CU
NY
The City
University
of
New York

Report of the CUNY Task Force On System-Wide Assessment Of Undergraduate Learning Gains

August 2011

Table of Contents

Executive Summary

In January 2011, the CUNY Task Force on System-Wide Assessment of Undergraduate Learning Gains (Assessment Task Force) was convened by Executive Vice Chancellor Alexandra Logue and charged as follows:

> *The Chancellery wishes to identify and adopt a standardized assessment instrument to measure learning gains at all of CUNY's undergraduate institutions. The instrument should be designed to assess the ability to read and think critically, communicate effectively in writing, and measure other learning outcomes associated with general education at CUNY. It must be possible for each college and the University to benchmark learning gains against those of comparable institutions outside CUNY. It is the responsibility of the Task Force to identify the most appropriate instrument and to advise the Chancellery on how best to administer the assessment and make use of the results.*
>
> *The Task Force is charged with the following specific responsibilities:*
>
> 1. *Taking into account psychometric quality, the alignment of the domain of the instrument with broad learning objectives at CUNY colleges, cost, facility of obtaining and using results, and the ability to benchmark results externally, select an assessment instrument from among those commercially available at this time.*
> 2. *Develop recommendations for the chancellery on how the assessment should best be administered so as to*
>   a. *represent each college's undergraduate student body;*
>   b. *generate a valid assessment of learning;*
>   c. *facilitate comparisons across CUNY colleges and between CUNY and other postsecondary institutions.*
> 3. *Develop recommendations on how the colleges and the chancellery can best use the results to improve teaching and learning throughout CUNY.*

This report primarily addresses the first component of the charge—selection of an assessment instrument. A companion report will present recommendations for administering the instrument and using the results to improve undergraduate education at CUNY. This report details the process by which the Assessment Task Force defined the cognitive abilities that the test should measure, identified a rubric for measuring those abilities, developed criteria for selecting the most appropriate test, and applied those criteria to recommend a test.

The Task Force began its work by reviewing the current general education requirements and learning outcomes at CUNY's 17 undergraduate colleges. Given the impossibility of measuring all outcomes with a single instrument, the Task Force identified five common core learning outcomes: reading, critical thinking, written communication, quantitative reasoning and information literacy. These five outcomes do not represent the full range deemed essential by all CUNY colleges. Consequently, an assessment instrument that measures these five abilities well can be just one component of a comprehensive assessment strategy.

The next step for the Task Force was to specify these outcomes so that they can be measured. For this purpose, the Task Force adopted five of the LEAP rubrics, developed by the AAC&U as part of its VALUE project. By so doing the Task Force did not endorse the rubrics, but merely adopted them as a convenient means of comparing the ability of the candidate assessment tests to discriminate the appropriate levels on the learning outcomes of interest.

The Task Force reviewed the following commercially-available tests:

1) The Critical Thinking Assessment Test;
2) The Collegiate Assessment of Academic Proficiency;
3) The Collegiate Learning Assessment;
4) The ETS Proficiency Profile (formerly known as the Measure of Academic Proficiency and Progress, MAPP).

The candidate tests were evaluated on the basis of 1) alignment of test purpose and design with the Task Force charge, 2) psychometric quality with respect to reliability and validity of test results, and 3) quality of the test development and administration process.

The Task Force divided into three panels, each of which reviewed one assessment and presented the results to the full Task Force. Subsequently, each member of the Task Force rated each of the candidate tests on the entire set of evaluation criteria. If two-thirds or more of Task Force members (i.e., 8 or more) assigned the same rating on any criterion, consensus was achieved.

The CLA was the only test to receive a consensus “outstanding” rating in any of the evaluation items – for content validity. The design of the CLA tasks requires students to demonstrate the higher-order critical thinking and analysis skills called for in the VALUE rubrics. The CLA also employs scoring rubrics that are similar in range and scaling to those of the VALUE rubrics. In all test specification areas related to purpose and design, the CLA received strong consensus agreement on acceptability.

The Task Force also reached consensus ratings of “acceptable” on all matters related to test development and logistics for the CLA, noting the need to conduct research on the validity of the electronic scoring methodology to be fully implemented soon by the Council for Aid to Education (CAE), the organization that develops and scores the CLA.

In a unanimous vote, with one abstention, the Task Force recommended adoption of the CLA by CUNY.

In addition to recommending an assessment instrument, the Task Force began to discuss how to administer the CLA so as to produce a valid measurement of learning gains and permit benchmarking against gains at non-CUNY colleges.

The CLA will be administered to samples of students who are just beginning their undergraduate studies and to students who are nearing the end of their undergraduate career. The sampling must be done randomly to produce representative results; yet random sampling will pose logistical challenges. CUNY may be able to learn from other institutions how best to motivate randomly selected students to demonstrate their true ability on the assessment.

The Task Force emphasizes that the CLA assesses a limited domain and should not be regarded as a comprehensive measure of general education outcomes defined by CUNY colleges. The test is not intended to evaluate all aspects of institutional effectiveness and is not designed to assess individual student or faculty performance.

Finally, the Task Force calls attention to the fact that the national sample of colleges that have administered the CLA differs in important respects from the CUNY student body, and that only a handful of community colleges have administered the community college version of the CLA to date. This lack of comparability may initially hamper CUNY's ability to interpret its learning gains with reference to national averages. All of the other candidate tests are characterized by this important constraint.

Report of the CUNY Assessment Task Force

Introduction

Driven by the dual mandates of external accountability and a desire for improvement, colleges across the nation have been strengthening their ability to assess learning for the past several decades (see Ewell, 2009). Accreditation bodies now uniformly require credible evidence of assessment. So too have legislatures, parents, students and other stakeholders demanded proof that higher education delivers on its promises. From an institutional perspective, a significant goal of outcomes assessment is to measure student learning gains, that is, to determine the “value added” by the college experience and to use that information to improve the quality of instruction. The question of how much American undergraduates are learning reached new urgency in 2011, with the publication of *Academically Adrift: Limited Learning on College Campuses*, which reported that a sizeable percentage of students manifest no measurable gain in critical thinking skills during their first two years of college (Arum and Roksa, 2011).

To be successful, assessment initiatives must be built around the regular, ongoing work of teaching and learning, firmly rooted in the college, its departments and the classroom (Hutchins, 2010). Faculty define the learning goals of general education, degree programs and courses, develop an array of appropriate metrics for measuring progress toward those goals, and draw upon the results to improve instruction. Ideally, assessment employs a variety of methods, both qualitative and quantitative, both formative and summative. CUNY’s history of administering standard instruments system-wide enables its colleges to combine the information from system-level instruments with data derived from local assessments.

For ten years, the CUNY Proficiency Exam (CPE) served as a means of assessing individual student proficiency in writing and quantitative reasoning. Approved by the CUNY Board of Trustees in 1997 and implemented in 2001, the CPE was designed to certify that students who had reached the 45th credit were ready for upper division course work. Because every CUNY student was required to pass the test in order to graduate, it was a high-stakes examination. In November 2009, Executive Vice Chancellor Alexandra Logue convened the CPE Task Force to evaluate the strengths and limitations of the CPE.

After extensive deliberations, the CPE Task Force recommended that CUNY discontinue the use of the CPE (CUNY Proficiency Examination Task Force, 2010). As a certification exam, the CPE had become redundant. Nearly every student who was eligible to take the exam— by completing 45 credits with a 2.0 GPA or better— passed the exam. Further, given that the CPE was designed by CUNY and administered only within CUNY, it could not be used to benchmark achievements of CUNY students against those of students at comparable institutions. Because it was administered only at a single point in time, the CPE also did not measure learning gains over time. Finally, the development and administration of the test had become prohibitively expensive, projected at $5 million per year going forward. The Board of Trustees took action to discontinue the CPE in November 2010.

