DOCUMENT RESUME

ED 320 911 TM 014 327

AUTHOR Pike, Gary R.
TITLE A Comparison of the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) Exams.
INSTITUTION Tennessee Univ., Knoxville. Center for Assessment Research and Development.
PUB DATE 89
NOTE 29p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; *College Seniors; Comparative Testing; *Construct Validity; Content Validity; Curriculum Evaluation; *General Education; Higher Education; Instructional Effectiveness; Outcomes of Education; Test Reliability; *Test Validity
IDENTIFIERS *College Outcome Measures Project; *Collegiate Assessment of Academic Proficiency; University of Tennessee Knoxville

ABSTRACT

The College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) examinations were evaluated as measures of the effectiveness of the general education program of the University of Tennessee (Knoxville). The criteria used to evaluate these examinations focused on their construct validity. The data were collected during the 1988-89 academic year. Scores on the 60-item COMP examination were obtained from 1,973 seniors and scores on the CAAP examination were obtained from 735 seniors. Approximately 100 of these students took both examinations. To evaluate the content representativeness of the examinations, a seven-member faculty committee rated both tests for their coverage of basic skills goals. Students who took the tests also rated the tests. The evaluation by the faculty indicated that neither test covered even one-third of the general education goals established at the university. The CAAP examination was weaker than the COMP examination in providing useful information in specific content areas. Students were somewhat more favorable than the faculty in their evaluations of the CAAP and less favorable in their evaluations of the COMP. Neither test was judged to be adequately reliable. Both tests had a large standard error of measurement related to the scores and subscores, and both were found to be insensitive to the general education coursework at the university. Nine tables and one graph provide the study data. (SLD)

A Comparison of the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) Exams

Gary R. Pike

Associate Director

Center for Assessment Research and Development

University of Tennessee, Knoxville

1819 Andy Holt Avenue

Knoxville, TN 37996-4350

(615) 974-2350

A Comparison of the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) Exams

Gary R. Pike
Associate Director
Center for Assessment Research and Development
University of Tennessee, Knoxville

During 1988-1989, staff of the Center for Assessment Research and Development located at the University of Tennessee, Knoxville (UTK) conducted a study designed to evaluate the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) examinations as measures of general education program effectiveness. This research parallels a study of the COMP exam and the Academic Profile that was presented to the Tennessee Higher Education Commission (THEC) last year (Pike, 1988).

The Evaluation Criteria

The criteria used to evaluate the COMP and CAAP exams are based on the theories of construct validity developed by Loevinger (1957) and Messick (1989) as applied to educational outcomes measures by Pike and Banta (1989). These criteria focus on three components of construct validity: substantive, structural, and external.

The substantive component of construct validity is concerned with the extent to which test items are accounted for by the construct (Loevinger 1957). Accordingly, an important aspect of the substantive component is the content representativeness of the instruments being evaluated. In addition, the dependability of measurement is a basic part of this component (Messick, 1989). According to Messick, a trade-off frequently occurs between content representativeness and dependability of measurement because attempts to faithfully represent all aspects of a construct frequently introduce items with large errors of measurement into a test. Similarly, efforts to develop highly reliable measures often neglect important content areas.

As its name implies, the structural component of construct validity focuses on the extent to which relationships among test items accurately reflect the structure of the construct (Loevinger, 1957). At one level, this component addresses the appropriateness of the scoring model used to represent the construct. At another level, the structural component focuses on whether relationships among scales and subscales are consistent with relationships assumed by the construct (Messick, 1989).

The final component of construct validity, the external component, is concerned with the extent to which relationships between test scores and other variables are consistent with theories of the construct (Loevinger, 1957). According to Pike (in press), this component is drawn from the concepts of convergence and discrimination proposed by Campbell and Fiske (1959). Specifically, test scores should be related to (converge with)

variables that can be shown to be related to the construct. At the same time, test scores should be unrelated to (discriminate among) variables that are not theoretically related to the construct (Campbell, 1960; Fiske, 1982).

Although the THEC has not provided a clear indication of what the construct is that is represented by the performance funding standard on general education, UTK has defined general education outcomes in terms of a multidimensional-multifaceted construct containing three primary goals and numerous subgoals (Humphreys, 1986). (These goals and subgoals are presented later in this paper in Table 1.) Given this construct, the specific criteria for evaluating the construct validity of the COMP and CAAP examinations are as follows:

(1) The extent to which general education goals and subgoals are represented by test content (substantive);

(2) The reliability of the exams in measuring general education goals and subgoals (substantive);

(3) The appropriateness of the scoring models used in the two tests (structural);

(4) The factor structure of the two tests (structural); and

(5) The extent to which test scores and subscores are sensitive to students' educational experiences and insensitive to background characteristics (external).

The Instruments

Both the COMP and the CAAP exams were developed by the American College Testing Program (ACT) as measures of general education program effectiveness. Because it is difficult, if not impossible, to identify a core of knowledge that is common to general education programs at most colleges and universities, both exams minimize the need to recall specific facts. However, the staff at ACT argues that familiarity with content does improve test performance.

The objective form of the COMP exam takes approximately 2 1/2 hours to administer and contains 60 questions, each with two correct answers (American College Testing Program, 1987). These questions are divided among 15 separately timed activities drawing on materials (stimuli) from television programs, radio broadcasts, and print media. Students taking the COMP exam are instructed that there is a penalty for guessing (i.e., incorrect answers will be subtracted from students' scores) (Forrest & Steele, 1982). The combination of two correct answers for each question and the guessing penalty means that each question is worth 4 points, and the maximum possible score on the COMP exam is 240 points.

