Poly-DIMTEST analysis neither permitted determination of exact dimensionality structures of the tests, nor offered explanation for sometimes peculiar outcomes.

References

Ackerman, T. A. & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Breland, H. M. & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Bridgeman, B., & Rock, D. (1993). Relationships among multiple-choice and open-ended analytical questions. Journal of Educational Measurement, 30, 313-329.

Frederiksen, N. (1984). The real test bias: influence of testing on teaching and learning. American Psychologist, 39, 193-202.

Hogan, T. P., & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 17, 219-227.

Hoover, H. D., & Bray, G. B. (1995, April). The research and development phase: can a performance assessment be cost effective? Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hoover, H.D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). Iowa Tests of Basic Skills, Form M, Interpretive Guide. Chicago: The Riverside Publishing Company.

Li, H. & Stout, W. (1995). A version of DIMTEST to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data. Paper presented at the annual meeting of the NCME, San Francisco.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores (pp. 359-382). Reading, MA: Addison-Wesley.

Lukhele, R., Thissen, D. & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 31, 234-250.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses--comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Nandakumar, R. & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 18, 41-68.

Nandakumar, R. & Yu, F. (1995). Assessing unidimensionality of polytomous data. Applied Psychological Measurement, 22, 99-115.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15, 361-373.

Roussos, L. A., Stout, W. F., & Marden, J. I (1993). Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis. Paper presented at the annual NCME meeting, Atlanta, GA.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-326.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST Manual. Department of Statistics, University of Illinois at Urbana-Champaign.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos L., Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. Applied Psychological Measurement, 20, 331-354.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31, 113-123.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1, 355-369.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. Applied Psychological Measurement, 14, 1-12.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward Marxist theory of test construction. Applied Measurement in Education, 6(2), 103-118.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6, 1-11.

Wilson, M. & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. Applied Psychological Measurement, 19, 51-71.

Zhang, J. & Stout, W. F. (1996a, April). A new theoretical DETECT index of dimensionality and its estimation. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Zhang, J. & Stout, W. F. (1996b). Conditional covariance structure of generalized compensatory multidimensional items with applications. Manuscript submitted for publication.

Table 1

Number of Items in Content Categories of Multiple-Choice Tests

	Number of Items	
Test and Content Category	Grade 7	Grade 8
Language	55	57
Spelling	5	4
Capitalization	8	8
Punctuation	8	8
Usage	3	7
Expression	23	26
Multiple Skills	8	4
Mathematics	87	92
Concepts	30	32
Estimation	22	24
Problem Solving	19	20
Data Interpretation	16	16

Table 2

Number of Items in Content Categories of Constructed-Response Tests

	Number of Items	
Test and Content Category	Grade 7	Grade 8
Language	26	30
Spelling	4	5
Capitalization	5	5
Punctuation	4	6
Usage	5	6
Expression	5	5
Writing	3	3
Mathematics	18	13
Concepts/Estimation	8	6
Problem Solving/ Data Interpretation	10	7

Table 3

Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Multiple-Choice Integrated Writing Skills Tests

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
Content-based						
Spelling	5	-0.683	.753	4	0.992	.160
Capitalization	8	-0.580	.719	8	1.477	.070
Punctuation	8	-0.520	.698	8	1.764	.039
Usage	3	0.035	.468	7	1.259	.104
Multiple Skills	8	2.089	.018	4	0.574	.283
Expression[a]	11	2.969	.002	11	0.732	.232
Stimulus-based						
Passage 1	8	2.544	.006	8	-0.428	.517
Passage 2	8	0.578	.282	8	1.542	.062
Passage 3	8	2.094	.018	8	1.659	.048
Passage 4	8	0.677	.249	8	0.819	.206
Passage 5	8	0.279	.390	8	1.812	.036
Passage 6	6	4.209	.000	6	2.666	.004

Note. Number of items in MC language tests was 55 at grade 7 and 57 at grade 8.

[a] Twelve items at grade 7 and 15 items at grade 8 were excluded from the analysis to satisfy Poly-DIMTEST's restriction on the length of AT1.