Following Board action on the CPE, Executive Vice Chancellor Logue established a faculty-based task force to identify a test to assess student learning that would shift the focus from high-stakes assessment of individual students to institutional assessment of learning gains. In January 2011, the CUNY Task Force on System-wide Assessment of Undergraduate Learning Gains (Assessment Task Force) was charged as follows:

> *The Chancellery wishes to identify and adopt a standardized assessment instrument to measure learning gains at all of CUNY's undergraduate institutions. The instrument should be designed to assess the ability to read and think critically, communicate effectively in writing, and measure other learning outcomes associated with general education at CUNY. It must be possible for each college and the University to benchmark learning gains against those of comparable institutions outside CUNY. It is the responsibility of the Task Force to identify the most appropriate instrument and to advise the Chancellery on how best to administer the assessment and make use of the results.*
>
> *The Task Force is charged with the following specific responsibilities:*
> 1. *Taking into account psychometric quality, the alignment of the domain of the instrument with broad learning objectives at CUNY colleges, cost, facility of obtaining and using results, and the ability to benchmark results externally, select an assessment instrument from among those commercially available at this time.*
> 2. *Develop recommendations for the chancellery on how the assessment should best be administered so as to*
>   a. *represent each college's undergraduate student body;*
>   b. *generate a valid assessment of learning;*
>   c. *facilitate comparisons across CUNY colleges and between CUNY and other postsecondary institutions.*
> 3. *Develop recommendations on how the colleges and the chancellery can best use the results to improve teaching and learning throughout CUNY.*

Candidates for the Assessment Task Force were nominated by the campuses on the basis of their assessment and psychometric expertise as well as their familiarity with undergraduate education at CUNY. Panel members were named by the Chancellery and included representatives from community and senior colleges, and the Central Office. One member, Kathleen Barker, was named by the University Faculty Senate. Three additional members are UFS senators: Lisa Ellis, Dahlia Remler and Ellen Belton. A complete list of Task Force members follows:

Mosen Auryan, Director of Assessment, Hunter College
Kathleen Barker, Professor, Department of Psychology, Medgar Evers College
Ellen Belton, Professor, Department of English, Brooklyn College

David Crook, University Dean for Institutional Research and Assessment, CUNY
Margot Edlin, Faculty Fellow in Academic Affairs, Basic Educational Skills Department, Queensborough Community College
Lisa Ellis, Professor, Department of Library, Baruch College
Richard Fox, Dean for Institutional Effectiveness and Strategic Planning, Kingsborough Community College
Howard Everson, Professor and Research Fellow, Center for Advanced Study in Education, CUNY Graduate Center
Raymond Moy, Director of Assessment, CUNY
Dahlia Remler, Professor, School of Public Affairs, Baruch College and Department of Economics, Graduate Center
Karrin Wilks, University Dean for Undergraduate Studies, CUNY

This report summarizes the work of the Assessment Task Force from its inception in January 2011 through deliberations to the point of recommending an instrument in May 2011. The Task Force discussed the methodological issues associated with assessing learning gains, and this report contains some initial recommendations for administering the test. However, these questions merited additional deliberation, and more detailed recommendations will be presented in a supplementary report.

Test Requirements

The formal charge to the Assessment Task Force set forth a series of requirements related to test content, including: 1) the domain of the test will align with broad learning objectives at CUNY colleges; 2) the test must be capable of measuring learning gains over time; 3) the test must allow CUNY to benchmark college performance against that of comparable institutions outside of CUNY; and 4) test scores must provide information specific enough to inform the design of policy and practice to improve teaching and learning at individual CUNY colleges.

To define the optimal domain of the test, the Task Force began with a review of the current general education requirements and learning outcomes at CUNY's 17 undergraduate colleges. Although general education learning outcomes are structured in various ways across the campuses, requirements for the most part can be classified in six categories: communication skills (reading, writing, speaking), quantitative and scientific reasoning, critical thinking, research and information literacy, knowledge of arts and humanities, and civic and personal responsibilities. Not all CUNY colleges have articulated outcomes in all six categories, and there is significant overlap of desired outcomes across the categories.

Given the impossibility of capturing all outcomes with a single instrument, the Task Force identified the core learning outcomes common across CUNY: reading, critical thinking, written communication, quantitative reasoning and information literacy. The Task Force acknowledges that these competencies do not represent the full range of learning outcomes deemed essential by CUNY colleges and institutions across the country (see Liberal Education and America's

Promise, 2007). Nor do they adequately represent discipline-specific knowledge and competencies. The assessment instrument best aligned with this restricted domain must therefore be seen as one component of a more comprehensive assessment system comprised of the many formative and summative measures tailored to assess general education learning outcomes.

After identifying the core skills that the new assessment should measure, the Task Force sought to define each skill area comprehensively and from a developmental perspective in order to evaluate the capacity of candidate tests to measure the outcomes. In 2007, as part of its Liberal Education and America's Promise Initiative, the Association of American Colleges and Universities launched the VALUE project to explore the articulation and assessment of broad standards for undergraduate learning (VALUE: Valid Assessment of Learning in Undergraduate Education Project, 2007). The VALUE project brought together hundreds of faculty and assessment experts from every type of postsecondary institution to develop rubrics to assess learning at beginning, intermediate and advanced levels of accomplishment across fifteen domains (Rhodes, 2010). The rubrics were extensively field-tested (LaGuardia Community College was a participant), and currently are used by institutions across the country including several CUNY colleges. The Task Force adopted the VALUE rubrics in reading, critical thinking, written communication, quantitative literacy and information literacy as a means of defining learning outcomes for progressively more sophisticated performance in each area. The intent of the Task Force was to evaluate the power of the candidate tests to discriminate the skills and skill levels represented in these rubrics.

Candidate Tests

Given the requirement for benchmarking CUNY institutional performance against that of comparable institutions outside CUNY, the number of candidate tests was limited to standardized tests that are nationally administered. According to the National Institute for Learning Outcomes Assessment (2011), only five such tests are currently available:

1) The Critical Thinking Assessment Test;
2) The Collegiate Assessment of Academic Proficiency;
3) The Collegiate Learning Assessment;
4) The ETS Proficiency Profile (formerly known as the Measure of Academic Proficiency and Progress, MAPP); and
5) WorkKeys.

An overview of each test is provided below (see Appendix A for sample test items). Of the five, only the CAAP, CLA and ETS Proficiency Profile are used to measure student learning gains in the Voluntary System of Accountability (VSA). Sponsored by the American Association of State Colleges and Universities and the Association of Public and Land-grant Universities, the VSA was developed in 2007 for public four-year institutions to provide comparable information on the undergraduate experience through a standard "college portrait." Currently, over 520 institutions participate in the VSA (Voluntary

System of Accountability, 2007). By adopting one of the three sponsored tests, CUNY would gain the option of participating in the VSA.

Critical Thinking Assessment Test (CAT)

With support from the National Science Foundation, the CAT was developed in 2001 to assess and promote the improvement of critical thinking and real-world problem solving skills. Six universities were involved in its development: Howard University, Tennessee Technological University, University of Colorado, University of Hawaii, University of Southern Maine, University of Texas, and the University of Washington.

The CAT is designed to assess critical thinking skills by having students evaluate information, demonstrate creative thinking, solve problems, and write critically. Students are allowed one hour to complete the two-part test. Part I is a series of questions about real-world topics to which students respond in short essay format. Part II is another series of questions that must be answered using a packet of eight short readings (four relevant, four irrelevant).

Students are awarded up to 38 points, with questions varying in value from 1-5 points each. All tests are scored by the administering college's faculty, who are trained to use a detailed scoring guide.

Collegiate Assessment of Academic Proficiency (CAAP)

The CAAP was developed by ACT and has been in use since 1990 by two and four-year colleges to measure academic progress in six areas: writing skills (usage/mechanics and rhetorical skills), writing (essay), mathematics, reading, critical thinking, and science. Except for the Writing Essay, all items are multiple-choice.

Writing Skills is a 72-item, 40-minute assessment of punctuation, grammar, sentence structure, appropriateness to audience and purpose, organization of ideas, and style. The test is based on six prose passages.

The Writing Essay is comprised of two 20-minute writing tasks. Student essays are scored independently by two trained raters on a holistic scale of 1-6.

Mathematics is a 35-item, 40-minute test with questions drawn from pre-algebra, elementary algebra, intermediate algebra, coordinate geometry, college algebra, and trigonometry. Approximately half of the items are at the basic algebra level, with the other half at the college algebra level.

