The Validity and Usefulness of Three National Standardized Tests for Measuring the Communication, Computation, and Critical Thinking Skills of Washington State College Sophomores

General Report

Report by
Interinstitutional Committee of Academic Officers (ICAO)
Task Force on Assessment and Community College Sophomore Assessment Task Force

May, 1989

This report should be cited as:

Council of Presidents and State Board for Community College Education. (1989, May). *The validity and usefulness of three national standardized tests for measuring the communication, computation, and critical thinking skills of Washington State college sophomores: General report.* Bellingham, WA: Western Washington University Office of Publications.

ICAO Task Force on Assessment

University of Washington

Frederick Campbell
Associate Dean for Undergraduate Studies

Gerald Gillmore*
Director, Educational Assessment Center

Washington State University

Donald Bushaw
Vice Provost for Instruction

Malcolm J. Campbell
Associate Physicist, Laboratory for Atmospheric Research

Central Washington University

Donald Schliesman
Dean of Undergraduate Studies

Gregory Trujillo
Associate Provost and Director of Institutional Research and Assessment

Kenneth O. Gamon
Professor of Mathematics

Eastern Washington University

Phillip Beukema*
Professor of Management, and Chairman,
ICAO Task Force on Assessment

Grant Smith
Professor of English

Frank Kazemek
Professor of Education

Western Washington University

George Mariz
President, Faculty Senate

Robert M. Thorndike*
Professor of Psychology

The Evergreen State College

Steve Hunter*
Director of Research and Planning

Carolyn Dobbs
Academic Dean

Council of Presidents' Office

Judith Gill*
Associate Director, Academic Affairs

Higher Eduation Coordinating Board

Neil Uhlman
Associate Director, Academic Affairs

Sandra Wall
Policy Associate, Academic Affairs

Community College Task Force

Jan Yoshiwara*
Assistant Director, SBCCE

Loretta Seppanen
Systems Information Manager, SBCCE

Barbara Adams*
Executive Vice President
Shoreline Community College

Stephen C. Goetz*
Faculty, Biological Sciences
Shoreline Community College

**Research and Writing Team for the Pilot Study*

Research Associate for Pilot Study
Jacqueline Parker
Western Washington University

Community College Sophomore Assessment Task Force

Big Bend Community College

Robert L. Mason
Dean of Student Services

LeRoy Johnson
Faculty, Math/Physics

Lower Columbia College

Donald E. Fuller
Dean for Instruction

George A. Dennis
Developmental Education Director

Pierce College

Nancy Wallis
Chair, Social Science Division

Seattle Central Community College

Mildred Ollee
Dean of Student Personnel Services

Shoreline Community College

Barbara A. Adams*
Executive Vice President

Stephen C. Goetz*
Faculty, Biological Sciences

South Seattle Community College

Michael McCrath
Division Chair, College Transfer

Joan C. Stover
Division Chair, Science/Chemistry

Tacoma Community College

Frank E. Garratt*
Chair, Community College Sophomore Assessment Task Force
Vice President - Academic and Student Affairs

Paul Jacobson
Faculty, Chemistry

Yakima Valley Community College

Donald Hughes
Dean of Student Services

Council Representatives

Michael J. Grubiak
(Counseling and Guidance Directors)
Associate Dean, Student Services

Joan M. Ray
(Minority Affairs Directors' Council)
Director, Student Assistance Center and Academic Director of Minority Affairs

State Board for Community College Education

Jan Yoshiwara*
Assistant Director

Ron Crossland
Associate Director

David Habura
Deputy Director

Jane Retherford
Secretary to the Deputy Director

Loretta Seppanen
Systems Information Manager

Higher Education Coordinating Board

Neil Uhlman
Associate Director, Academic Affairs

Sandra Wall
Policy Associate

ICAO Task Force

Judith I. Gill*
Associate Director, Academic Affairs
Council of Presidents' Office

Robert M. Thorndike*
Professor of Psychology
Western Washington University

* *Research and Writing Team for the Pilot Study*

EXECUTIVE SUMMARY

The Washington state institutions of higher education have a long-standing commitment to assess student learning and to discern the value of a college education. They agree that assessment helps enhance the quality of programs. Currently, faculty and administrators across the state are involved in discussions and studies to determine the best methods for obtaining appropriate and useful information from assessment activities.