In addition to a total score, the COMP exam provides three content subscores (Functioning within Social Institutions, Using Science and Technology, and Using the Arts) and three process subscores (Communicating,

Solving Problems, and Clarifying Values) (American College Testing Program, 1987). It is difficult to determine precisely which constructs these scales are designed to measure because the technical manual for the COMP exam provides only one-paragraph descriptions of the subscales (Forrest & Steele, 1982).

The CAAP examination consists of four separately timed tests (Writing, Mathematics, Reading, and Critical Thinking), each lasting approximately 40 minutes (American College Testing Program, 1988). While the test-taking time for the CAAP exam is 2 hours and 40 minutes, the need to collect and distribute test booklets, as well as to read instructions, means that the actual time required to administer four modules of the exam is in excess of three hours. ACT provides raw, scaled, and percentage correct scores for all four modules. As with the COMP exam, descriptions of the constructs being measured by the modules are quite limited (American College Testing Program, 1988).

The Data

The data for this research were gathered during the 1988-1989 academic year. During the year, useable scores on the COMP exam were obtained from 1973 seniors and useable scores on the CAAP exam were obtained from 735 seniors. Assignment to a testing group was based on two criteria: First, all students who had taken the COMP exam as freshmen were assigned to the COMP testing group as seniors. Second, students who were not tested as freshmen were randomly assigned to either the COMP or the CAAP testing groups. In addition, approximately 100 students agreed to take both the COMP and the CAAP exams. These students were all volunteers and were compensated for their participation.

An examination of the background characteristics of the COMP and the CAAP testing groups reveals that the two groups are quite similar. Slightly more than 51% of the COMP testing group are males and 49% are females. For the CAAP testing group, 56% of the students are males and 44% are females. Although these gender differences are statistically significant, they account for less than 3% of the variance.

Both age and race differences are less noteworthy. The average age of students in the COMP testing group is 23.6 years, as compared to a mean of 23.9 years for the CAAP testing group. Approximately 94% of the students in the COMP testing group are white, and 92% of the students in the CAAP testing group are white.

Differences in ability levels also are relatively minor. The average ACT Assessment score is 21.8 for the COMP testing group, as compared to 21.5 for the CAAP testing group. The mean high school grade point average for students in the COMP testing group is 3.14 and the mean college GPA is 2.87. Mean high school and college grade point averages for the CAAP testing group are 3.09 and 2.90 respectively. Overall, these differences account for approximately 1% of the variance in ability scores.

The Results

Content Representativeness

In order to evaluate the content representativeness of the COMP and CAAP exams, a select faculty committee and students who took the exams independently rated the two tests. Initially, seven faculty members from five undergraduate colleges evaluated the tests using the UTK general education goals and subgoals. The results of these ratings, expressed as percentages are presented in Table 1.

Insert Table 1 about here

An examination of the data in Table 1 reveals that the CAAP exam is somewhat superior to the COMP exam in its coverage of basic skills goals (46% versus 36%). More specifically, the CAAP exam is judged to be superior in its coverage of English composition (50% versus 0%) and computational skills (75% versus 25%). The COMP exam is judged to be superior in its coverage of spoken English (25% versus 0%). Neither test covers the areas of foreign language or computer skills.

In its coverage of specific knowledge (content) components, the COMP exam is judged to be far superior to the CAAP (29% versus 0%). The reason the CAAP exam is rated as not covering any content areas is the absence of any content subscores. Even though the CAAP exam contains questions related to science, history, and the social sciences, the absence of any subscores for these areas makes the evaluation of content outcomes impossible.

In the area of attitudes and judgments, the COMP exam is judged to be slightly superior to the CAAP exam (20% versus 15%). The principal difference in the two exams on this criterion is the greater coverage of values by the COMP exam (50% versus 25%). Neither test covers the subgoals of personal wholeness, life-long learning, or experience in learning.

Overall, the faculty consider the COMP exam to be slightly superior to the CAAP exam (29% to 21%). primarily due to its better coverage of content areas. Most significant, however, is the fact that neither test covers even one-third of the general education goals at UTK.

Students who took the COMP and CAAP examinations also were given the opportunity to rate the exams as measures of general education, basic skills, and critical thinking. The results of these ratings are presented in Table 2.

Insert Table 2 about here

An examination of the data in Table 2 reveals that students tend to be more positive in their ratings of the CAAP exam as a measure of general education and basic skills. Only 17% of the students rate the COMP exam as

an "excellent" or "good" measure of general education, while 25% of the students taking the CAAP exam give that test a good or excellent rating. For both the COMP and the CAAP exams, a substantial proportion of the students give the tests a fair or poor rating (52% and 42% respectively).

Concerning the value of the two tests as measures of basic skills, 35% of the students rate the CAAP exam as excellent or good, compared to 18% for the COMP exam. Nearly 50% of the students rate the COMP exam as a fair or poor measure of basic skills, compared to 28% for the CAAP exam.

Students tend to be somewhat more favorably disposed toward the COMP exam as a measure of critical thinking. Almost 35% rate the COMP exam as excellent or good, compared to 30% for the CAAP exam. Interestingly, 33% of the students give the COMP exam a fair or poor rating, as compared to 29% for the CAAP exam.