Reading is a 36 item, 40-minute test based on four prose passages, each about 900 words in length. Approximately 30% of the items refer directly to the text while the other 70% require making inferences beyond the text.

The Science Test is a 45-item, 40-minute test consisting of questions drawn from biological sciences, chemistry, physics, and the physical sciences. There are eight passages of varying perspective, including data representation (33%), research summaries (54%) and conflicting viewpoints (13%). The test items themselves are classified by area of scientific inquiry: understanding (23%), analyzing (51%) or generalizing (27%).

The Critical Thinking Test is a 32-item, 40-minute test that measures students' skills in analyzing (59%), evaluating (22%), and extending (19%) arguments. The items are linked to one of four passages that present a series of sub-arguments in support of more general conclusions.

Collegiate Learning Assessment (CLA)

The CLA was developed by the Council for Aid to Education (CAE) as an alternative to multiple-choice tests of critical thinking and written communication skills. The CLA is designed to evaluate student skills through cognitively challenging and realistic tasks. It consists of the Performance Task and two types of Analytic Writing Tasks. Student responses are evaluated according to analytic rubrics that can be scored by outside readers or computer.

The Performance Task is a 90-minute test that requires students to answer several open-ended questions about a hypothetical but realistic situation. The Performance Task includes a document library consisting of a range of sometimes conflicting information sources, such as letters, memos, and summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes. Students are expected to base their responses on an analysis and synthesis of information presented.

There are two types of Analytic Writing tasks – Make-an-Argument, which asks students to support or reject a position on an issue; and Critique-an-Argument, which requires students to evaluate the validity of an argument presented in a prompt. The tests are 45 and 30 minutes long respectively.

CAE has recently begun administering the CLA at community colleges, but here the test is referred to as the CCLA (Community College Learning Assessment), mainly because performance comparisons are limited to two-year institutions. Otherwise, the design and content of the CCLA and the CLA are the same.

ETS Proficiency Profile (ETSPP)

The ETSPP was developed in 1990 for use by two and four-year colleges and universities. It is designed to assess learning outcomes of general education programs in order to improve the quality of instruction and learning.

The ETSPP consists of 108 items, with 27 items for each subtest area measuring Critical Thinking, Reading, Writing, and Mathematics. The Critical Thinking and Reading subtest items are linked to brief reading selections, pictures, or graphs representing three academic contexts— humanities, social sciences, and natural sciences. The Writing multiple choice items are based on sentence-level texts with answer alternatives that focus on the test-taker's knowledge of grammar, syntax, and usage. The Mathematics section contains word problems, computations, and algebraic equation solving at varying levels of difficulty. The test can be administered in a single two-hour session or in separate testing sessions of one hour each. Colleges have the option to add up to 50 of their own multiple-choice items and/or an essay

WorkKeys

WorkKeys is a job skills assessment system developed by ACT. It tests nine foundational skills needed for success in the workplace, including applied mathematics, locating information, reading for information, applied technology, business writing, listening, teamwork, workplace observation, and writing. There is a subtest for each skill area with a series of six to thirty-eight work-based questions that are of increasing levels of difficulty. Most of the questions are multiple-choice. Each subtest is timed and lasts between 30 and 64 minutes.

Test Selection Specifications

The candidate tests were evaluated on the basis of: 1) alignment of test purpose and design with the Task Force charge, 2) psychometric quality with respect to reliability and validity of test results, and 3) quality of the test development and administration process. These evaluation criteria are described more fully below and formed the basis for detailed test evaluation guidelines developed by OIRA (see Appendix B).

Test Purpose and Design

The Task Force first evaluated candidate tests on the extent to which the publisher's stated test purposes and design align with CUNY's purposes for the test. The Task Force charge identified three purposes: 1) measure learning gains, 2) benchmark college performance against that of comparable institutions, and 3) use the results to improve teaching and learning throughout CUNY.

Psychometric Quality

The psychometric quality of a test depends on the validity and reliability of its scores. Validity refers to the degree to which evidence and theory support the interpretations of test scores within the proposed uses of the test; reliability refers to the consistency of scores when the testing procedure is repeated on different populations (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Validity. There are three types of validity that the Task Force reviewed for each test: 1) content validity, 2) external criterion validity, and 3) validity generalization.

Content validity was evaluated in relation to the VALUE rubrics in reading, critical thinking, written communication, quantitative literacy and information literacy as described earlier. The Task Force assessed how well candidate tests covered the skills and competencies in the rubrics, as well as the tests' ability to differentiate among the performance levels described in the rubrics (see Appendix C for the VALUE rubrics).

External criterion validity depends on how well a test's results correlate with other known measures of the construct of interest. The Task Force evaluated the extent to which candidate test scores detected learning gains as measured by external criteria, including scores on other tests.

Validity generalization is the extent to which the observed validity relationships are generalizable to different test takers, test sessions or time periods, or other conditions in which a test might be administered. The Task Force evaluated how candidate test scores were to be interpreted and used, and the demographic profiles of the colleges included in the norming and benchmarking of test results.

Reliability. To assess reliability, the Task Force reviewed the candidate tests' technical materials for evidence of how stable test scores are over different forms of a test, as well as the internal consistency of the test items that make up a total test score. When test scores are assigned by multiple human graders, inter-rater consistency was reviewed as well.

Test Development and Administration

The process for evaluating test development and administration took into account the four principal phases of test development detailed in the *Standards for Educational and Psychological Testing* (American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 1999). The four phases are: 1) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; 2) development and evaluation of the test specifications; 3) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and 4) assembly and evaluation of the test for operational use.

The Task Force reviewed the technical manuals of candidate tests and any other documentation available on the test development process, looking specifically for details about the measurement construct being evaluated, test design specifications, scoring methodology and test quality assessments gained from the field testing of multiple forms of the instrument.

The Task Force also reviewed the administrative manuals of candidate tests or other documentation that specified the standard procedures for orienting students to the test, the test proctor scripts, and instructions for handling the test forms from the point at which they are delivered by the test publisher, to scoring and generation of score reports, to final disposition of used exam booklets. Administering the test to properly oriented students under standard and secure conditions is essential for gathering quality data.

Finally, the Task Force reviewed the quality of the candidate tests' score reports to the colleges. Evaluation criteria included the organization and lay-out of the report, the interpretability of results, and the availability of comparative data on both internal and external groups. Additionally, the Task Force looked for the ability of the test publisher to provide customized reports.

Test Evaluation Methodology

Each of the candidate tests was reviewed using the test evaluation guidelines in Appendix B. Evaluations were based primarily on the materials provided by the test publishers, including sample tests. A list of these materials appears in Appendix D, along with webpage links where available. Although WorkKeys was designed to assess some of the same skill areas as the other candidate tests (e.g., mathematics, reading, and writing), the target audience is high school graduates intending to enter the workplace. Consequently, the Task Force did not regard WorkKeys as a viable test for assessing college-level learning or for benchmarking performance with institutions outside CUNY, and did not formally evaluate the test using the methodology below.

For the CAT, OIRA demonstrated the review and scoring methodology to the Task Force. To review the CAAP, CLA and ETSPP, the Task Force was divided into three teams. Each team consisted of at least one member with test assessment or development expertise; two or three additional team members were assigned by lottery. Each team reviewed the materials associated with its test and met to evaluate the materials against the criteria described above. Further information was requested from test publishers as needed— primarily samples of test forms and score reports. Each team presented its findings to the entire Task Force for discussion and further analysis. Summaries of team presentations appear in Appendix E.

Test Evaluations

After all four presentations were complete, each member of the Task Force individually rated the candidate tests on the entire set of evaluation criteria, assigning scores of 1 (unacceptable--serious lack or deficiency), 2 (acceptable), or 3 (outstanding or highly desirable feature) to each criterion (see Appendix F for the test evaluation score sheet). If two-thirds or more of Task Force members (i.e., 8 or more) assigned the same rating on any criterion, consensus was achieved. Thus, if 8 or more members assigned a 1 rating on any criterion, a consensus of "unacceptable" was achieved for that criterion. If 8 or more members assigned a 2 or 3 rating (acceptable or outstanding) on any criterion, consensus on "acceptable" was achieved. If a criterion ranking received fewer than 8 votes, consensus was not achieved. Based on the compilation of individual ratings by Task Force members, each test received one of three possible designations on each criterion: Not Acceptable (consensus on a rating of 1), Acceptable (consensus on a rating of 2 or 3) or No Consensus.