Some states use standardized tests to measure the academic performance of students. In its master plan (December, 1987), the Higher Education Coordinating (HEC) Board recommended that both two-year and four-year institutions conduct a pilot study to evaluate the appropriateness of using standardized tests as a means for measuring the communication, computation, and critical thinking skills of sophomores. The purpose for such a testing program would be for institutions to: (a) strengthen their curricula, (b) improve teaching and learning, and (c) provide accountability data to the public.

To design and implement the study requested by the master plan, two task forces were established; one representing the public baccalaureate institutions and one representing the community colleges. Both task forces included faculty and academic administrators from each participating institution and two HEC Board staff members. The two task forces worked in parallel and ultimately conducted a joint study.

Only three tests met the criteria of the HEC Board recommendation for study: the Academic Profile (AP), the College Outcome Measures Program (COMP), and the Collegiate Assessment of Academic Proficiency (CAAP). Over 1,300 sophomore students from the public four-year institutions and from eight two-year colleges were tested, with each student taking two of the three tests. More than 100 faculty members from the same institutions took shortened versions of the tests and critiqued them for appropriateness of content and usefulness.

The results of the pilot study strongly suggest that the three tests do not provide an appropriate or useful assessment of the communication, computation, and critical thinking skills of Washington college sophomores:

- None of the tests studied measured the separate academic skills (communication, computation, and critical thinking). Rather, these tests primarily measured verbal and quantitative aptitude.

- The tests added little reliable new information about students' academic performance. Results essentially reiterated what is already known from admissions test data and student grades.

- Test scores were not sensitive to specific aspects of the college experience, such as estimated time spent studying and credits earned.

- None of the tests was judged by faculty to provide an adequate match with curricular content or as being an appropriate or useful measure of communication, computation and critical thinking.

- Norms for making comparisons with peer institutions are currently unavailable. Furthermore, student performance is affected by differences in the manner in which institutions administer tests, the timing of testing, the selection of students, and the students' motivation. Thus, comparisons with future norms based on tests given under differing conditions will be misleading.

Analyses of costs associated with conducting the pilot study suggest that the projected expense associated with state-wide implementation, either by testing a sample of sophomores or all sophomores, would be high and would likely exceed the value of the results.

Both two-year and four-year faculty participants in the study recognized the importance and value of having public as well as institutional access to appropriate measures of student performance. They reaffirmed the value of assessment activities for strengthening the curriculum, improving teaching and learning, and enhancing overall instructional quality. They also shared the view that the development of meaningful assessment measures is both difficult and time consuming, that measures should be institution-specific, and that national standardized multiple-choice tests have serious limitations in the assessment of teaching and learning.

INTRODUCTION

Quality in American higher education has been an enduring concern for over 200 years. Throughout their history, institutions of higher learning have dealt with effectiveness issues -- the achievement of their students, the contributions of their graduates to the society, and the satisfaction of their alumni with their college experience. The assessment of this effectiveness has been and remains an ongoing, integral component of the teaching/learning process and of institutional activity.

Recently, three major national reports have critically addressed the issue of quality in American higher education. These reports and other observers have challenged higher education institutions to demonstrate that they are making a difference in the intellectual and personal development of their students. Regional accrediting agencies, for example, which traditionally have based their evaluations of colleges and universities on institutional resources and processes, are adopting new criteria for institutional excellence and establishing standards which include assessment of educational outcomes.

State boards of higher education have recognized the need to address the effectiveness of their public higher education systems. Most state boards believe their primary responsibility to be one of ensuring system quality and providing accountability, but they recognize that responsibility for designing assessment systems should rest with the individual institutions.

Washington's Agenda

Discussion of assessment as a policy issue began in Washington in 1986 with the Higher Education Coordinating (HEC) Board's preparation of a master plan for higher education. Arguments supporting greater financial autonomy for the institutions were balanced by statements calling for accountability in the use of state dollars. Proponents of a state-mandated assessment program argued that student performance data would provide students, legislators, and the public with important information on how well institutions and the system as a whole are performing. This information, they asserted, is essential to gaining increased funding for high-quality higher education.

The State Board for Community College Education (SBCCE) and the Council of Presidents (COP), representing two-year and four- year institutions respectively, publicly supported assessment. Their statements of support stressed that assessment had always been an integral component of academic management practices.

