ED 390 895 TM 024 177

AUTHOR Wainer, Howard; And Others
TITLE How Unidimensional Are Tests Comprising Both Multiple-Choice and Free-Response Items? An Analysis of Two Tests. Program Statistics Research Technical Report No. 93-32.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY North Carolina State Dept. of Public Instruction, Raleigh.
REPORT NO ETS-RR-93-28
PUB DATE May 93
NOTE 16p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Advanced Placement; Chemistry; Computer Science; High Schools; *High School Students; *Multiple Choice Tests; Test Construction; *Test Items
IDENTIFIERS Advanced Placement Examinations (CEEB); Confirmatory Factor Analysis; *Free Response Test Items; Multidimensionality (Tests); *Unidimensionality (Tests)

ABSTRACT

The relationship between the multiple-choice and free-response sections of the Computer Science and Chemistry tests of the College Board's Advanced Placement program was studied. Confirmatory factor analysis showed that the free-response sections measure the same underlying proficiency as the multiple-choice sections for the most part. However, there was a significant, if relatively small, amount of local dependence among the free-response items that produced a small degree of multidimensionality for each test. Data from the Computer Science test came from 2 random samples of 1,000 students each from a previous study of the 1988 test administration. Data from the Chemistry test came from a sample of 2,000 students choosing one free-response problem and 2,000 choosing another. (Contains 4 tables, 1 figure, and 17 references.) (Author/SLD)

RR-93-28

How Unidimensional Are Tests Comprising Both Multiple-Choice and Free-Response Items? An Analysis of Two Tests

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina

Xiang-Bo Wang
Law School Admissions Council

ETS®

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 93-32

Educational Testing Service
Princeton, New Jersey 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

How Unidimensional Are Tests Comprising Both Multiple-Choice and Free-Response Items? An Analysis of Two Tests

Howard Wainer
Educational Testing Service

David Thissen
University of North Carolina

Xiang-Bo Wang
Law School Admissions Council

Program Statistics Research
Technical Report No. 93-32

Research Report No. 93-28

Edcational Testing Service
Princeton, New Jersey 08541

May 1993

How Unidimensional are Tests Comprising both Multiple-Choice and Free-Response Items? An Analysis of Two Tests[1]

David Thissen
University of North Carolina at Chapel Hill

Howard Wainer
Educational Testing Service

Xiang-Bo Wang
Law School Admission Services

Abstract

We consider the relationship between the multiple-choice and free-response sections on the Computer Science and Chemistry tests of the College Board's Advanced Placement program. Confirmatory factor analysis shows that the free-response sections measure the same underlying proficiency as the multiple-choice sections for the most part. However, there is also a significant, if relatively small, amount of local dependence among the free-response items that produces a small degree of multidimensionality for each test.

[1] This research was supported, in part, by the Educational Testing Service through the Research Statistics Project and by the North Carolina Department of Public Instruction. Xiang-Bo Wang's work on this project was done while he was an Educational Testing Service Summer Predoctoral Fellow. The Chemistry AP data, and advice about how to interpret them were supplied by Walter MacDonald, Behroz Maneckshana, Joe Stevens, and Hessie Taft. Without their expert help and kind cooperation we would have had little to write about. We thank Mark Reckase for the suggestion that led to this paper. Any errors that remain are, of course, our own.

How Unidimensional are Tests Comprising both Multiple-Choice and Free-Response Items? An Analysis of Two Tests

There is increasing interest in the development of tests that combine multiple-choice and free-response items. Such tests raise the following two questions:

(1) Are we measuring the same thing with the free-response items as we are with the multiple-choice questions?

(2) Is it meaningful to combine the scores on the free-response sections with the multiple choice score, to yield a single reported total score?

Answers to these questions are necessary to construct appropriate score-reporting strategies for such tests, as well as to consider the thornier problems that arise when the examinee is permitted to choose to answer a subset of the time-consuming free-response questions (Wainer, Wang & Thissen, 1991; 1993). The use of item response theory (IRT) to score the test, or to equate forms comprising chosen questions, requires that the test (or forms) be essentially unidimensional—that all the items measure more-or-less the same thing.