The results of the Task Force evaluations of the candidate tests appear in Appendix F. For each test under consideration, the number of Task Force members assigning a score of “1”, “2” or “3” is presented for each evaluation criterion. The table below provides a summary of consensus patterns.

Table 1: Summary of Consensus Patterns of Task Force Members’ Rankings

Test	Total “acceptable” consensus ranking	Total “unacceptable” consensus rankings	Total “no consensus” ranking
CAT	1	0	12
CAAP	7	1	5
CLA	12	0	1
ETSPP	8	2	3

As indicated above, the consensus agreement on the acceptability of the CLA was greater than for any other test, and it did not receive any consensus ratings of unacceptable. Further, and as detailed in Appendix F, the CLA received by far the fewest number of individual “unacceptable” ratings from the members of the Task Force. The total number of individual member ratings of 1 (serious lack or deficiency) across criteria follows: CAT (58), CAAP (32), CLA (6), and ETSPP (43).

CAT

The Task Force found the content validity of the CAT to be its strongest feature as it requires test-takers to engage their critical thinking and problem solving skills as reflected in the VALUE rubrics. However, the strength of the test design could not compensate for the serious deficiencies the Task Force found with its test development and benchmarking characteristics. To date, only one form of the test has been developed. Only one sample test prompt has been released to the public, and no test specifications have been published. The Task Force determined that it lacked sufficient information to conduct a thorough evaluation of the CAT. Finally, available benchmark data were limited to a sample of 7 colleges whose demographics were not comparable to CUNY’s.

CAAP

Although the members of the Task Force found the CAAP to be acceptable on many of the criteria, they determined that the test did not adequately reflect the VALUE rubrics. Most of the items on the CAAP—including those purporting to measure skills in reading, critical thinking, science and even mathematics— can more accurately be described as measuring reading comprehension. All items are multiple-choice questions to be answered based on a reading prompt. Even as reading items, the level of performance required by the CAAP does not go beyond the lowest levels of reading skills described in the VALUE rubric, for example, focusing on identifying the author's point rather than reacting to ideas in the text.

The Writing Essay section of the CAAP requires students to respond to a prompt that identifies a hypothetical situation and audience, the same format that was formerly used by CUNY to assess basic skills proficiency and readiness for freshman composition. The Task Force viewed this as highly problematic given the recent work to significantly revise the CUNY basic skills test in writing to better reflect expectations of faculty. Overall, the Task Force found the CAAP to be unacceptable for measuring core learning outcomes at CUNY, and inadequate for measuring the full range of skills described in the VALUE rubrics.

CLA

The CLA was the only test to receive a consensus "outstanding" rating in any of the evaluation items – for content validity. The design of the CLA tasks requires students to demonstrate the higher-order critical thinking and analysis skills reflected in the VALUE rubrics. The CLA also employs scoring rubrics that are similar in range and scaling to those of the VALUE rubrics. In all test specification areas related to purpose and design, the CLA received strong consensus agreement on acceptability.

In terms of psychometric quality, the CLA had acceptable ratings for all evaluation items except for comparability of the colleges available in the benchmarking sample to CUNY. This lack of comparability, especially in terms of minority and English language learning status, was found with all candidate tests, whose norming populations were predominantly white and native speakers of English. Another concern has to do with the paucity of community colleges in the norming population for the CLA. The community-college version of the CLA, the CCLA, (which consists of the same prompts as the CLA) has been given at only 6 community colleges as of this writing. The CAE will soon report results in terms of performance levels on the scoring rubrics rather than against norming populations, a fact not reflected in the rankings by the Task Force. The CAE will report the percent of test takers at each score point on the rubric, and in particular the percent reaching a "proficient" level of performance. Neither the CAAP nor the ETSPP can report scores in this way since their results are all norm based.

The Task Force also reached consensus ratings of "acceptable" on all matters related to test development and logistics for the CLA. However, because the CAE has recently implemented machine scoring for all of its unstructured response tests, the Task Force recommends that the University obtain more information about the validity of the scoring process and consider the possible implications for the interpretation of test scores.

ETSPP

The strength of the ETSPP is principally in its technical execution. It is closest to the CAAP in design - focusing on multiple-choice answers in response to texts. However, the texts are much shorter and the number of items far fewer. The items in the ETSPP are developed for their efficiency in discriminating total test score differences in the norming population, and do not reflect the learning outcomes in the VALUE rubrics. The Task Forces gave the ETSPP low ratings for content validity, and for the capacity of the test to measure CUNY's core learning outcomes.

Cost

The costs of administering the tests were obtained from the respective test websites. Task Force members reviewed cost information, but cost did not emerge as a primary consideration in the evaluation of candidate tests. Table 2 provides a cost comparison for testing 200 freshmen and 200 seniors per college. The cost of incentives is not included in these estimates.

Table 2
Estimated Annual Cost of Administering CAT, CAAP, CLA and ETSPP to a Sample of 200 Freshmen and 200 Seniors per College

	CAT	CLA	CAAP	ETSPP
Number tested[a]	7,200	7,200	7,200	7,200
Set up per college	$550	$6,500[b]		
Cost per unit	$5.00	$25.00	$19.20	$14.80
Scoring of essay	$11.00[c]	NA	$13.50	$5.00
Total cost of scoring	$125,100	$207,000	$235,440	$142,560

[a] 18 colleges with 400 students each

[b] Includes testing and scoring of 200 students per college. Additional students are $25 each.

[c] CAT trains college faculty to score. The scoring cost is an estimate based on the average per paper scoring cost for the CATW.

Task Force Recommendations

The primary objective of the charge to the Task Force was to identify a standardized assessment instrument to measure learning gains at all of CUNY's undergraduate institutions. The selection of a test is but a first step in the implementation of an assessment system designed to gather reliable and valid data, and to interpret and use the results to inform the teaching and learning process. This report contains the Task Force's recommendations for a test instrument. A supplementary report will provide guidance on test administration and use of test results by faculty and academic administrators.

Test Selection

After a review and discussion of the tallies of rankings for all criteria, as well as the patterns of consensus across candidate tests, the Task Force voted unanimously— with one abstention— to recommend adoption of the CLA.

Of the tests commercially available, the CLA is the only instrument that adequately meets design and quality requirements identified by the Task Force. Most significantly, the CLA addresses the following core learning outcomes for general education programs across CUNY: reading, critical thinking, written communication, quantitative literacy, and information literacy. Further, the CLA is the only test that can adequately measure the range of abilities described by the VALUE rubrics.

The Task Force does not, however, endorse the CLA for all purposes. CLA results are intended for use in evaluating learning outcomes only at the institutional level and primarily as a "signaling tool to highlight differences in programs that can lead to improvements in teaching and learning" (from the introduction to the sample 2009-2010 CLA Institutional Report). As indicated earlier, the CLA assesses learning in a limited domain and cannot be regarded as a comprehensive measure of general education outcomes as currently defined by CUNY colleges or as may be defined by the Pathways initiative. The test is not intended to evaluate all aspects of institutional effectiveness and is not designed to assess individual student or faculty performance. The Task Force also urges caution with respect to interpreting the available benchmarking data. In its standard report to participating colleges, the CAE provides data comparing the learning gains at each college to gains measured in the national sample. The validity of these comparisons may be affected by the extent to which the colleges comprising the benchmark sample resemble CUNY and the degree to which the sample of tested students in the benchmark colleges reflects the total population of undergraduates in those colleges.