In late 1987 the HEC Board adopted its master plan, entitled "Building A System." It includes recommendations providing for greater management flexibility and increased accountability for the institutions. The recommendation on perfor-

mance evaluation included a requirement for (a) institutionally designed assessment programs and (b) a statewide pilot study on the use of nationally normed standardized tests. This paper reports the results of that study.

Study Planning

In August, 1987, the HEC Board, together with the Interinstitutional Council of Academic Officers (ICAO) and the State Board for Community College Education (SBCCE), established two task forces--one for the baccalaureate institutions and one for the community colleges. Both were composed of faculty and administrators from each participating institution and HEC Board staff members. The two task forces also shared members to assure comparability of the studies and to provide a mechanism for sharing ideas and developments.

The charge to each task force was to conduct a study of the usefulness and validity of nationally normed tests of communication, computation, and critical thinking skills to be administered to students in the last term of their sophomore year. The two task forces worked in parallel and ultimately conducted a joint study.

This report contains a brief description of the study, a summary of the general findings, and the broad conclusions to which those findings lead. It includes the methodology, student and faculty results, and a discussion of the feasibility of using standardized tests for the purposes envisioned in the HEC Board master plan. A complete description of the study, including data supporting the conclusions, is available in the technical report.

DESIGN OF THE STUDY

Overview

Published tests of academic achievement were examined, and only three met the HEC Board criteria for measuring the communication, computation, and critical thinking skills of students at the end of their sophomore year. All three tests were used in the pilot study, which was conducted at all public four-year institutions and eight two-year colleges. Students from the pilot institutions took the tests and responded to a questionnaire. Faculty took portions of the three tests and critiqued the tests through a questionnaire and group discussions. Important research questions addressed by the study included:

1. The relationship of the tests to the curriculum;
2. The amount of new information gained as a result of the test data;
3. The relationship of the test data to other measures of academic performance;
4. The relationship of the test data to students' college experience; and
5. The usefulness of the test results for curriculum improvement and student advising.

Methods

Tests. The three tests used in the study were the College Outcome Measures Program (COMP), the Collegiate Assessment of Academic Proficiency (CAAP), and the Academic Profile . The COMP, published by the American College Testing Program (ACT), is intended to measure knowledge and skills related to successful functioning in society. It includes questions based on passages from popular magazines and clips from video and audio presentations. It has been used by over 350 institutions to help assess various aspects of college programs and has been available for about 10 years.

The CAAP, also published by ACT, and the Academic Profile, published by the Educational Testing Service (ETS), were still under refinement when this study was conducted. Both tests were intended to measure student achievement in the foundational skills of reading comprehension, writing, mathematics usage, and critical thinking. The questions on the AP and the CAAP were more academic in content than the questions on the COMP, with many being preceded by a written passage to provide the context.

All test materials were of high technical quality. Test items were generally clearly worded and the wrong alternatives were plausible. Publishers' analyses of data from national samples indicate that all of the tests have adequate reliability. The COMP was the only instrument that provided results in a national normative form (percentile ranks based on institutions that have used the COMP) at the time of this study.

In addition to the three tests, a one-hour writing essay based on writing prompts provided by ACT was administered. This measure was intended to evaluate students' writing skills, but technical problems with the scoring criteria used by the publisher resulted in data that were not useful in the present study. Thus, results of the writing essay are not included in the following discussion.

Student Samples. Students were selected for participation in the study according to the following criteria:

> Baccalaureate students must have completed between 75 quarter (45 semester) hours and 105 quarter (60 semester) hours, with no more than 15 quarter (10 semester) transfer credits.
>
> Community college students must have completed at least 70 college-level quarter hours in an associate degree transfer program, with no more than 15 credits earned outside the state community college system.

Students who met the criteria were sent a letter inviting them to participate in the study. The letter stressed the importance of the study and students were offered $25 or $35 for their participation, depending on the time involved. Letters of invitation were sent to 5752 students, of whom 1302 (23%) volunteered and were tested.

Because completing all tests required eight hours, individual students were asked to take only two of the tests. About one-third of the students took the COMP and the CAAP, another third took the COMP, the Academic Profile, and the writing sample, and the final third took the CAAP, the Academic Profile, and the writing sample. Following each test, students responded to questions about test difficulty and content and the seriousness of their own effort. Student background data were obtained from the registrar's records at each institution.