In-depth follow-up interviews were conducted with the students who took both the COMP and the CAAP exams. These students were asked to rate each test on four criteria using a scale from 0 (low) to 10 (high). Results indicate that students feel that the CAAP is superior to the COMP exam as a measure of general education coursework (6.51 versus 4.05) and basic skills (7.25 versus 5.05). Interestingly, students who took both exams also rate the CAAP exam as a better measure of critical thinking (6.36 versus 5.71). Only in its ability to measure attitudes and values is the COMP exam judged superior to the CAAP (5.51 versus 4.27).

Reliability

In order to assess the dependability of measurement for the COMP and CAAP exams, several analyses were conducted. The dependability of individual scores on the COMP exam were evaluated using Cronbach's alpha, and CAAP scores were evaluated using a form of the KR-21 formula that assumes homogeneity of item difficulty levels. In addition, group means generalizability coefficients were calculated for institutional means on the COMP exam. Because ACT does not provide item responses for the CAAP exam it was impossible to calculate alpha reliability or generalizability coefficients for this test. Measures of dependability, along with their standard errors of measurement are presented in Table 3

Insert Table 3 about here

An examination of the alpha reliability coefficients presented in Table 3 indicates that while total score on the COMP exam has marginally acceptable reliability (.74), the reliability of subscores is unacceptably low (.44 to .60). These low reliability levels can be attributed to substantial differences in subjects' item response profiles.

The KR-21 coefficients for CAAP scores reveal that three of the tests (Writing [.77], Mathematics [.72], and Reading [.70]) have low, but acceptable, levels of reliability. However, the KR-21 coefficient for Critical Thinking is unacceptably low (.53). What cannot be determined from the

data provided by ACT is the effect of variance in item difficulty levels on KR-21 reliability estimates. As Gulliksen (1950) notes, the KR-21 coefficient used in this research may underestimate reliability where there are substantial differences in the difficulty levels of test items.

In addition to the relatively poor reliability estimates for both the COMP and CAAP exams, standard errors of measurement are quite high. These results strongly suggest that little confidence can be placed in scores and subscores for individuals on these two tests.

Steele (1988) notes that evaluating the effectiveness of general education programs requires the use of group (institutional) means. Accordingly, he contends that estimates of dependability should focus on means, not individual scores, as the units of analysis. Drawing on the work of Cronbach, Gleser, Nanda, and Rajaratnam (1972), Pike and Phillippi (1989) have identified procedures for assessing the dependability of group means using generalizability theory. Using this method, generalizability coefficients were calculated for COMP total score and subscore means.

The data presented in Table 3 indicate that, while the use of group means does improve dependability somewhat (particularly for total score), the dependability of subscore means is still unacceptably low (from .59 to .61). Moreover, standard errors for total score and subscore means are still relatively large.

Scoring Model

Messick (1989) argues that questions about the appropriateness of a scoring model are central to evaluations of the construct validity of an instrument. When an instrument is based on an unusual scoring scheme, providing evidence to support the assumptions underlying that scheme are incumbent on a test developer. In the case of the COMP exam, which uses a cumulative scoring model and item scores ranging from 0 to 4, it is essential that item scores meet the assumptions of a graded model (Samejima, 1969).

In order to test whether the COMP exam meets the assumptions of a graded model, responses to a typical item on Form 8 of the exam were constructed by summing over all 60 items. The summed frequency counts then were analyzed using a polychotomous item response model (Thissen, 1988). Figure 1 depicts the trace lines for the five possible scores (0 to 4) on the typical COMP item. Each trace line represents the probability of receiving a particular score on the item at a given level of ability (theta).

Insert Figure 1 about here

An examination of the graph in Figure 1 suggests that while the scoring method used by the COMP exam meets the requirements of a graded model, it may not accurately represent the ability levels of students. More specifically, although the peak of each trace line representing a

higher score is at a higher level of ability, it is evident that the probability of receiving a score of 0 on the item is greater than or equal to the probability of receiving a score of 1 at all ability levels. Likewise, the probability of receiving a score of either 2 or 4 on the item is greater than or equal to the probability of receiving a score of 3 at all levels of ability.

It is unfortunate that ACT does not provide item responses for the CAAP exam. Because these responses are not available, it is not possible to conclude that the scoring model developed for the CAAP exam is superior to that used for the COMP.

Previous research on the COMP exam and the Academic Profile raises questions about the difficulty levels of the COMP and CAAP exams and the appropriateness of using percentile ranks in a scoring model (Pike, 1988; Pike & Banta, 1989). Table 4 presents scale scores, percentage correct scores, and standard deviations for the scales and subscales of the two tests. Overall means are presented for the COMP and CAAP testing groups, and means for those students taking both tests are included.

Insert Table 4 about here

The data in Table 4 clearly show that the COMP exam is a significantly less difficult test than the CAAP. Percentage correct scores on the COMP exam range from 75% correct to 81% correct for those students who took only the COMP exam and range from 75% correct to 82% correct for students who took both tests. In contrast, percentage correct scores on the CAAP exam consistently are 50% correct for those students who only took the CAAP exam and range from 49% correct to 58% correct for students who took both exams. Although the variability in scores is greater for the COMP exam than the CAAP, the relative variability in COMP scores is quite small given the range of the COMP scoring metric.

The fact that the COMP exam is not a difficult test, coupled with relatively low levels of variability for the exam and the fact that each question on the test can be worth as many as four points, creates a situation in which small changes in student performance can have enormous effects on percentile ranks. For example, a change in students' responses on 1 of the 60 questions (2 responses) would produce a change in the mean total score of 4 points (out of a possible 240 points). This 4 point score change is less than 2% of the possible score but it translates into approximately a 10 percentile point gain or loss for scores between the 30th and 70th percentiles.