Implementation and Logistics

The Task Force identified sampling design, motivation of students, and involvement of faculty as keys to the successful implementation of the CLA. Sampling must be conducted carefully so that the test results accurately reflect the level of learning and unique demographics at each CUNY institution. Because the test is not high stakes, CUNY must devise a strategy for encouraging test takers to demonstrate their true abilities on the test. Finally, unless faculty believe that the test is a valuable tool for assessing the learning goals they are attempting to advance in their own classrooms, the information generated by the assessment will not become a resource for improving learning outcomes of undergraduate students.

Sampling Design. To measure learning gains, CUNY must choose either a cross-sectional or a longitudinal design. In a cross-sectional study, random samples of freshmen and seniors are drawn during the school year— freshmen in the fall and seniors in the spring. In a longitudinal study, a group of freshmen is tested in their first year, and then again as seniors. In theory, the two designs should yield equivalent results. However, both designs present challenges associated with the treatment of drop-outs and transfer students, and solutions to these issues must be standardized if the measurement of gains is to be benchmarked across institutions. Because of the multi-year period required to execute a longitudinal design, the Task Force endorses a cross-sectional design. Moreover, because CUNY wishes to use the same instrument to test learning outcomes at all of its colleges—community and senior—the Task Force recommends testing students at the beginning of their academic career, at roughly the 60th credit, and for students pursuing the bachelors degree, when approaching the 120th credit. Finally, in developing a sampling scheme, analysts must take into account the numbers of ESL and remedial students, and the appropriateness of including them in the college's representative sample. Both groups may face special challenges in a timed testing situation.

The methodological issues of sampling will have a direct effect not only on assessments of learning at the institutional level, but also on calculations of learning gains and subsequent derivations of the learning gains to be ascribed to the college rather than to natural maturation. A further complication to measuring learning gains is determining the nature and significance of any gain. The assessment of learning gains must take into account changes in performance from one point in time to the next, as well as gain relative to specific standards. With both methodological and substantive complexities in play, the Task Force recommends caution in the initial administrations of the test and the use of multiple alternative measures to help in the interpretation of results.

Motivation of Students. Students must be motivated to demonstrate their true abilities on the assessment. The challenges associated with this goal have been the focus of research on the CLA as well as the subject of implementation surveys among colleges participating in CLA testing. It is recommended that the University review these studies and consult with colleges that have demonstrated success administering the CLA using a sampling scheme to recruit test takers.

Engaging Faculty and Academic Administrators. Some institutions have reported success in fostering a campus culture for assessment and improvement. When faculty and students are committed to assessment, the challenges of student motivation are reduced. At CUNY, we should integrate the CLA into existing assessment initiatives to garner support for the test. The University should create a communication campaign to convince faculty and students that the results of the CLA can be used to improve the quality of undergraduate education at CUNY.

Works Cited

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Arum, R. and J. Roksa (2011). *Academcially Adrift: Limited Learning on College Campuses*. University of Chicago Press.

CUNY Proficiency Examination Task Force. (2010). *Report of the CUNY Proficiency Examination Task Force.* New York: City University of New York.

Ewell, P. (2009). *Assessment, Accountability and Improvement: Revisiting the Tension.* University of Illinois and University of Indiana: National Institute for Learning Outcomes Assessment.

Hutchins, P. (2010). *Opening Doors to Faculty Involvement in Assessment*. University of Illinois at Urbana-Champaign: National Institute for Learning Outcomes Assessment.

Liberal Education and America's Promise. (2007). *Liberal Education and America's Promise (LEAP) - Essential Learning Outcomes*. Retrieved June 21, 2011, from AAC&U - Association of American Colleges and Universities: http://www.aacu.org/leap/vision.cfm

National Institute for Learning Outcomes Assessment. (2011). Tool Kit: Tests. Retrieved July 13, 2011, from http://www.learningoutcomesassessment.org/tests.htm

Rhodes, T. (. (2010). *Assessing Outcomes and Improving Achievement: Tips and Tools for Using Rubrics.* Washington, D.C.: Association of American Colleges and Universities.

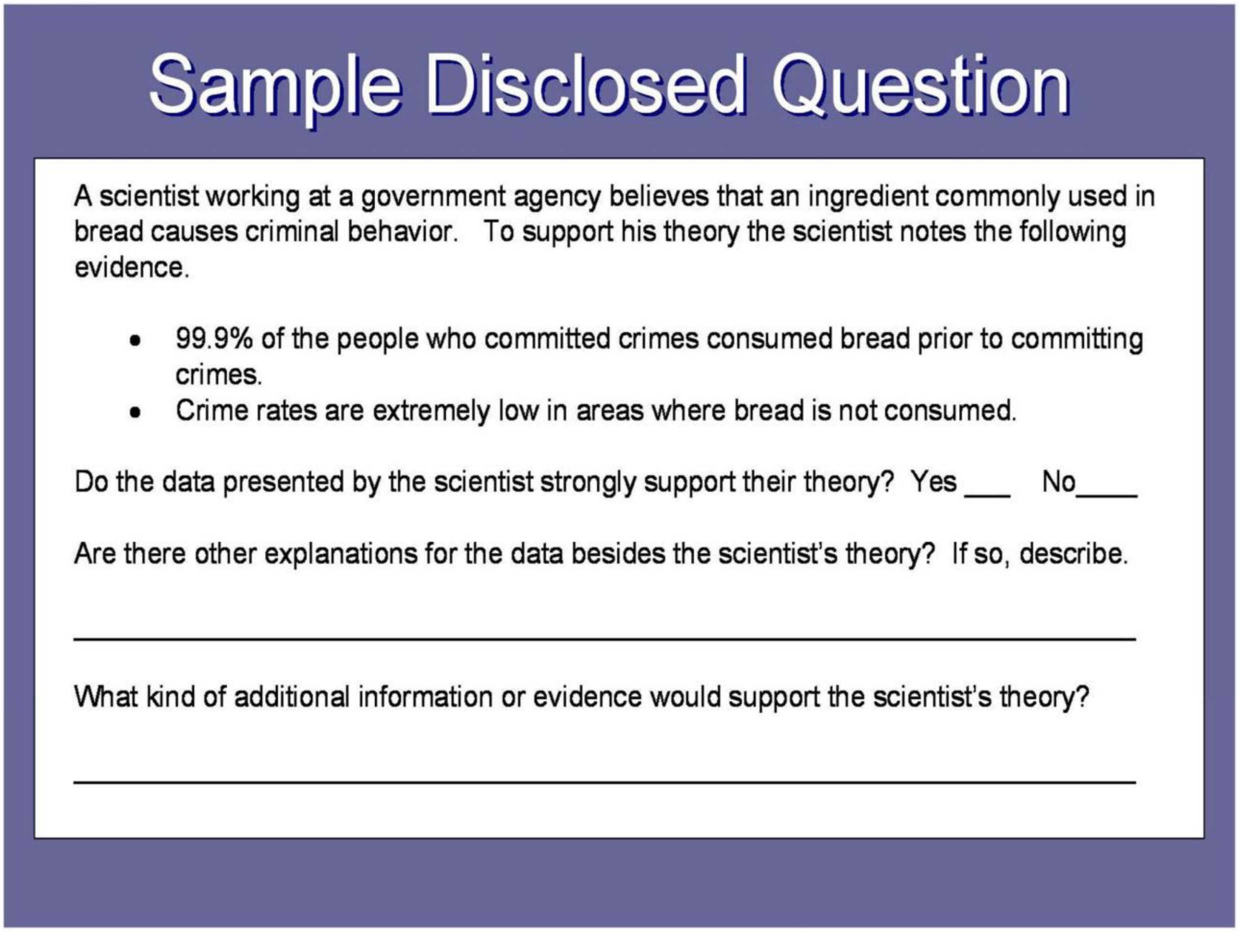
VALUE: Valid Assessment of Learning in Undergraduate Education Project. (2007). *VALUE: Valid Assessment of Learning in Undergraduate Education Overview*. Retrieved June 21, 2011, from AAC&U Association of American Colleges and Universities: http://www.aacu.org/value/index.cfm

Voluntary System of Accountability (2007). About VSA. Retrieved June 21, 2011, from Voluntary System of Accountability: http://www.voluntarysystem.org

Appendix A

Sample Test Items

CAT



Sample Disclosed Question

A scientist working at a government agency believes that an ingredient commonly used in bread causes criminal behavior. To support his theory the scientist notes the following evidence.