Table 4

Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Constructed-Response Language Tests

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
Content-based						
Spelling	4	4.563	.000	5	4.554	.000
Capitalization	5	6.138	.000	5	4.583	.000
Punctuation	4	2.483	.006	6	1.341	.090
Usage	5	2.906	.002	6	2.055	.020
Expression	5	3.016	.001	5	3.501	.000
Skill-based						
Revision	4	2.383	.009	4	2.990	.001
Stimulus-based						
Passage 1	5	1.299	.097	7	1.245	.106
Passage 2	6	0.336	.368	4	-0.634	.737
Passage 3	8	2.097	.018	7[a]	3.483	.000

Note. Number of items in CR language tests was 26 at grade 7 and 30 at grade 8.

[a] Five items were excluded from the analysis to satisfy Poly-DIMTEST's restriction on the length of AT1.

Table 5

Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Multiple-Choice Spelling, Capitalization, Punctuation, Usage, and Multiple Skills Item Subsets vs. Expression Subset

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
Spelling	5	-0.566	.714	4	0.567	.285
Capitalization	8	-0.179	.571	8	0.381	.352
Punctuation	8	1.159	.123	8	1.298	.097
Usage	3	-1.808	.965	7	0.020	.492
Multiple Skills	8	0.988	.162	4	0.836	.202

Note. Number of MC expression items was 23 at grade 7 and 26 at grade 8.

Table 6

Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Combined Language Tests

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
CR items	26	1.216	.112	30	0.606	.272
CR items[a]	20	-0.560	.712	20	0.317	.376
Content-based						
Spelling	9	4.290	.000	9	6.125	.000
Capitalization	13	3.688	.000	13	2.070	.019
Punctuation	12	2.302	.011	14	2.947	.002
Usage	8	3.359	.000	13	4.424	.000
Expression	28	4.817	.000	31	2.890	.002
Expression[b]	23	1.969	.024	24	2.097	.018

Note. Number of items in combined language tests was 81 at grade 7 and 87 at grade 8.

[a] Six items at grade 7 and ten items at grade 8 were excluded from the analysis to satisfy Poly-DIMTEST's restriction on the length of AT1.

[b] Five items at grade 7 and seven items at grade 8 were excluded from the analysis to satisfy Poly-DIMTEST's restriction on the length of AT1.

Table 7

Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Multiple-Choice Mathematics Tests

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
Content-based						
Concepts	30	1.372	.085	32	3.693	.000
Concepts[a]	20	1.911	.028	21	3.197	.001
Estimation	22	1.491	.068	24	3.856	.000
Problem Solving	19	2.809	.002	20	2.021	.022
Data Interpretation	16	3.917	.000	16	3.821	.000
Stimulus-based						
PS 1	4	0.152	.440	8	1.595	.055
PS 2	4	2.746	.003	6	2.476	.007
PS 3	3	3.191	.001	6	1.790	.037
PS 4	8	2.985	.001	--	--	--
DI 1	4	0.547	.292	4	-0.600	.726
DI 2	4	0.462	.322	4	2.272	.011
DI 3	4	0.539	.295	4	1.920	.027
DI 4	4	1.304	.096	4	2.151	.016

Note. Number of items in MC mathematics tests was 87 at grade 7 and 92 at grade 8. PS 1 through PS 4 = item clusters related to common problem settings. DI 1 through DI 4 = item clusters related to common graphs or tables. Dashes indicate there was no fourth problem solving cluster at grade 8.

[a] Ten items at grade 7 and 11 items at grade 8 were excluded from the analysis to satisfy Poly-DIMTEST's restriction on the length of AT1.