Faculty. The study also sought faculty perceptions about the validity and usefulness of the tests used. Twelve faculty members from each of the four-year institutions and six from each of the participating community colleges were selected. All faculty had a demonstrated commitment to general education and taught in undergraduate programs. Special care was taken to include writing and mathematics faculty as well as scientists and philosophers with special expertise in critical thinking.

The faculty took shortened versions of the tests and responded to questions about test validity -- what the tests measured and the appropriateness of the content for assessing student performance in communication, computation, and critical thinking. They then met to discuss the test validity in a group setting. Summaries of these discussions were prepared.

In a second session, about half of the faculty participants from each institution (two-year and four-year) met together to hear presentations from testing company representatives about the qualities and uses of their tests. Also at this session a nationally recognized expert on assessment presented an overview of current issues in the area. Faculty then independently completed questionnaires about the usefulness of the tests and finally met in groups with trained focus-group leaders for discussion. Summaries of these discussions were also prepared.

STUDENT RESULTS*

Test Difficulty

The question of whether the tests were of an appropriate level of difficulty, neither too easy nor too difficult for sophomore students, was approached from three different perspectives:

> First, in terms of student performance, average scores on the various scales ranged from 50 to 75 percent correct, indicating that the levels of difficulty were appropriate.
>
> Second, student responses to the questionnaires indicated that few found the tests either very easy or very difficult. The majority of the students felt they had adequate time to complete all tests.

**Unless otherwise indicated, reported results refer to students in both two-year and four-year institutions.*

Finally, faculty reviewers from the participating institutions also rated the difficulty level of each instrument. While there were differences among faculty, they generally rated the difficulty of the tests as slightly on the easy side of "about right."

All three lines of evidence--student performance, student perceptions, and faculty perceptions--indicate that the tests appear to be of an appropriate level of difficulty for Washington two-year and four-year sophomore students.

What New Information Do the Tests Provide?

A major reason for using any test of student performance is that it will provide new or better information than is available from existing sources, or that information can be obtained in a more convenient or economical manner. A decision to administer standardized tests should presuppose that such test results will yield meaningful information about student performance that is different from that already available from other sources and that is in a form that will be useful both to institutions and to policy makers. To examine the ability of these tests to provide new information, a series of multiple regression analyses was performed. The specific purpose of these analyses was to determine the overlap of background characteristics (demographic data such as age, sex, and ethnicity and academic achievement data such as Washington Pre-College (WPC) Test scores and grade- point averages) with student test results.

For four-year college students, analysis of the Academic Profile showed that about 65 percent of the variability in students' total scores was explained by differences in their background characteristics and by their self reported effort in taking the test. For the CAAP scales from 45 to 60 percent of student variability was explained by these factors. For the COMP about 50 percent of the differences in total scores was explained by the students' background characteristics.

Because community college students were not as likely to have taken the Washington Pre-College Tests, WPC scores were available for only about one-third of the two-year student sample. The results of multiple regression analyses for this subgroup of students were similar to those of the four-year students. When the general background variables (excluding the WPC) were analyzed with the total sample of two-year students, the major predictors of test scores were college GPA and immigrant/refugee status. The presence of the latter variable as a major predictor is almost certainly due to the fact that the two-year sample, unlike the four year sample, contained a number of immigrants with a limited facility in English.

In summary, *the results of these analyses indicate the Academic Profile, the CAAP, and the COMP added relatively little reliable new information about students.*

Relationship of Test Scores to College Experience

To be useful for planning and evaluation, test results must relate to the specific aspects of the college experience. Student test scores were compared to five dimensions of the college experience.

Number of credits earned. It is reasonable to expect that students have who completed more credits, and hence a greater proportion of their undergraduate education, would receive higher scores on the tests. The study did not support this expectation. *Correlations of credits earned with test scores tended to be small and negative.*

Completion of general education requirements. At two-year and four-year institutions, distribution requirements in the humanities, social sciences, and natural sciences are the foundation of general education. These courses should foster development of the cognitive skills that the CAAP, the Academic Profile, and the COMP claim to measure. Accordingly, test scores should correlate positively with the number and percent of distribution credits completed. This expectation was not confirmed. *Correlations were uniformly small and often negative.*

Grade point average (GPA). The study found substantial positive correlations between test scores and GPA's. However, GPA is also correlated with academic aptitude as measured by the Washington Pre-College Test (WPC). To determine the degree to which the relationship between GPA and the sophomore test results was due to their common association with the WPC, a partial correlation analysis was employed. Results of the analysis indicated that the relationship between grades and test scores diminished considerably once the variability both had in common with the WPC was removed. Thus, *the correlation of college grades with the skills test scores was due mainly to the influence on both variables of general academic aptitude as measured by college admissions tests.*

Student reports of academic effort. It is also reasonable to anticipate that test scores might be positively related to students' academic effort as indicated by their reports about hours spent on studying or in the library. These relationships were not supported; correlations between test scores and these two variables were negligible.