- 99.9% of the people who committed crimes consumed bread prior to committing crimes.
- Crime rates are extremely low in areas where bread is not consumed.

Do the data presented by the scientist strongly support their theory? Yes ___ No____

Are there other explanations for the data besides the scientist's theory? If so, describe.

____

What kind of additional information or evidence would support the scientist's theory?

____

APPENDIX A: SAMPLE ITEMS

CAAP

Sample CAAP Critical Thinking Item:

Directions: There are four passages in this test. Each passage is followed by several questions. After reading a passage, choose the best answer to each question by circling the corresponding answer option. You may refer to the passages as often as necessary.

Senator Favor proposed a bill in the state legislature that would allow pharmacists to prescribe medications for minor illnesses, without authorization from a physician (i.e., a "prescription"). In support of her proposal, Favor argued:

Doctors have had a monopoly on authorizing the use of prescription medicines for too long. This has caused consumers of this state to incur unnecessary expense for their minor ailments. Often, physicians will require patients with minor complaints to go through an expensive office visit before the physician will authorize the purchase of the most effective medicines available to the sick.

Consumers are tired of paying for these unnecessary visits. At a recent political rally in Johnson County, I spoke to a number of my constituents and a majority of them confirmed my belief that this burdensome, expensive, and unnecessary practice is widespread in our state. One man with whom I spoke said that his doctor required him to spend $80 on an office visit for an uncommon skin problem which he discovered could be cured with a $2 tube of prescription cortisone lotion.

Anyone who has had to wait in a crowded doctor's office recently will be all too familiar with the "routine": after an hour in the lobby and a half-hour in the examining room, a physician rushes in, takes a quick look at you, glances at your chart and writes out a prescription. To keep up with the dizzying pace of "health care," physicians rely more and more upon prescriptions, and less and less upon careful examination, inquiry, and bedside manner.

Physicians make too much money for the services they render. If "fast food" health care is all we are offered, we might as well get it at a good price. This bill, if passed into law, would greatly decrease unnecessary medical expenses and provide relief to the sick: people who need all the help they can get in these trying economic times. I urge you to vote for this bill.

After Senator Favor's speech, Senator Counter stood to present an opposing position, stating:

Senator Favor does a great injustice to the physicians of this state in generalizing from her own health care experiences. If physicians' offices are crowded, they are crowded for reasons that are different from those suggested by Senator Favor. With high operating costs, difficulties in collecting medical bills, and exponential increases in the costs of malpractice insurance, physicians are lucky to keep their heads above water. In order to do so, they must make their practices more efficient, relying upon nurses and laboratories to do some of the patient screening.

No one disputes the fact that medical expenses are soaring. But, there are issues at stake which are more important than money—we must consider the quality of health care. Pharmacists are not trained to diagnose illnesses. Incorrect diagnoses by pharmacists could lead to extended illness or even death for an innocent customer. If we permit such diagnoses, we will be personally responsible for those illnesses and deaths.

Furthermore, since pharmacies make most of their money by selling prescription drugs, it would be unwise to allow pharmacists to prescribe. A sick person who has not seen a physician might go into a drugstore for aspirin and come out with narcotics!

Finally, with the skyrocketing cost of insurance, it would not be profitable for pharmacists to open themselves up to malpractice suits for mis-prescribing drugs. It is difficult enough for physicians with established practices to make it; few pharmacists would be willing to take on this financial risk. I recommend that you vote against this bill.

Favor's "unofficial poll" of her constituents at the Johnson County political rally would be more persuasive as evidence for her contentions if the group of people to whom she spoke had:

I. been randomly selected.
II. represented a broad spectrum of the population: young and old, white and non-white, male and female, etc.
III. not included an unusually large number of pharmacists.

(A) I only
(B) II only
(C) III only
(D) I, II, and III

Sample CAAP Science Item (Biology, Data Representation Format):

Directions: There are eight passages in this test. Each passage is followed by several questions. After reading a passage, choose the best answer to each question by circling the corresponding answer option. You may refer to the passages as often as necessary.

A scientist investigated the factors that affect seed mass in the plant species Desnodium poniculatum. Some results of this study are summarized in the two tables below.

Table 1

Daylight hours	Other variable	Average seed mass (in mg) of plants raised at:	
		23°C	29°C
14	—	7.10	5.63
14	Leaves removed	7.15	6.11
14	Reduced water	4.81	5.81
8	—	6.12	—

Table 2

A. Number of seeds per fruit	Average seed mass (mg)
1	6.62
2	6.28
3	5.97
4	6.00
5	5.59
B. Position of seed in fruit*	Average seed mass (mg)

Their rooms were shrines of upholstery and lace. Silent radios standing under stacks of magazines. Did they work? Could I turn the knobs? Questions.I wouldn't ask here. Windows with shades pulled low, so the light peeping through took on a changed quality, as if it were brighter or dimmer than I remembered. And portraits, photographs, on walls, on tables, faces strangely familiar, as if I was destined to know them. I asked no questions and the women never questioned me. Never asked where the money went, had the price gone up since last year, were there any additional flavors. They bought what they remembered—if it was peanut-butter last year, peanut-butter this year would be fine. They brought the coins from jars, from pocketbooks without handles, counted them carefully before me, while I stared at their thin crops of knotted hair. A Sunday brooch pinned loosely to the shoulder of an everyday dress. What were these women thinking of?

And the door would close softly behind me, transaction complete, the closing click like a drawer sliding back, a world slid quietly out of sight, and I was free to return to my own universe, to Grandma standing with arms folded in the courtyard, staring peacefully up at a bluejay or sprouting leaf. Suddenly I'd see Grandma in her dress of tiny flowers, curly gray permanent, tightly laced shoes, as one of them—but then she'd turn, laugh, "Did she buy?" and again belong to me.

Gray women in rooms with the shades drawn . . . weeks later the cookies would come. I would stack the boxes, make my delivery rounds to the sleeping doors. This time I would be businesslike, I would rap firmly, "Hello Ma'am, here are the cookies you ordered." And the face would peer up, uncertain . . . cookies? . . . as if for a moment we were floating in the space between us. What I did (carefully balancing boxes in both my arms, wondering who would eat the cookies—I was the only child ever seen in that building) or what she did (reaching out with floating hands to touch what she had bought) had little to do with who we were, had been, or ever would be.

Naomi Shihab Nye, "The Cookies." © 1982 by Naomi Shihab Nye.

Which of the following statements represents a justifiable interpretation of the meaning of the story?
(A) The girl's experience selling Girl Scout cookies influenced her choice of careers.
(B) The girl's experiences with elderly women made her aware of the prospect of aging.
(C) Because she spent so much time with her grandmother, the girl preferred the company of older people to that of other children.
(D) The whole experience of selling Girl Scout cookies was a dream or hallucination and had nothing to do with who the girl really was.

Sample CAAP Writing Skills Item:

Directions: In the six passages that follow, certain words and phrases are underlined and numbered. In the right-hand column, you will find alternatives for each underlined part. You are to choose the one that best expresses the idea, makes the statement appropriate for standard written English, or is worded most consistently with the style and tone of the passage as a whole. If you think the original version is best, choose "NO CHANGE."

In the end, everyone gives up jogging. Some find that their strenuous efforts to earn a living **drains (1)** away their energy.
(A) NO CHANGE
(B) drain
(C) has drained
(D) is draining

Sample CAAP Writing Essay Prompt:

Your college administration is considering whether or not there should be a physical education requirement for undergraduates. The administration has asked students for their views on the issue and has announced that its final decision will be based on how such a requirement would affect the overall educational mission of the college. Write a letter to the administration arguing whether or not there should be a physical education requirement for undergraduates at your college.
(Do not concern yourself with letter formatting; simply begin your letter, "Dear Administration.")

Sample CAAP Mathematics Item (Pre-Algebra Application):

Directions: Solve each problem, then choose the correct answer by circling the corresponding answer option. Do not linger over problems that take too much time. Solve as many as you can; then return to the others in the time you have left for this test. You may use a calculator for any of the problems on this test. However, all problems can be solved without using a calculator, and some of the problems may in fact be simpler if done without a calculator.