Are such tests unidimensional? The literature on this subject is equivocal. Bennett, Rock, Braun, Frye, Sprohrer & Soloway (1991) fitted different factor structures to two relatively similar combinations of multiple-choice, free-response, and constrained free-response items; a one-factor model was sufficient for one set of data, but a two-factor model was required for another similar set of data. Bennett, Rock & M. Wang (1991) examined a particular two-factor model for the combined multiple-choice and free-response items on the College Board's Advanced Placement (AP) test in Computer Science, and concluded that the one-factor model provided a more parsimonious fit.

Here, we re-analyze the Computer Science AP data reported by Bennett et al. (1991), and show that significant, albeit relatively small, factors explain some of the observed local dependence among the free-response items. We replicate this finding using data from the AP test in Chemistry.

Example I: Advanced Placement in Computer Science

Bennett, Rock & M. Wang (1991) considered data obtained from two random samples of 1000 students each, drawn from among the 7,372 students taking the 1988 administration of the AP Computer Science "AB" examination. "The "AB" examination is intended to assess mastery of topics covered in a college-level introductory course in computer science" (Bennett et al., 1991. p. 78). The test includes a 50-item multiple-choice section and five free-response items. The free-response exercises require the student to write or design a brief Pascal program; the responses are graded on a 0-9 scale.

To factor analyze the entire test, Bennett et al. (1991) divided the multiple-choice section into five "item parcels" (Cattell, 1956, 1974; Cattell & Burdsal, 1975; Cook, Dorans, Eignor & Peterson, 1985; Dorans & Lawrence, 1987, 1991). Each parcel comprised 10 multiple-choice items, and produced a summed score ranging from 0-10. Using Maximum Likelihood (ML) factor analysis of the five multiple-choice item parcels and the five free-response problems, Bennett et al. concluded that a single-factor model produced a more parsimonious fit than the two-factor model they considered. However, the one-factor model did not fit the data particularly well, as measured by the likelihood ratio goodness of fit test. And the only two-factor model considered by Bennett et al. was a model in which the multiple-choice items loaded only on one factor, and the free-response items loaded only on the other (correlated) factor.

We consider an alternative model that includes a general factor for all of the items, plus two orthogonal specific factors for the free-response items. Factor loadings for this model are shown in Table 1. (Many of the factor loadings in Table 1 are greater than one, because these are loadings for the covariances among the item parcels and free-response items, not the correlations.) The models are the same for the two samples, except for the small negative loading of free-response item B on factor 2 and an even smaller error covariance for multiple choice parcels A and E (not shown in the table), both in the first

Table 1

Loadings for Three-Factor Solutions for the Computer Science AP Data

Item	*Sample 1*			*Sample 2*		
Parcel	λ_1	λ_2	λ	λ_1	λ_2	λ_3
A	1.90	—	—	1.78	—	—
B	1.80	—	—	1.70	—	—
C	1.67	—	—	1.69	—	—
D	1.43	—	—	1.36	—	—
E	1.74	—	—	1.58	—	—
Problem						
A	2.13	—	.95	1.98	—	1.01
B	1.91	−.28	.95	1.80	—	1.01
C	1.85	.56	—	1.76	.64	—
D	1.72	.72	—	1.56	1.02	—
E	1.60	1.08	—	1.35	1.07	—

sample only.[2] The models fit the data, as measured by the likelihood-ratio goodness of fit statistic: For sample 1, $G^2(29) = 41, p = .06$; and for sample 2, $G^2(31) = 36, p = .23$. All of the loadings shown in Table 1 have t-values (for the loading divided by its standard error) greater than 4.0, except for the aforementioned loading of free-response item B on factor 2 and the error covariance for multiple choice parcels A and E, both in the first sample only. Because the latter two parameters differ significantly from zero in only one of the two random samples, and because they are small for that sample, it is likely that they represent Type I errors.

The model shown in Table 1 fits the Computer Science AP data considerably better than the two-factor models reported by Bennett et al. (1991), which had values of G^2 over 100 with 34 *d.f.* It appears that the general-plus-specific-factors model used here is a more parsimonious way to examine the dimensionality of mixed multiple-choice and free-response tests than the two-factor model that postulates two correlated factors: one for the multiple-choice items only, and the other for the free-response items only.