Table 8

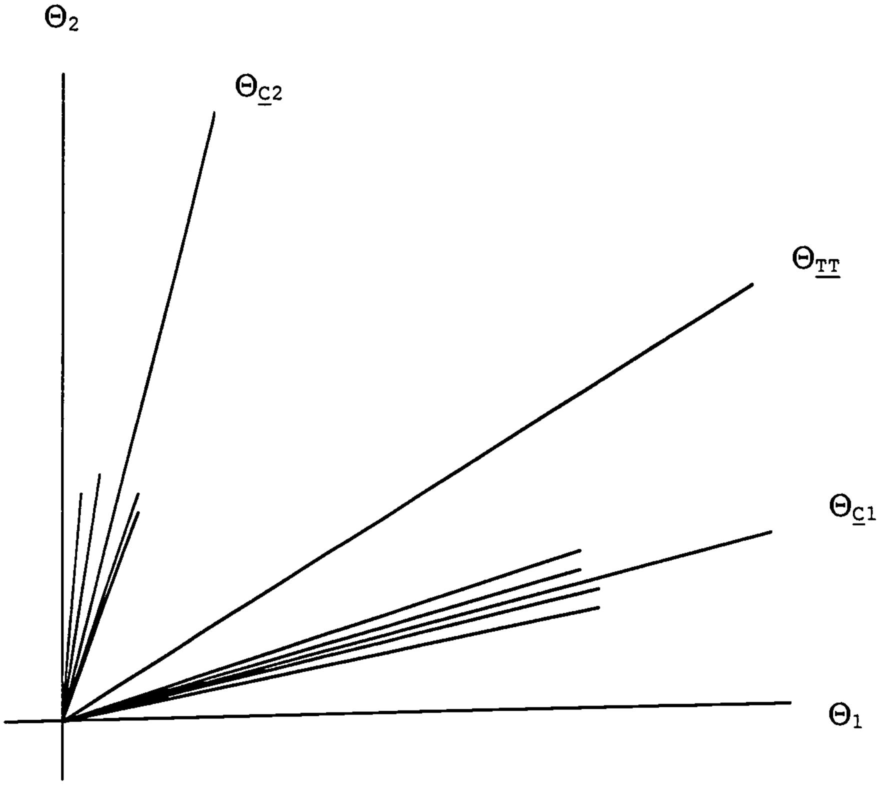
Values of T and Observed Significance (p) from Poly-DIMTEST Analysis of Constructed-Response Items vs. Multiple-Choice Items for Mathematics Subskill Categories

	Grade 7			Grade 8		
AT1	n_{AT1}	T	p	n_{AT1}	T	p
Concepts/Estimation	8	4.339	.000	6	2.461	.007
Concepts/Estimation[a]	5	3.645	.000	--	--	--
Problem Solving/ Data Interpretation	10	0.025	.490	7	2.361	.009
Problem Solving/ Data Interpretation[a]	7	0.745	.228	--	--	--

Note. Number of MC concepts/estimation items was 52 at grade 7 and 56 at grade 8. Number of MC problem solving/data interpretation items was 35 at grade 7 and 36 at grade 8.

[a] Only dichotomous CR items were compared to the MC items of the corresponding subskill category. Dashes indicate that the number of dichotomous CR items was too small for the analysis.

Figure 1. A test demonstrating approximate simple structure


Θ_2
$\Theta_{\underline{C}2}$
$\Theta_{\underline{TT}}$
$\Theta_{\underline{C}1}$
Θ_1

Note. The figure was adapted from "Conditional Covariance-Based Nonparametric Multidimensionality Assessment," by W. Stout, B. Habing, J. Douglas, H. R. Kim, L. Roussos, and J. Zhang, 1996, Applied Psychological Measurement, 20, p. 334. Copyright 1996 by the Applied Psychological Measurement Inc.

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

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Influences of Item Content and Format on the Dimensionality of tests Combining Multiple-Choice and Open-Response Items: An Application of the Poly-DIMTEST procedure	
Author(s): Yelena Perkhounkova and Stephen B. Dunbar	
Corporate Source:	Publication Date: April, 1999

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 in the indicated space following.

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)	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)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
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.

Signature: Perkhounkova	Printed Name/Position/Title: Yelena Perkhounkova,	
Organization/Address: University of Iowa 224A LC Iowa City, IA 52242	Telephone: (319) 335-5560	Fax:
	E-mail Address: yelena-perkhounkova@uiowa.edu	Date: May 20, 1999