Student perception of college influence on test performance. To examine the effects of college on their test performance, students were asked to estimate how they would have performed had they taken the test right after high school. *On average students reported that their present performance was probably a little better than it would have been immediately after high school.*

What Do the Tests Measure?

The HEC Board Master Plan Recommendation (Chapter 4, p. 27) emphasizes assessment of the three skills: communication, computation, and critical thinking. If communication can be subdivided into reading and writing, then both the CAAP and the Academic Profile have scales that purport to measure the skills called for in the recommendation. The COMP offers scales purportedly measuring communication, computation, problem solving, and clarifying values, as well as additional scales that measure knowledge and skills relevant to functioning in adult society.

One way to address the question of what the scales measure is to examine the correlations among them. One would expect like-named scales from different tests to correlate highly with one another.

> Analyses showed that each CAAP scale had a correlation of 0.60 to 0.65 with the Academic Profile scale of the same name (e.g., CAAP Critical thinking with Academic Profile Critical thinking). Because the Academic Profile scales are relatively short (12 items) and are therefore of modest reliability, these correlations can be considered quite high. Thus the two instruments seem to be measuring the same skill areas.

However, the correlations among the supposedly different scales within both instruments (e.g., Academic Profile Critical Thinking with Academic Profile Writing) were too high to be consistent with a hypothesis that each skill, as measured, represents a separate dimension. In addition, although the COMP scale labels suggest that they measure quite different things, the correlations among those scales were also very high.

In order to examine the underlying structure of the tests, selected sets of scales were submitted to factor analysis, a statistical technique that determines common underlying "factors" or dimensions.

> Results for the CAAP and the Academic Profile indicated that both tests measure two factors: a verbal ability factor composed of the Writing, Reading, and Critical Thinking scales, and a quantitative factor consisting of the Math scale alone for the CAAP and the Math scale and a portion of the Critical Thinking scale for the Academic Profile. Thus, neither the CAAP nor the Academic Profile measured the separate skills identified by its scale labels.
>
> Similarly, the factor analysis of the COMP revealed a quantitative component, as well as a values component that is not addressed by either the Academic Profile or the CAAP.

A second series of factor analyses was performed responding to the unique opportunity made available because students took pairs of tests.

> Factor analysis of the Academic Profile and the CAAP combined resulted once again in a large verbal ability factor and a smaller quantitative factor that has some small relationship with critical thinking. The substantive conclusions are identical to those above.
>
> Pairing the COMP with the other tests showed that it measures something similar to the quantitative dimension of the CAAP or the Academic Profile. The definite and independent values component remained but was unrelated to the academic skills that the Academic Profile and the CAAP claim to measure. The COMP did not share a verbal component with the CAAP or the Academic Profile.

Based on the factor analyses reported above, it appears that *none of the tests studied actually measured the separate academic skills identified in the HEC Board Master Plan.* In particular, it has proven to be almost impossible to separate reading, writing, and critical thinking. All seem to be part of a general verbal

academic ability that probably is largely reading comprehension. An analysis of the test tasks is consistent with this interpretation because the reading, writing, and critical thinking scales all involve reading passages of text and answering questions based on an understanding of what was read.

Student Perceptions of Tests

On the post-test questionnaire, students were asked to judge how well the test they had just completed measured a number of skills, including communication, computation, critical thinking, and general education and other skills purportedly measured by the tests.

Students gave the highest average ratings to all three tests' ability to measure reading comprehension. In general, they also gave fairly positive ratings to the tests' ability to assess critical thinking.

In contrast, the tests' measurement of communications (writing and/or speaking) ability and overall general knowledge received lower ratings.

Student Motivation

As an incentive to participate in the study, students were offered monetary compensation for their time. It was hoped that the payment, along with the letter of invitation, the introduction to the testing session, and the feedback of individual results, would provide adequate motivation for students to perform at or near their best. Several lines of evidence suggest that a high proportion of the examinees took the tests very seriously and worked diligently to achieve their best performance.