There are clearly free-response factors. While these free-response factors have loadings that are significantly different from zero, they account for relatively little of the observed covariance among the items: The loadings of the free-response items on the free-response factors are uniformly smaller than the loadings of the free-response items on the general factor, indicating that the free-response items measure the same proficiency as the multiple-choice items, for the most part. The two free-response factors appear to represent content-based local dependence (Yen, 1992) among the items; free-response items C, D, and E cover programming material that is usually taught and learned graphically, while free-response items A and B are numerical or algebraic problems.

Example II: Advanced Placement in Chemistry

The 1989 Advanced Placement Examination in Chemistry is divided into two sections with 90 minutes allotted for each. Section I consists of 75 five-option multiple choice questions and accounts for 45% of the total grade. Section II consists of problems and essay questions, and has four parts:

- Part A is a single problem (Problem 1) that all examinees must answer, and accounts for 14% of the total grade.

- Part B has two problems (Problems 2 and 3), and the examinee must answer exactly one of those. This part accounts for 14% of the total grade.

- Part C is treated as a single problem (Problem 4), but has eight components, from which the examinee must answer five. This part accounts for 8% of the total grade.

[2] The loadings on factor 3 were constrained to be equal, to identify the model.

- Part D has five problems (Problems 5, 6, 7, 8 and 9) of which the examinee must answer three. This part accounts for 19% of the total grade.

Figure 1 shows the structure of the test graphically.

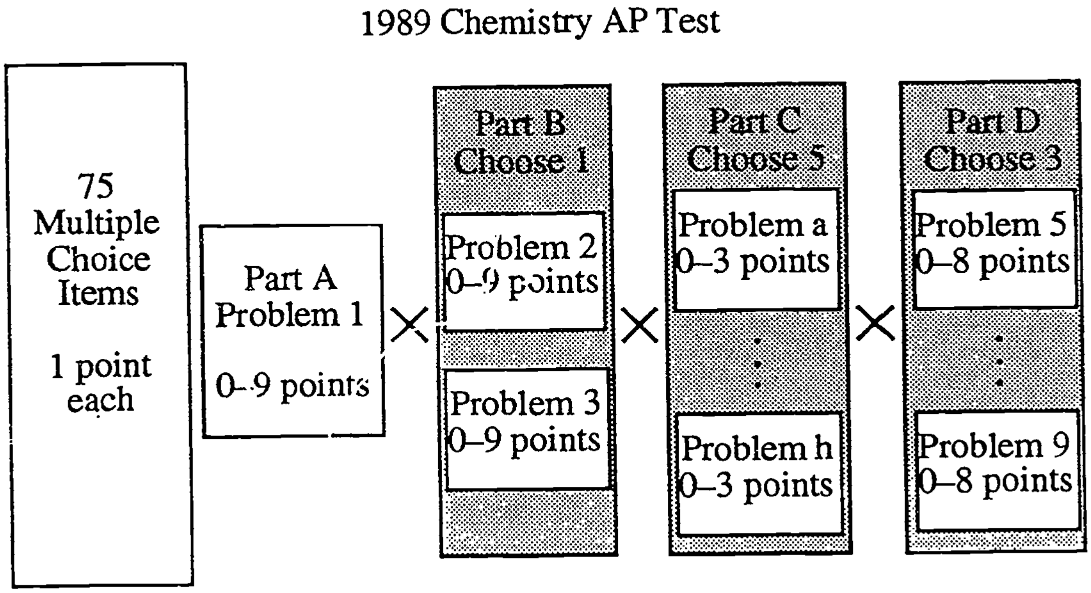

Figure 1. Graphical representation of the 1990 Chemistry AP test, showing the arrangement of items and problems into sections, and the choices.

This form of the exam was taken by approximately 18,000 students[3] in 1989. The test form has been released and interested readers may obtain copies with the answers and a full description of the scoring methods from the College Board.

We examined the dimensionality of the multiple-choice items using full-information item factor analysis (Bock, Gibbons, & Muraki, 1988) and discovered that, although three dimensions were required to obtain an acceptable fit, those three dimensions were highly correlated and a one-dimensional solution did very well indeed. Trimming some of the late-appearing items that were not reached by a significant proportion of the sample further strengthened the one-dimensional solution.