While national norms for the CAAP exam are not yet available, the fact that small changes in scaled scores reflect very large changes in percent correct scores (a 1 point scaled score change translates into a 5 to 7 point percentage correct score change) suggests that small scaled score changes on the CAAP also may reflect very large changes in percentile ranks. Here again, undue reliance on percentile ranks runs the risk of making trivial score changes seem significant.

Differences in the scores of students who were single or double tested using the CAAP exam also have important implications for the use of percentile ranks based on national norms. Basically the observed score differences between the single and double tested groups suggest that the CAAP is highly sensitive to student motivation. (Students who took both tests had a mean motivation score of 3.5, as compared to a mean motivation score of 3.1 for students who only took the CAAP exam.) If an institution which tests all of its students is compared to institutions which only test samples of highly motivated students, the institution testing all students may be unfairly disadvantaged simply because some of the students will be less motivated. In effect, national norms may reflect sampling differences rather than differences in the quality of education programs. While this may be true for all standardized tests, it is a particular problem for the CAAP exam because of its sensitivity to the motivation of test takers.

Factor Structure

If general education outcomes are indeed multidimensional, it is critical that the structure of tests like the COMP and the CAAP accurately reflect these outcomes dimensions. In order to determine if the factor structure of either the COMP or the CAAP exams reflects the dimensions of general education outcomes identified as important by UTK, two sets of analyses were performed. First, the COMP and the CAAP exams were analyzed independently using principal components analysis in order to identify their factor structures. Second both tests were simultaneously analyzed using canonical variate analysis in order to determine if there were relationships between the two tests. These two sets of analyses represent forms of internal and external factor analysis (Thorndike, 1978).

Regarding the internal factor analyses, a principal components analysis of the six subscales of the COMP exam strongly suggests that the test is a unidimensional measure. The first principal component produces an eigenvalue of 3.89 and is able to explain 65% of the total variance, while the second principal component produces an eigenvalue of .60 and is able to explain 10% of the variance. Even though a unidimensional structure is clearly indicated, the fact that the second principal component is able to explain 10% of the variance suggests that a two factor solution is worth examination. Therefore, two factors are retained and rotated. Table 5 presents both the unrotated and rotated loadings for these two factors.

Insert Table 5 about here

An examination of the unrotated loadings clearly shows that the first principal component is a general factor with all six COMP subscores having significant positive loadings. The second principal component is a more specialized factor. Only Using the Arts, and to a lesser extent Solving Problems, are positively related to the second principal component. When the factors are rotated, all six subscales are still positively related to the first principal component. However, both Using the Arts and Solving Problems have their strongest loadings on the second principal component.

Clarifying Values also has a strong positive loading on the second principal component, although the strongest loading for this subscale is still on the first principal component.

Substantively, these results suggest that COMP subscores are primarily defined by a strong general factor containing all six subscales. At the same time, there is a much weaker and more specialized factor representing critical thinking in the arts and humanities. While it is possible to debate about the presence or absence of the second dimension, it is clear that neither the one-component nor the two-component solutions provide evidence of measuring the model of general education outcomes at UTK.

Principal components analyses for the four CAAP scores also suggest a unidimensional structure. The first principal component produces an eigenvalue of 2.54 and explains 63% of the total variance, while the second principal component produces an eigenvalue of .82 and explains 21% of the score variance. Once again, two components are retained and rotated because a substantial proportion of the variance is explained by the presence of the second principal component. Both unrotated and rotated loadings for the two component solution are presented in Table 6.

Insert Table 6 about here

Unrotated loadings on the first principal component clearly indicate that this factor is a general outcomes dimension, while loadings on the second principal component suggest that this factor represents a mathematics dimension. When the two components are rotated, the first component seems to represent a verbal dimension with Writing, Reading, and Critical Thinking scores having significant positive loadings on this factor. The second principal component clearly represents a mathematics dimension because only math scores have a positive loading on this factor. Here again, neither the one-component nor the two-component solutions correspond well to the UTK model.

Given the fact that a strong general factor underlies scores on both the COMP and CAAP exams, the question naturally arises as to whether the same dimension is common to both exams. In order to evaluate relationships between scores on the two tests, a canonical analysis was performed. Results indicate the presence of two significant roots. The first root has a canonical correlation of .69 and explains 74% of the variance. The second root has a canonical correlation of .44 and explains 20% of the variance. Table 7 presents the canonical component loadings for the two roots. In this context canonical component loadings represent the correlations of measured variables with the canonical variates.

Insert Table 7 about here

An examination of the canonical component loadings in Table 7 clearly indicates that the first root represents a common general outcomes dimension that transcends both COMP and CAAP scores. All six COMP subscales are

positively correlated with the first COMP variate, and this variate explains approximately 51% of the variance in COMP scores. Similarly, all four CAAP scores are significantly correlated with the first CAAP variate and this variate explains almost 60% of the variance in CAAP scores. Given a canonical correlation of .69 between the COMP and CAAP variates, the proportion of variance (redundancy) in COMP subscores that is explained by the first CAAP variate is .24, and the redundancy of CAAP scores with the first COMP variate is approximately .29 (Stewart & Love, 1968).