First, observers reported that all of the students appeared to be working conscientiously.

Second, very few test scores were so low as to suggest random responding.

Finally, over 92 percent of the students indicated that they either "tried my best on every item" or only "occasionally guessed rather than thinking hard." The majority of the students chose the former response.

FACULTY RESULTS

Research and experience in the assessment of college student performance have demonstrated consistently that faculty acceptance of and involvement in such efforts are absolutely critical for their success. Thus, the present study gathered faculty opinion on the validity (appropriateness) and usefulness of the tests reviewed The emphasis was on assessment of the instruments as measures of sophomore-level skills in communication, computation, and critical thinking, the three areas explicitly cited in the HEC Board Master Plan. General or liberal education was added as a fourth major category of evaluation, since all institutions include this as a major goal in the education of lower-division students.

The following sections describe faculty opinions of the validity and usefulness of the instruments, followed by other concerns which emerged during faculty discussions. As above, results pertain to both baccalaureate and community college faculty unless otherwise noted.

Validity

Communication. None of the tests received a positive overall rating as a valid measure of communication skills. At best, 46 percent of the faculty rated one of the tests an appropriate measure.

> The Academic Profile and the CAAP tended to be rated more favorably than the COMP.
>
> The content of the reading section of the Academic Profile was viewed more favorably than the equivalent section on the CAAP.
>
> The CAAP writing section was viewed somewhat more favorably than that of the Academic Profile.
>
> Many faculty viewed the sections which purportedly measure writing to be measures of editing skills rather than of the skills necessary to produce good written communication.

Computation. The CAAP was the only test to receive a positive rating of overall validity as a measure of computational skills by a majority of the faculty.

> Opinions were mixed regarding the presence of test items requiring calculus in the CAAP.
>
> There was a tendency for two-year faculty to rate the CAAP's level of difficulty as too great.

Critical thinking. A majority of the faculty rated the Academic Profile (68 percent) and the CAAP (56 percent) positively with regard to their overall validity. Only 35 percent rated the COMP favorably.

> Faculty ratings suggested a higher level of satisfaction with both the Academic Profile and the CAAP as measures of critical thinking rather than as measures of communication and computation.
>
> Faculty preferred the Academic Profile over the CAAP. The appropriateness of the Academic Profile content and its overall validity as a measure of critical thinking were judged more favorably by the faculty than were the similar features of the CAAP.
>
> Despite higher satisfaction with the Academic Profile and the CAAP as measures of critical thinking, faculty comments indicated a range of concerns relating to a definition of critical thinking that they judged to be too narrow.

General Education. No test was rated by a majority of the faculty as a valid measure of general education.

For each test, a significantly smaller proportion of faculty from four-year institutions provided favorable ratings as compared to two-year faculty.

Four-year faculty tended to regard the range of skills and the methods employed for measuring these as too narrowly defined.

Usefulness

Improvement of curriculum and specific courses. At best, 50 percent of the faculty rated test results of any test in any area "very" or "moderately" useful to the review and evaluation of the curriculum and specific courses. None of the tests received a strong endorsement by faculty.

Of the three tests, the CAAP was rated most useful.

The absence of a clear relationship between test results and the curriculum was viewed as the most serious limitation to usefulness.

Need for additional student work. For all tests, fewer than 25 percent of the faculty provided ratings of "very" or "moderately" useful for determining the need for additional coursework by students.

Institutional effectiveness. The CAAP was the only test to receive favorable ratings from a majority of the faculty reviewers; these favorable ratings occurred for computation and critical thinking.

The COMP and the Academic Profile received a favorable rating by no more than 30 percent of the faculty on any dimension.

Faculty reservations about the use of these tests as measures of institutional effectiveness were grounded in their inability to see a clear relationship between test results and the curriculum.

Other Considerations

Legitimacy of assessment efforts. There was agreement in each faculty discussion group that the state had a right to seek information documenting the effectiveness of the higher education system. While they had serious reservations about the three tests, faculty supported the legitimacy of the larger goal of institutional assessment.

Test bias toward non-native speakers of English. Faculty noted that reading comprehension and speed seemed essential to high performance on the tests. In fact, the tests were viewed as largely measures of reading comprehension. Many observed that non-native speakers of English would be placed at a distinct disadvantage because their reading speed was slow in comparison with native speakers. Faculty recommended that this factor be taken into account if any of the tests were implemented. Also, faculty observed some insensitivity to diversity in culture and gender.