Because full information item factor analysis is currently implemented in software only for binary items, in order to consider the factor structure of the multiple-choice items and the free-response items jointly we followed the same strategy as that employed by Bennett et al. (1991). We divided the 75 multiple-choice items into 15 five-item parcels; the first parcel comprised items 1, 16, 31, 46, and 61; the second parcel included items 2,

[3] The actual number of examinees was 18,462; however 31 tests were handed-in essentially blank and so were excluded from the analysis.

17, ...; and so on. We constructed the parcels in this way because, while there may be some slight content-based multidimensionality in the multiple-choice section of the test, we wanted to examine the relation of the free-response items with the principal axis. We hoped that constructing the parcels explicitly without respect for item content would produce parcels that would be approximately equally correlated. The items appear on the test roughly in order of their difficulty: the equal-spacing of the items in each parcel was chosen so that the parcels would be approximately equally difficult with maximum variation in parcel summed scores.

Because of the probable non-normality of the summed parcel scores (ranging only from 0 to 5), and the clear non-normality of the free-response problem scores (that range 0-8 or 0-9, but are highly skewed), we used weighted least squares (WLS) estimation for the factor models, computed from the matrix of polychoric correlations among the parcels and problems. WLS estimation is asymptotically distribution free (Browne, 1982; 1984). The computations were done using PRELIS and LISREL 7 (Jöreskog & Sörbom, 1986; 1988).

The examinee-choice among the free-response items on the Chemistry AP examination makes it impossible to simultaneously factor analyze all of the free-response problems with the multiple-choice item parcels: No examinees responded to all of the free-response items, and the number of examinees that chose any particular subset is too small to provide data for factor analysis. Therefore, we divided the examinees into groups based on their choices of free-response items, and performed factor analyses within those groups, considering only the free-response items chosen.

First, we consider free-response problems 1, 2, and 3. All of the examinees responded to problem 1; 14,290 chose problem 2 and 4,172 chose problem 3. We analyzed data from a sample of $N = 2000$ from the group choosing problem 2, and from another sample of $N = 2000$ from the group choosing problem 3. The results are shown in Table 2. For each of the samples, we fit a one-factor model, and then tested the significance of the residual covariance (θ_δ) between the two free-response items (problems 1 and 2 for the first sample, and problems 1 and 3 for the second sample). The residual covariance is algebraically equivalent to the squared factor loading of the two free-response items on a second, uncorrelated free-response factor, if those two factor loadings are constrained to be equal. For a factor related to only two items, this single parameter is the only one that is identified.

Table 2

Loadings for Two-Factor Solutions for the Chemistry AP Data, Including Problems 1 and 2 or 3

Item	*Choice: 2*		*Choice: 3*	
Parcel	λ_1		λ_1	
1	.62		.64	
2	.62		.65	
3	.54		.60	
4	.63		.67	
5	.69		.69	
6	.60		.63	
7	.62		.65	
8	.68		.70	
9	.60		.65	
10	.72		.74	
11	.72		.78	
12	.57		.66	
13	.74		.73	
14	.66		.62	
15	.59		.64	
Problem				
1	.72		.75	
2	.74	$\theta_\delta(1,2)$:	—	$\theta_\delta(1,3)$:
3	—	.10	.73	.05
$G^2(1)$		47.1		11.6

For both samples, the residual covariance (θ_δ) between the two free-response items is significantly greater than zero; the G^2 tests of significance are given in Table 2, and the estimates of the covariances are 0.10 for problems 1 and 2 for the first sample, and 0.05 for problems 1 and 3 for the second sample. These values correspond to loadings of about 0.3 for the first sample, and 0.2 for the second sample. The loadings of the free-response items on the first (general) factor are much higher than these values, so the results are very much the same as those for the Computer Science examination: The free-response items are mostly related to the same general factor as are the multiple-choice items, but there is significant (and small) local dependence between the free-response items as well.