Mark bought 3 shirts at a clothing store. If he paid a total of $15.00 for 2 shirts and the average (arithmetic mean) cost of the 3 shirts was $8.00, how much did Mark pay for the third shirt?
(A) $7.00
(B) $7.67
(C) $8.50
(D) $9.00
(E) $11.50

MAPP

Sample MAPP Reading and Critical Thinking Items:

Directions: Each stimulus (a passage, poem, graph, or table, for example) is followed by a question or questions based on that stimulus. Read each stimulus carefully. Then choose the best answer to each question following a stimulus.

Certain literary theorists claim to see no difference between literature and criticism. They rest their case on two similarities between the genres: both are impassioned and both use "literary language." The critical essays of John Ruskin (1819–1900) are surely impassioned, and surely full of literary language. However, we do recognize a difference, not in the use of language, but in the internal organization of parts between the literary genres (the novel, drama, poetry), which tend to be organized around a central, defining symbol or set of symbols, and the nonliterary ones (homily, criticism, the philosophical essay), which tend to be linear and discursive in nature. It is by some such structural principle, and not by any remarks about language, that we distinguish the critical essay from literary genres such as poetry.

Reading

The primary purpose of the passage is to
(A) analyze a major trend in recent literary theory
(B) point out the distinguishing features of certain important literary genres
(C) question the claim that there are significant differences between literary and nonliterary genres
(D) identify a means of differentiating between literary and nonliterary genres

Critical Thinking

Which of the following claims, if true, would be most difficult to reconcile with the argument made by the author of the passage?

(A) Few essayists are as skilled in their use of literary language as Ruskin was.
(B) Many prose poets tend to avoid the use of impassioned literary language in their work.
(C) The use of the symbol as a structuring device in poetry is more common in certain literary periods than in others.
(D) The essay form was invented in the late sixteenth century as a way for writers to articulate personal thoughts and feelings.

Sample MAPP Writing Item:

Directions: The following question tests your ability to rewrite a given sentence. You will be told exactly how to revise your new sentence. Keep in mind that your new sentence should have the same meaning as the sentence given to you. In choosing an answer, follow the requirements of standard written English; that is, pay attention to acceptable usage in grammar, diction (choice of words), sentence construction, and punctuation. Choose the best answer; this answer should be clear and exact, without awkwardness, ambiguity, or redundancy.

Being a female jockey, she was often interviewed.

Rewrite, beginning with

She was often interviewed

The next words will be

(A) on account of she was
(B) by her being
(C) because she was
(D) being as she was

Sample MAPP Math Item:

Directions: Solve each problem, using any available space on the page for scratchwork. Then decide which is the best of the choices given and select that answer.

A train traveled at a constant rate of f feet per second. How many feet did it travel in x minutes?

(A) $\frac{60f}{x}$

(B) $\frac{fx}{60}$

(C) $\frac{x}{60f}$

(D) $60fx$

CLA

Sample CLA Performance Task:

> Directions: Please read the instructions in Document 1 located in the Document Library (see right side of screen). Your answers to the questions that follow should describe all details necessary to support your position. Your answers will be judged not only on the accuracy of the information you provide, but also on how clearly the ideas are presented, how effectively the ideas are organized, and how thoroughly the information is covered. While your personal values and experiences are important, please answer all questions solely on the basis of the information above and in the Document Library.

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

1: Newspaper articles about the accident

2: Federal Accident Report on in-flight breakups in single engine planes

3: Pat's e-mail to you & Sally's e-mail to Pat

4: Charts on SwiftAir's performance characteristics

5: Amateur Pilot article comparing SwiftAir 235 to similar planes

6: Pictures and description of SwiftAir Models 180 and 235

Do the available data tend to support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups? What is the basis for your conclusion? What other factors might have contributed to the accident and should be taken into account? What is your preliminary recommendation about whether or not DynaTech should buy the plane and what is the basis for this recommendation?

Sample CLA Make-an-Argument Prompt:

Directions: You will have 45 minutes to plan and write an argument on the topic on the next screen. You should take a position to support or oppose the statement. Use examples taken from your reading, coursework, or personal experience to support your position. Your essay will be evaluated on how well you do the following:
1. State your position
2. Organize, develop, and express your ideas
3. Support your ideas with relevant reasons and/or examples
4. Control the elements of standard written English

Government funding would be better spent on preventing crime than in dealing with criminals after the fact.

Sample CLA Critique-an-Argument Prompt:

Directions: There is something wrong with the argument presented below. It is your job to explain what is wrong with the argument. Discuss any flaws in the argument, any questionable assumptions, any missing information, and any inconsistencies. What we are interested in is your critical thinking skills and how well you write your response. You will have 30 minutes to respond to the argument. You will be judged on how well you do the following:
1. Explain any flaws in the points the author makes
2. Organize, develop, and express your ideas
3. Support your ideas with relevant reasons and/or examples
4. Control the elements of standard written English

The number of marriages that end in divorce keeps growing. A large percentage of them are from June weddings. Because June weddings are so popular, couples end up being engaged for a long time just so that they can get married in the summer months. The number of divorces gets bigger with each passing year, and the latest news is that more than 1 out of 3 marriages will end in divorce. So, if you want a marriage that lasts forever, it is best to do everything you can to prevent getting divorced. Therefore, it is good advice for young couples to have short engagements and choose a month other than June for a wedding.

WorkKeys

Level 5 Reading for Information Sample Item

Goldberg's Auto Parts is served by more than fifty different accounts, each with its own sales representative, company name, corporate address, and shipping address. As a shipping and receiving clerk at Goldberg's, you are required to return defective merchandise to the manufacturer.

Standard procedure for returning an item begins with your written request to the company for authorization. Always send the request to the corporate address, not to the shipping address. Unless the company file folder contains a form for this procedure, write a business letter to the manufacturer supplying the item's stock number, cost, and invoice number; the date it was received; and the reason for its return. The manufacturer's reply will include an authorization number from the sales representative, a sticker for you to place on the outside of the box to identify it as an authorized return, and a closing date for the company's acceptance of the returned item. If you do not attach the provided sticker, your returned box will be refused by the manufacturer as unauthorized, and you will need to obtain a new letter, authorization, sticker, and closing date. Always send a returned box to the shipping address, not to the company's corporate address.

According to the policy shown, what should you do if you lose an authorization sticker?

1. Send a request for a return authorization along with the rejected part directly to the manufacturer's shipping address.
2. Send a request for return authorization along with the rejected part directly to the manufacturer's corporate address.
3. Repeat the standard procedure to obtain a new letter, authorization, sticker, and closing date.
4. Use a sticker from another company's folder.
5. Send the rejected part to your sales representative.

Appendix B
Test Evaluation Scoring Guide

Task Force on Learning Outcomes Assessment
Test Evaluation Scoring Guide

Specification	Basis of evaluation	Test evaluation scoring
Test purpose and design are consistent with the Task Force charge and with CUNY learning objectives	Tests are to be used to: • Measure learning gains • Benchmark college performance against that of comparable institutions outside CUNY • Improve teaching and learning throughout CUNY Core learning outcomes across CUNY: • Reading • Critical thinking • Written communication • Quantitative literacy • Information literacy	1. Test has a significant misalignment with task force purposes 2. Test is mostly aligned with task force purposes. 3. Test is aligned with task force purposes with some outstanding feature(s) that deserve attention.
Psychometric quality	**Content Validity** Do the test tasks require the test-taker to use the skills and competencies described in the relevant LEAP VALUE rubrics? Does the scoring of the tasks reflect the progression of rubric skill levels?	1. Test content has a significant misalignment with the VALUE rubrics 2. Test content is mostly aligned with the VALUE rubrics 3. Test content is closely aligned with the VALUE rubrics.
	External Criterion Validity What evidence is there that the test detects learning gains at the institutional level?	1. Little or no evidence of external validity with other indicators of the learning outcome(s) of interest. 2. Consistent evidence of external validity. 3. Strong evidence of external validity.