Motivation of students. The difficulty of ensuring high-quality student effort on the tests was cited routinely. If the student had no stake in the test results, faculty believed that quality of effort would suffer.

Use of results. It was not clear to faculty how test results would be used. The absence of clear plans for the use of results and the potential adverse consequences to institutions caused apprehension.

Alternatives to standardized tests. Consistent with the faculty opinions that (a) the state deserves measures of institutional accountability, and (b) the measures reviewed fell short of the mark, some faculty group discussions turned to an exploration of alternatives to the tests. Development of institutionally-focused measures rather than measures standardized across institutions was a dominant theme. Also, faculty predicted that development of meaningful measures would be difficult and time-consuming.

FEASIBILITY

The feasibility of testing sophomores within the Washington State system of higher education was examined based on the costs of conducting this pilot project, the reported costs of various testing programs, the availability of national norms, and the experiences of other states engaged in statewide assessment programs.

Cost of the Pilot Study

The total cost of the pilot study was estimated at $396,000 -- $113,000 for travel, testing expenses, payments to students, and so forth, and $283,000 for faculty and staff time. Of the total, the HEC Board contributed $8,000 for the purchase of tests, and the remainder was contributed by the cooperating institutions, the State Board for Community College Education, and the Council of Presidents.

Projected Implementation Costs

The cost of implementation is difficult to predict because the parameters of a testing program are not defined. However, estimated per-student expenditures are provided based on two hypothetical plans. First, if all students in participating institutions were tested, using a single test which includes a writing sample and a locally developed questionnaire, the cost would be approximately $24.00 to $31.50 per student, depending upon the test chosen. If such a program were implemented across the state for all two-year and four-year college sophomores, the total annual cost would be over one-half million dollars. This estimate assumes that students' participation would be a result of an institutional testing requirement and that payment for participation would not be necessary.

Second, if a representative sample of students were tested rather than all students, the cost per student would increase because student compensation probably would be necessary to elicit participation. Furthermore, expenditures would be necessary

for student recruitment. Using the same student compensation rates as were used in the current study and the same package as outlined above, the average cost would be $56.00 to $73.50 per student, depending upon the test chosen.

It is especially difficult to project a total cost for testing a representative sampling because the size of the sample will depend upon the intent of the assessment. A conservative estimate for a sample might be that 100 students will be tested at each two-year school and 200 students at each four-year school. The total annual cost under these assumptions would be approximately $175,000 to $200,000.

Student Motivation

Student effort represents a major, unresolved problem in the assessment process. Experience from other states has demonstrated that students must be externally motivated to perform at their best. *If students are not motivated, the results of assessment will be seriously flawed, yielding misleading conclusions and loss of credibility.*

For this study the evidence indicates that a large proportion of the students were reasonably motivated to do their best. However, only one- fourth of the students who were invited actually took part in the study. This low response rate occurred in spite of the monetary incentive. Had there been no payment, few students would have volunteered, and those who did would not have been representative of all college sophomores. Furthermore, based on experience from other states, cooperation would have been low and results meaningless if students had been conscripted into participation.

Test results from assessment activities outside of the normal classroom processes must be meaningful to individual students for the data to provide useful information to the institution. Requiring a test of the whole student body, for example, does not solve the motivation problem unless performance on the test makes a valued difference to the students. Administering a test to a sample of students is particularly troublesome because offering either reward or negative consequences based on test outcome involves the question of fairness relative to those students not in the sample. In addition, solutions to the motivation problem can have direct and indirect effects on the college's curriculum and requirements. Thus, any consideration of feasibility must give serious attention to the issue of student motivation.

Comparative Norms

One possible use for a sophomore-level testing program is to provide data which would allow comparisons with peer institutions. Unfortunately, none of the tests under investigation currently has an adequate norm base for making these comparisons, including the COMP which has had a decade of use. The COMP norms are derived from the performance of students at institutions where the test is used. The basis of these norms changes at varying academic levels because different institutions test students at different points in their careers. In order for such comparisons to be valid, peer institutions would have to adopt the same test and testing conventions as those used by the State of Washington.