If the first canonical root represents a general dimension, the second canonical root represents a very specialized outcomes dimension. The positive correlations of the Using the Arts and Solving Problems subscales with the second COMP variate, coupled with the negative correlation between mathematics scores and the second CAAP variate, suggest a dimension with critical thinking in the arts and humanities at one pole and mathematical ability at the opposite pole. It must be stressed, however, that the power of this relationship is very weak. The redundancy of COMP subscores with the second CAAP variate is .02, and the redundancy of CAAP scores with the second COMP variate is .04.

Sensitivity to Education

As Pike (in press) notes, an important element in determining whether a test is a valid assessment instrument is its sensitivity to the effects of education. Specifically, a valid assessment measure should be more sensitive to the effects of education than to the effects of background characteristics, such as demographics (gender and race), ability (ACT Assessment scores, college GPA, and high school GPA), and motivation when taking the test.

In order to evaluate the sensitivity of the COMP and CAAP exams to education, several stepwise multiple regression analyses were performed. Each of the regression analyses drew on a different COMP or CAAP score as the dependent variable and independent variables consisted of the background characteristics identified above and three measures of general education coursework (calculus, social science, and humanities course taking). These coursework measures represent empirical patterns of coursetaking at UTK that have been used in previous research (Pike & Banta, 1989). The results of the regression analyses for COMP total score and subscores are presented in Table 8.

Insert Table 8 about here

An examination of the data in Table 8 reveals a consistent pattern of results. For all of the COMP scores, ability, as measured by students' ACT Assessment scores, is the primary determinant of performance on the COMP exam, and ACT scores are able to explain from 17% to 37% of the variance in COMP scores. The second strongest predictor of performance on the COMP exam is self-reported motivation when taking the test. Motivation consistently accounts for between 1% and 2% of the variance in COMP scores. Two other background characteristics, college GPA and gender, also are signifi-

cantly related to COMP scores, and these variables each account for approximately 1% of the score variance. (The negative standardized regression coefficient for gender indicates that males perform better than females.)

These results provide little evidence to indicate that the COMP exam is sensitive to educational effects. For only two of the subscales, Using the Arts and Solving Problems, are coursework variables related to COMP scores. In the case of the Using the Arts subscale, humanities coursework is positively related to test performance, but it is able to account for only 1% of the score variance. For the Solving Problems subscale, calculus coursework is negatively related to test performance and is able to explain 1% of the subscore variance.

Table 9 presents the results of the regression analyses for the four CAAP scores. As was the case for the COMP exam, the CAAP does not appear to be sensitive to the effects of education, at least as measured by patterns of course taking. Indeed, coursework variables were not related to any of the CAAP scores.

Insert Table 9 about here

An examination of the data in Table 9 reveals that students' ACT Assessment scores are the primary determinant of performance on the CAAP exam, accounting for between 30% and 33% of the score variance. Self-reported motivation is the second most powerful predictor for three of the four CAAP scales and accounts for between 6% and 9% of the variance in these scores. On the mathematics test, motivation is the fourth strongest predictor and accounts for 1% of the score variance. Gender also is significantly related to performance on the CAAP, accounting for between 1% and 3% of the score variance. (As with results for the COMP exam, positive standardized regression coefficients indicate that females perform better on a test, while negative betas indicate that males perform better on a test.)

Discussion

Based on the results of the present research, it is clear that neither the content of the COMP exam nor the content of the CAAP accurately represents UTK general education goals. A content analysis of the two tests by faculty members reveals that neither exam covers more than 30% of the general education goals established at UTK. The CAAP exam is particularly weak in providing useful information in specific content areas, and both tests provided only limited information about goals related to Understanding Attitudes and Judgments.

Students tend to be somewhat more favorable than faculty in their evaluations of the CAAP and much less favorable in their evaluations of the COMP exam. In part, this may be due to the fact that the structure of the COMP exam is unusual from the perspective of traditionally educated stu-

dents. For whatever reason, approximately half of the students taking the COMP exam give it an unfavorable rating (fair or poor), and almost one-third of the students taking the CAAP exam rate it as fair or poor.

Messick notes that there usually is a trade-off between content representativeness on the one hand and dependability of measurement on the other. In this case, the COMP and CAAP exams seem to have achieved the worst of both worlds. Neither test adequately represents the general education goals of UTK, nor does either provide highly dependable measures of educational outcomes.

The COMP exam does provide a marginally reliable total score, but reliability estimates for all six of the subscores are unacceptably low. Although the use of group means does improve the dependability of scores and subscores on the COMP exam, errors of measurement for the subscores remain too large to allow the use of these subscores in program improvement efforts. For the CAAP exam, three of the scores evidence marginally acceptable reliabilities; however, the reliability of the Critical Thinking scale is very poor.

Perhaps the most troubling aspect of the lack of dependable outcomes measures is the large standard error of measurement associated with the scores and subscores. For example, the 95% confidence interval (± 1.96S.E.) for COMP total score means is approximately ± 6.86, which covers virtually the entire range of point awards for the performance funding standard on general education.

The present research also raises serious questions about the appropriateness of the scoring models for the COMP and CAAP exams. Concerning the COMP exam, this study strongly suggests that the 0 to 4 scoring scheme does not meet the assumptions of a graded model, and these assumptions must be met when using cumulative (additive) scoring procedures with polychotomous data. The results of this research indicate that higher (or lower) item scores are not associated with higher (or lower) levels of the underlying construct (ability).