Overall, the models shown in Table 2 do not completely fit the data; the goodness of fit G^2s (on 118 *d.f.*) are 193 for the problem 1-2 group and 177 for the problem 1-3 group. This lack of fit appears to be due to some degree of local dependence (multidimensionality) among the multiple-choice item parcels themselves; for several pairs of parcels, the modification index computed by LISREL 7 indicates that the G^2 test for these error covariances would be as large as that obtained for the free-response items. Because our goal here is not to determine the detailed factor structure of the Chemistry items themselves, and because we did not construct the item parcels to be interpretable in a multidimensional solution, we did not pursue extensive model modification for the multiple-choice part of the test. However, we note that it appears that the free-response items do not add substantially more multidimensionality to the test than already exists in the multiple-choice section.

Next we considered groups of test-takers choosing various triples from among free-response items 5-6-7-8-9 in Part D of the test. Table 3 shows the numbers of examinees choosing each of the five most popular triples; all of those choice patterns yield $N > 1000$. For the other five triples, $N < 1000$; a sample size too small for WLS factor analysis when there are 19 variables (15 item parcels, free-response item 1, and the triple). The word *asymptotically* in *asymptotically distribution free* is best taken seriously. Therefore, we factor analyzed the appropriate set of variables for each of the five groups with sufficiently large sample sizes.

We used the same type of restricted factor analysis model as was used for the Computer Science data in the previous section to produce a general factor and a factor specifically for the free-response items. For all five samples, the joint G^2 test of significance of the loadings on the specifically free-response factor was significant; those values are given at the bottom of Table 4. The overall G^2 measures of goodness of fit of the two-factors models are in Table 3; as is to be expected with the likelihood ratio goodness of fit test, the values are roughly proportional to sample size. All of the values in Table 3 indicate the same lack of fit as was observed in the analyses of the multiple-choice item parcels with free-response problems 1, 2, and 3, for the same reason.

Table 3

Numbers of Test-Takers Choosing Five Triples from among Problems 5-6-7-8-9 on the Chemistry AP Examination

Problems chosen in Part D	*N*	2-factor-model G^2 (148 *d.f*)
567	2,555	249
568	5,227	403
578	4,918	457
589	1,392	186
678	1,707	197

For all of the choice-combinations, the free-response items have substantial loadings on the general factor, as shown in Table 4. Interestingly, free-response problem 1 has tiny (often negative) loadings on the free-response factor, which thus appears not to be a free-response factor at all but rather a content-based factor, specific to a particular triple chosen from among items 5-6-7-8-9. Indeed, the second factor often appears to be either a doublet (with only two substantial loadings), or nearly unique to a single item. As was the case with the Computer Science free-response items, there appears to be some degree of local dependence among the Chemistry free-response items, but the amount of local dependence appears to be small.

Table 4

Loadings for Two-Factor Solutions for the Chemistry AP Data, Including Problems 1 and Three Chosen from 5-6-7-8-9

Item	*Choices: 567*		*Choices: 568*		*Choices: 578*		*Choices: 589*		*Choices: 678*	
Parcel	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2
1	.66	—	.60	—	.64	—	.63	—	.61	—
2	.65	—	.63	—	.64	—	.67	—	.64	—
3	.54	—	.54	—	.56	—	.60	—	.59	—
4	.68	—	.64	—	.65	—	.66	—	.64	—
5	.69	—	.67	—	.68	—	.71	—	.63	—
6	.60	—	.59	—	.60	—	.64	—	.57	—
7	.64	—	.63	—	.64	—	.65	—	.59	—
8	.70	—	.68	—	.68	—	.71	—	.67	—
9	.58	—	.62	—	.63	—	.62	—	.58	—
10	.73	—	.71	—	.74	—	.74	--	.72	—
11	.73	—	.75	—	.74	—	.74	—	.74	—
12	.62	—	.61	—	.61	—	.66	—	.56	—
13	.67	—	.72	—	.72	—	.77	—	.71	—
14	.65	—	.61	—	.64	—	.66	—	.65	—
15	.65	—	.59	—	.60	—	.64	—	.58	—
Problem										
1	.80	.04	.74	–.07	.76	–.02	.74	–.00	.73	–.04
5	.64	.17	.57	.04	.57	–.02	.60	–.03	—	—
6	.68	.28	.58	.12	—	—	—	—	.59	.11
7	.50	.40	—	—	.46	.52	—	—	.38	.27
8	—	—	.45	.50	.46	.15	.49	.87	.40	.58
9	—	—	—	—	—	—	.62	.08	—	—
$G^2(4)$		45		30		34		13		47