Specifically, the method by which students are selected, the time at which students are tested, and, especially, the motivation of the students can seriously affect the results. For example, both two-year and four-year students tested in the current study had higher average COMP scores than seniors at an eastern state research university, and much higher average scores than the entire community college and baccalaureate norm base, respectively. In interpreting these higher averages, it is not possible to separate the effects of the quality of the education the students received, the average motivation of the students tested at each institution, and the method whereby students were selected for testing. It is very unlikely that a sufficient number of peer institutions will adopt a program sufficiently like one that might be adopted in Washington to provide an adequate normative base.

CONCLUSIONS

The specific purpose of this study was to examine the validity and usefulness of three standardized examinations for assessing communication, computation, and critical thinking abilities of Washington State college sophomores. The study focused on three national, standardized tests, the CAAP, the Academic Profile, and the COMP, which at the time of the study were the only tests available for the purpose.

Data were obtained from tests and questionnaires from over 1300 students and from questionnaires and group discussion with over 100 faculty members. While many of the attributes of the tests themselves were judged as positive and a wealth of detailed information emerged, the results of the study casts doubt on the validity and usefulness of the three tests under investigation and of nationally standardized tests in general for the purposes guiding this study.

The tests failed to meet reasonable standards for validity and usefulness for several interrelated reasons. At the most basic level, the tests appear largely to measure the general academic ability of students. At best, they measure the same verbal aptitude or intelligence and quantitative aptitude or intelligence that the college admissions tests measure. Administering the sophomore-level tests appears to be effectively equivalent to administering college admissions exams a second time, although the sophomore-level tests are somewhat more difficult. Furthermore, because the test scores would likely be highly related to the entering abilities of students, these instruments would serve more as a measure of the quality of incoming students than as an evaluation of the college's instructional programs.

Given that the tests appear to measure only general aptitude, the test results would be of limited value for curricular planning and evaluation. Furthermore, the study failed to find any empirical evidence of a relationship between test results and college experiences. Thus, test results would be unlikely to provide clues as to where improvements might be needed and would be unlikely to detect gains in student learning if improvements were made.

In addition, the faculty participants tended to feel that the results would not help them in planning or evaluating the curriculum or their specific courses. There was a shared perception that the tests failed to measure the goals and content of the curriculum. This perception was most strongly felt in the area of writing but held true for other areas as well. Because faculty did not feel that the tests measured important outcomes of the curriculum, test results would not be useful and would lack credibility.

The test results would not be useful for the second major purpose stated in the master plan, external accountability. The absence of meaningful normative information makes it highly unlikely that such test results can be used to compare the state's higher educational institutions with their national peers. Such norms do not exist and are unlikely to exist in the future because there is no reason to expect a significant number of state systems or individual campuses to adopt a common test. Even with an adequate norm base, the questions of student motivation, differences in the methods of selecting students for testing, and the timing of the testing still might render such comparisons meaningless.

Although of very high quality, the reason the tests examined in this study failed to meet the HEC Board's criteria is probably inherent in the development of national standardized tests. These tests must be constructed to satisfy a national audience; thus, the content which can be included is necessarily general and comes to resemble measures of basic aptitude or intelligence.

Furthermore, to equalize the content knowledge of students who come to the tests from a wide variety of experiences, test developers commonly provide the examinees with written situations or contexts, followed by a number of derivative questions. The unfortunate side-effect is that a very high premium is placed on reading comprehension -- the ability to read and understand a passage quickly. Students in this study judged that the tests measured reading comprehension more effectively than any other ability, and faculty consistently criticized the test for that characteristic. There was also the concern, validated by the community college test results, that the emphasis on reading would hurt the performance of students for whom English is a second language.

The state's colleges and universities do not view the teaching of reading comprehension as part of their responsibility, other than remediation. They expect students to have already acquired this skill. Hence, faculty understandably felt this particular emphasis was inappropriate.

For the reasons stated above, any nationally developed and standardized test, other than those measuring very specific, narrowly defined abilities, is likely to be disappointing in its validity and usefulness at the college level. However, both two-year and four-year faculty participants in the study recognized the importance and value of having public as well as institutional access to appropriate measures of student performance. They reaffirmed the value of assessment activities for strengthening the curriculum, improving teaching and learning, and enhancing overall instructional quality. They also shared the view that the development of meaningful assessment measures is both difficult and time-consuming, that measures should be institution-specific, and that national standardized multiple-choice tests have serious limitations in the assessment of teaching and learning.