In addition, the results of the present research also raise questions about the appropriateness of using norm-referenced percentile ranks to make judgments about the quality of general education programs. Given the low difficulty level of the COMP exam and its unusual scoring scheme, the use of percentile ranks to compare programs will inevitably make trivial score differences seem significant. This tendency to inflate minor score differences also may be a problem for the CAAP exam, but a final evaluation of the CAAP scoring model must await the publication of national norms.

One clear problem with norm-referenced comparisons of CAAP scores is the sensitivity of these scores to differences in sampling and administration procedures. The results of the present research clearly show that the scores of a non-random sample of students (e.g., student who volunteer to be double tested and are paid for their efforts) can differ significantly from scores of the population (e.g., all students who are tested).

This problem is underscored by regression results showing that ability and motivation exert an undue influence on CAAP scores. Given the fact

that CAAP scores are extremely sensitive to sampling differences, it is quite possible that norm-based comparisons will reflect differences in how samples are drawn at particular institutions instead of differences in the quality of general education programs at those institutions.

Regarding the factor structures of the COMP and CAAP exams, there is strong evidence to suggest that both tests measure a single dimension. Empirically, the scores and subscores provided by the test developers do not represent independent dimensions of student educational outcomes. Factor analysis results show that both tests can be represented by single dimensions that explain nearly two-thirds of the variance in students' scores. Furthermore, when scores on the two tests are interrelated using canonical analysis, the most significant relationship to emerge is a single dimension that transcends scores on a specific test. In addition, the redundancy of the two tests is substantial.

Using multivariate statistical procedures, such as principal components analysis and canonical analysis, it is always possible to create multidimensional representations of test scores. While the two-dimensional solutions derived from COMP and CAAP scores produce interesting results, they do not correspond to the dimensions of general education at UTK or to the specifications of the test developers.

More specifically, a two-dimensional representation of COMP subscores reveals a strong general outcomes dimension and a much weaker dimension representing critical thinking in the arts and humanities. A two-dimensional representation of CAAP scores presents a different picture. One dimension seems to represent verbal ability, while the other dimension represents mathematical ability. When canonical analysis is used to provide a two-dimensional representation of relationships between scores on the two tests, the first dimension reflects the general factor discussed previously and the second dimension contrasts the two specialized factors identified using factor analysis (i.e., one pole of the dimension is represented by critical thinking in the arts and humanities and the opposite pole is represented by mathematical ability).

The most disappointing result of the present research is the insensitivity of COMP and CAAP scores to patterns of general education coursework. None of the CAAP scores is related to coursework, and only two of the COMP subscales (Using the Arts and Solving Problems) are related to patterns of coursetaking. Moreover, the relationship between solving problems and calculus coursework is negative, and it certainly seems counterproductive to structure a curriculum to discourage advanced mathematics coursework in order to improve scores on a single COMP subscale.

The variables which do predict successful performance on the COMP and CAAP exams are the students' levels of ability when they enter college (ACT Assessment scores), self-reports of how hard they try on the tests, and to a lesser extent, their gender and grade point averages in college. These results suggest that the most effective methods of improving student performance are not to revise the general education curriculum, but rather to select more able students and seek ways to motivate them to try harder on the tests. While the latter may be construed as a reasonable educational goal, the former has serious limitations for a publicly supported institution such as the University of Tennessee.

References

American College Testing Program. (1987). College Outcome Measures Program: 1987-1988. Iowa City, IA: Author.

American College Testing Program. (1988). CAAP fact sheet. Iowa City, IA: Author.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait or discriminant validity. American Psychologist, 15, 546-553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

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: John Wiley.

Fiske, D. W. (1982). Convergent-discriminant validation in measurements in research strategies. In D. Brinberg & L. H. Kidder (Eds.), Form of validity in research (New directions for the methodology of social and behavioral science, No. 12, pp. 105-117). San Francisco: Jossey-Bass.

Forrest, A., & Steele, J. M. (1982). Defining and measuring general education knowledge and skills. Iowa City, IA: American College Testing Program.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley.

Humphreys, W. L. (1986). Measuring achievement in general education. In T. W. Banta (Ed.), Performance funding in higher education: A Critical analysis of Tennessee's experience (pp. 61-72). Boulder, CO: National Center for Higher Education Management Systems.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Messick, S. (1989). In R. Linn (ed.), Educational measurement (3rd ed., pp. 1-45). New York: MacMillan.

Pike, G. R. (1988). A comparison of the College Outcome Measures Program and the ETS Academic Profile (Research Report No. 88-06). University of Tennessee, Center for Assessment Research and Development, Knoxville.

Pike, G. R. (in press). Background, college experiences, and the ACT-COMP exam: Using construct validity to evaluate assessment instruments. Review of Higher Education.

Pike, G. R., & Banta, T. W. (1989, March). Using construct validity to evaluate assessment instruments: A comparison of the ACT-COMP exam and the ETS Academic Profile. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Pike, G. R., & Phillippi, R. H. (1989). Generalizability of the differential coursework methodology: Relationships between self-reported coursework and performance on the ACT-COMP exam. Research in Higher Education, 30, 245-260.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17.

Steele, J. M. (1988, May). Using measures of student outcomes and growth to improve college programs. Paper presented at the annual forum of the Association for Institutional Research, Phoenix.

Stewart, D., & Love, W. (1968). A general canonical correlation index. Psychological Bulletin, 70, 160-163.

Thissen, D. (1988). MULTILOG: Multiple categorical item analysis and test scoring using item response theory. Mooresville, IN: Scientific Software.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner Press.