Discussion

There is clear evidence that the free response problems on both the Computer Science and Chemistry AP tests "measure something different" than the multiple-choice sections of those tests: There are statistically significant factors for the free-response items, orthogonal to the general factor. However, there is also clear evidence that the free-response problems *predominantly* measure the same thing as the multiple choice sections: The factor loadings for the free-response items are almost always larger on the general (multiple-choice) factor than on the free-response factor(s). The loadings of the free-response items on the specifically free-response factors are small, indicating that the free-response items do not "measure something different" very well. Given the small size of the free-response factor loadings, it is clear that it would take many free-response items to produce a reliable score on the factor underlying the free-response items alone—many more items than are currently used.

Is it meaningful to combine the scores on the free-response sections with the multiple choice score, to yield a single reported score? It probably is; indeed, given the small size of the loadings of the free-response items on their own specific factors, it would probably not be meaningful to attempt to report a free-response score separately, because it would not be reliably distinct from the multiple-choice score.

If a test comprises multiple-choice items and a single free-response item, the scores surely may be combined, especially if IRT weights are used; the single free-response item would then simply be weighted according to its relation with the multiple-choice items. If the test comprises several free-response items, then the local dependence among the free-response items creates some problems for IRT analysis of the scale (Yen, 1992). Those problems can be solved by combining the locally-dependent free-response items into a single testlet (Wainer & Kiely, 1987; Yen, 1992). However, the degree of local dependence among the free-response items on the Computer Science and Chemistry AP tests considered here appears to be sufficiently small that it might also be ignored for many purposes; the local dependence among the free-response items is no larger than that which also exists among the multiple-choice items themselves.

We have considered only two tests, AP Computer Science and Chemistry; for both tests, the free-response items are intended to measure essentially the same proficiency as the multiple-choice items—and they appear to do so. There are certainly other tests that use multiple-choice items to measure one aspect and free-response items to measure some other aspect of proficiency; for such tests, we would not expect results such as those reported here. If the free-response items are genuinely intended to measure something different than the multiple-choice items, then a different factor model, such as that used by Bennett et al. (1991) would be appropriate, as would separately-reported scores. As the multiplicity of both the formats and the uses of tests increase, it is important that procedures for item analysis and score-reporting advance apace, so that the reported test scores accurately portray individual differences among those taking the tests.

References

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

Bennett, R.E., Rock, D.A., Braun, H.I., Frye, D., Spohrer, J.C., & Soloway, E. (1991). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151-162.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Browne, M.W. (1982). Covariance structures. In D.M. Hawkins (Ed.), *Topics in applied multivariate analysis* (Pp. 72-141). Cambridge: Cambridge University Press.

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

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 205-214.

Cattell, R. B. (1974). Radial parcel factoring versus item factoring in defining personality structure in questionnaires: Theory and experimental checks. *Australian Journal of Psychology, 26*, 103-119.

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

Cook, L. L., Dorans, N. J., Eignor, D. R., & Peterson, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (RR-85-30). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (RR-87-35). Princeton, NJ: Educational Testing Service.

Dorans, N., & Lawrence, I. M. (1991, November). *The role of the unit of analysis in dimensionality assessment.* Paper presented at the International Symposium on Modern Theories in Measurement: Problems and Issues, Montebello, Quebec, Canada,.

Jöreskog, K.J., & Sörbom, D. (1986). *PRELIS: A program for multivariate data screening and data summarization.* Chicago, IL: Scientific Software, Inc.

Jöreskog, K.J., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications.* Chicago, IL: SPSS, Inc.

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

Wainer, H., Wang, X., & Thissen, D. (1991). *How well can we equate test forms that are constructed by examinees?* Princeton, NJ: Educational Testing Service Program Statistics Technical Report 91-15.

Wainer, H., Wang, X., & Thissen, D. (1993). Can we compare scores on test forms that are constructed by examinees? *Journal of Educational Measurement, 30,* xxx-xxx.

Yen, W.M. (1992). *Scaling performance assessments: Strategies for managing local item dependence.* Invited address presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca., April.