Table 1
Faculty Ratings of Content Representativeness for the CAAP and COMP Exams

CONTENT AREA	CAAP	COMP
Basic Skills		
Verbal Communication		
Composition	50%	0%
Speaking	0%	25%
Reading	100%	100%
Computation	75%	25%
Foreign Language	0%	0%
Computer Skills	0%	0%
Problem Solving	100%	100%
Total for Basic Skills	46%	36%
Knowledge		
Aesthetics	0%	75%
Science for Life	0%	25%
Techno'ogy	0%	50%
Western History	0%	0%
Foreign Culture	0%	0%
Economics	0%	25%
Social Science	0%	25%
Total for Knowledge	0%	29%
Attitudes/Judgments		
Values	25%	50%
Political Dynamics	50%	50%
Personal Wholeness	%	0%
Life-Long Learning	0%	0%
Experience in Learning	0%	0%
Total for Attitudes/Judgments	15%	20%
Grand Total	21%	29%

Table 2
Student Ratings of Content Representativeness for the CAAP and COMP Exams

CONTENT AREA	CAAP	COMP
Measure of General Education		
Excellent	2%	1%
Good	23%	16%
Satisfactory	33%	30%
Fair	24%	25%
Poor	18%	28%
Measure of Basic Skills		
Excellent	3%	2%
Good	32%	16%
Satisfactory	36%	33%
Fair	21%	27%
Poor	7%	22%
Measure of Logical/Critical Thinking		
Excellent	3%	5%
Good	27%	29%
Satisfactory	41%	33%
Fair	19%	20%
Poor	10%	13%

Table 3
Reliability Coefficients and Standard Errors of Measurement for the COMP and CAAP Exams

TEST/SCALE	RELIABILITY	STANDARD ERROR
CAAP		
Writing	.77	5.07
Mathematics	.72	2.82
Reading	.70	3.51
Critical Thinking	.53	3.29
COMP		
Total Score	.76	7.40
Functioning Within Social Institutions	.54	4.43
Using Science and Technology	.60	3.92
Using the Arts	.45	4.28
Communicating	.55	4.88
Solving Problems	.51	4.54
Clarifying Values	.44	4.90

	GENERALI-ZABILITY	STANDARD ERROR
COMP		
Total Score	.82	3.50
Functioning within Social Institutions	.60	1.99
Using Science and Technology	.60	1.99
Using the Arts	.60	1.99
Communicating	.59	1.96
Solving Problems	.61	2.05
Clarifying Values	.59	1.96

Table 4
Scaled Scores, Percentage Correct Scores and Standard Deviations for the CAAP and COMP Exams

TEST/SCALE	SINGLE-TEST GROUP MEAN	% MEAN	STANDARD DEVIATION
CAAP			
Writing	66.06	49.9	4.12
Mathematics	58.48	50.2	4.29
Reading	64.99	49.8	4.93
Critical Thinking	65.11	50.1	4.24
COMP			
Total Score	190.00	79.2	15.35
Functioning Within Social Institutions	63.50	79.4	5.84
Using Science and Technology	64.70	80.9	5.88
Using the Arts	61.80	77.2	6.37
Communicating	53.80	74.7	7.18
Solving Problems	77.70	80.8	6.05
Clarifying Values	58.30	81.0	5.69

Table 4 (Continued)

TEST/SCALE	DOUBLE-TEST GROUP MEAN	% MEAN	STANDARD DEVIATION
CAAP			
Writing	67.29	58.4	3.20
Mathematics	58.31	49.1	3.58
Reading	66.36	57.8	4.34
Critical Thinking	66.17	57.3	3.47
COMP			
Total Score	191.41	79.8	13.86
Functioning Within Social Institutions	64.11	80.1	5.60
Using Science and Technology	65.44	81.8	5.28
Using the Arts	61.86	77.3	5.97
Communicating	54.00	75.0	6.77
Solving Problems	78.44	81.7	5.43
Clarifying Values	58.95	81.9	5.39

Table 5
Factor Analysis Results for the COMP Exam

FACTOR MATRIX		
	I	II
Functioning Within Social Institutions	.81	-.31
Using Science and Technology	.83	-.23
Using the Arts	.78	.56
Communicating	.82	-.24
Solving Problems	.81	.29
Clarifying Values	.79	-.03

ROTATED FACTOR LOADINGS		
	I	II
Functioning Within Social Institutions	.83	.25
Using Science and Technology	.79	.34
Using the Arts	.26	.92
Communicating	.79	.32
Solving Problems	.46	.73
Clarifying Values	.64	.46

Table 6
Factor Analysis Results for the CAAP Exam

FACTOR MATRIX		
	I	II
Writing	.87	-.18
Mathematics	.52	.85
Reading	.87	-.23
Critical Thinking	.87	-.10

ROTATED FACTOR LOADINGS		
	I	II
Writing	.88	.15
Mathematics	.17	.98
Reading	.89	.10
Critical Thinking	.85	.22

Table 7
Results of the Canonical Variate Analysis of CAAP and COMP Scores

	STRUCTURE COEFFICIENTS	
	I	II
CAAP		
Writing	.83	.15
Mathematics	.48	-.87
Reading	.79	.02
Critical Thinking	.93	.11
COMP		
Functioning Within Social Institutions	.86	-.05
Using Science and Technology	.64	-.07
Using the Arts	.62	.47
Communicating	.94	-.18
Solving Problems	.57	.68
Clarifying Values	.58	-.12

Table 8
Multiple Regression Results for the COMP Exam

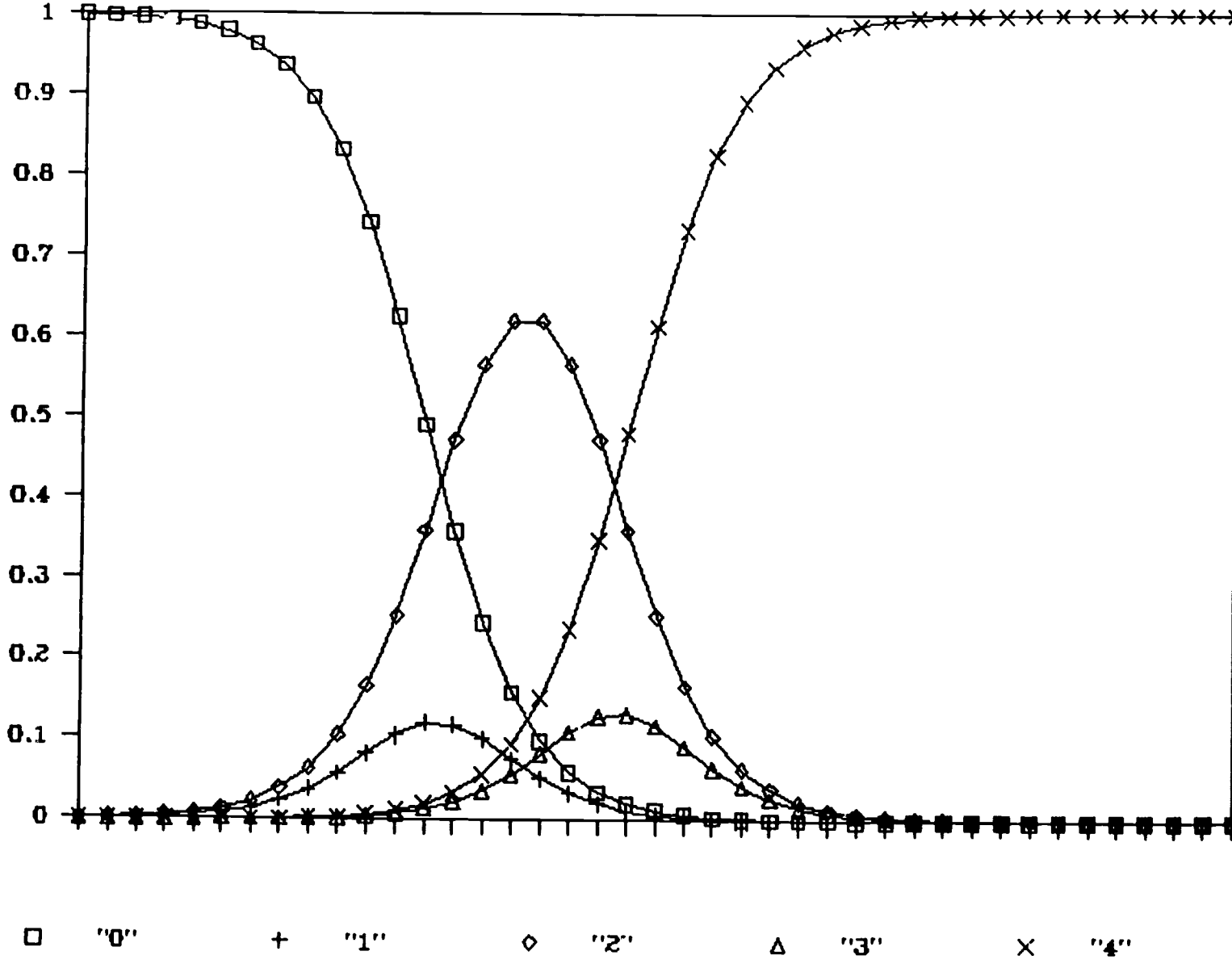
		B	R^2
Total Score	ACT	.61	.37
	Motivation	.15	.02
	College GPA	.08	.01
	Gender (M)	-.07	.01
Funct. Soc. Inst.	ACT	.50	.25
	Motivation	.11	.01
	College GPA	.07	.01
Using Science	ACT	.54	.29
	Motivation	.12	.01
	Gender (M)	-.12	.01
	College GPA	.07	.01
Using the Arts	ACT	.41	.17
	Motivation	.13	.02
	College GPA	.08	.01
	Humanities Coursework	.08	.01
Communicating	ACT	.56	.32
	Motivation	.14	.02
	Gender (M)	-.13	.02
	College GPA	.08	.01
Solving Problems	ACT	.45	.20
	Motivation	.11	.01
	Math Coursework	-.11	.01
	College GPA	.07	.01
Clarifying Values	ACT	.43	.18
	Motivation	.11	.01
	College GPA	.07	.01
	High School GPA	-.07	.01

Table 9
Multiple Regression Results for the CAAP Exam

		B	R^2
Writing	ACT	.58	.33
	Motivation	.27	.07
	Gender (F)	.21	.04
Mathematics	ACT	.57	.33
	High School GPA	.21	.03
	Gender (M)	-.14	.02
	Motivation	.09	.01
Reading	ACT	.56	.32
	Motivation	.30	.09
	Gender (F)	.11	.01
Critical Thinking	ACT	.54	.30
	Motivation	.25	.06
	College GPA	.10	.01

Figure Captions

Figure 1. Trace Lines for Possible Scores on a Typical COMP Item.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
"0"
"1"
"2"
"3"
"4"