Specification	Basis of evaluation	Test evaluation scoring
	Validity generalization Does the test developer clearly set forth how test scores are to be interpreted and used? Are there other participating colleges in its database of results that are comparable to those of CUNY and can serve in a benchmarking function?	1. Test scores have weak or faulty interpretability beyond the tested sample. 2. Test scores are linked to an interpretable scale. 3. Test scores are linked to an interpretable scale that has actionable implications.
	Score accuracy for institution-level comparisons **Reliability** What evidence is there for stability of scores over different items or forms of the test? If tests are scored by humans, what is the inter-rater reliability of scores? Does the test developer provide guidance for sampling covariates, e.g., ESL status, gender, race?	1. Weak evidence of reliability over test items or raters. 2. Acceptable evidence of reliability over test items (or forms) and raters. 3. Precision of measurement allows detection of small changes in ability.
Test Development & Logistics	Is there a technical manual that describes the test development process, test specifications, scoring rubrics, field testing, and availability of multiple parallel forms? Is there a test administration manual that describes the testing protocol and any special testing requirements, e.g., online administration, administrator certification, test-taker preparation materials, scoring protocols How are test results communicated to the colleges? What guidance is there for score interpretation with respect to benchmarking and learning gains?	1. Little or no documentation. 2. Documentation is adequate. 3. Documentation is detailed and complete.

Appendix C

VALUE Rubrics for the Core Learning Outcomes at CUNY

The VALUE (Valid Assessment of Learning in Undergraduate Education) rubrics were developed under the auspices of the Association of American Colleges and Universities (AAC&U) by teams of faculty and other academic and student affairs professionals from across the United States. Each VALUE rubric contains the most common and broadly shared criteria or core characteristics considered critical for judging the quality of student work in that outcome area. From the 15 rubrics developed by AAC&U, the 5 rubrics appearing in Appendix C, are those that are common to all CUNY colleges.

Critical Thinking VALUE Rubric

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Framing Language

This rubric is designed to be transdisciplinary, reflecting the recognition that success in all disciplines requires habits of inquiry and analysis that share common attributes. Further, research suggests that successful critical thinkers from all disciplines increasingly need to be able to apply those habits in various and changing situations encountered in all walks of life.

This rubric is designed for use with many different types of assignments and the suggestions here are not an exhaustive list of possibilities. Critical thinking can be demonstrated in assignments that require students to complete analyses of text, data, or issues. Assignments that cut across presentation mode might be especially useful in some fields. If insight into the process components of critical thinking (e.g., how information sources were evaluated regardless of whether they were included in the product) is important, assignments focused on student reflection might be especially illuminating.

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Ambiguity: Information that may be interpreted in more than one way.
- Assumptions: Ideas, conditions, or beliefs (often implicit or unstated) that are "taken for granted or accepted as true without proof." (quoted from www.dictionary.reference.com/browse/assumptions)
- Context: The historical, ethical. political, cultural, environmental, or circumstantial settings or conditions that influence and complicate the consideration of any issues, ideas, artifacts, and events.
- Literal meaning: Interpretation of information exactly as stated. For example, "she was green with envy" would be interpreted to mean that her skin was green.
- Metaphor: Information that is (intended to be) interpreted in a non-literal way. For example, "she was green with envy" is intended to convey an intensity of emotion, not a skin color.

Critical Thinking VALUE Rubric

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	Milestones		Benchmark
	4	3	2	1
Explanation of issues	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown.	Issue/problem to be considered critically is stated without clarification or description.
Evidence *Selecting and using information to investigate a point of view or conclusion*	Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis or synthesis. Viewpoints of experts are questioned thoroughly.	Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis or synthesis. Viewpoints of experts are subject to questioning.	Information is taken from source(s) with some interpretation/evaluation, but not enough to develop a coherent analysis or synthesis. Viewpoints of experts are taken as mostly fact, with little questioning.	Information is taken from source(s) without any interpretation/evaluation. Viewpoints of experts are taken as fact, without question.
Influence of context and assumptions	Thoroughly (systematically and methodically) analyzes own and others' assumptions and carefully evaluates the relevance of contexts when presenting a position.	Identifies own and others' assumptions and several relevant contexts when presenting a position.	Questions some assumptions. Identifies several relevant contexts when presenting a position. May be more aware of others' assumptions than one's own (or vice versa).	Shows an emerging awareness of present assumptions (sometimes labels assertions as assumptions). Begins to identify some contexts when presenting a position.
Student's position (perspective, thesis/hypothesis)	Specific position (perspective, thesis/hypothesis) is imaginative, taking into account the complexities of an issue. Limits of position (perspective, thesis/hypothesis) are acknowledged. Others' points of view are synthesized within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) takes into account the complexities of an issue. Others' points of view are acknowledged within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) acknowledges different sides of an issue.	Specific position (perspective, thesis/hypothesis) is stated, but is simplistic and obvious.
Conclusions and related outcomes (implications and consequences)	Conclusions and related outcomes (consequences and implications) are logical and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order.	Conclusion is logically tied to a range of information, including opposing viewpoints; related outcomes (consequences and implications) are identified clearly.	Conclusion is logically tied to information (because information is chosen to fit the desired conclusion); some related outcomes (consequences and implications) are identified clearly.	Conclusion is inconsistently tied to some of the information discussed; related outcomes (consequences and implications) are oversimplified.

Written Communication VALUE Rubric

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Framing Language

This writing rubric is designed for use in a wide variety of educational institutions. The most clear finding to emerge from decades of research on writing assessment is that the best writing assessments are locally determined and sensitive to local context and mission. Users of this rubric should, in the end, consider making adaptations and additions that clearly link the language of the rubric to individual campus contexts.

This rubric focuses assessment on how specific written work samples or collectios of work respond to specific contexts. The central question guiding the rubric is "How well does writing respond to the needs of audience(s) for the work?" In focusing on this question the rubric does not attend to other aspects of writing that are equally important: issues of writing process, writing strategies, writers' fluency with different modes of textual production or publication, or writer's growing engagement with writing and disciplinarity through the process of writing.

Evaluators using this rubric must have information about the assignments or purposes for writing guiding writers' work. Also recommended is including reflective work samples of collections of work that address such questions as: What decisions did the writer make about audience, purpose, and genre as s/he compiled the work in the portfolio? How are those choices evident in the writing -- in the content, organization and structure, reasoning, evidence, mechanical and surface conventions, and citational systems used in the writing? This will enable evaluators to have a clear sense of how writers understand the assignments and take it into consideration as they evaluate

The first section of this rubric addresses the context and purpose for writing. A work sample or collections of work can convey the context and purpose for the writing tasks it showcases by including the writing assignments associated with work samples. But writers may also convey the context and purpose for their writing within the texts. It is important for faculty and institutions to include directions for students about how they should represent their writing contexts and purposes.

Faculty interested in the research on writing assessment that has guided our work here can consult the National Council of Teachers of English/Council of Writing Program Administrators' White Paper on Writing Assessment (2008; www.wpacouncil.org/whitepaper) and the Conference on College Composition and Communication's Writing Assessment: A Position Statement (2008; www.ncte.org/cccc/resources/positions/123784.htm)

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Content Development: The ways in which the text explores and represents its topic in relation to its audience and purpose.
- Context of and purpose for writing: The context of writing is the situation surrounding a text: who is reading it? who is writing it? Under what circumstances will the text be shared or circulated? What social or political factors might affect how the text is composed or interpreted? The purpose for writing is the writer's intended effect on an audience. Writers might want to persuade or inform; they might want to report or summarize information; they might want to work through complexity or confusion; they might want to argue with other writers, or connect with other writers; they might want to convey urgency or amuse; they might write for themselves or for an assignment or to remember.
- Disciplinary conventions: Formal and informal rules that constitute what is seen generally as appropriate within different academic fields, e.g. introductory strategies, use of passive voice or first person point of view, expectations for thesis or hypothesis, expectations for kinds of evidence and support that are appropriate to the task at hand, use of primary and secondary sources to provide evidence and support arguments and to document critical perspectives on the topic. Writers will incorporate sources according to disciplinary and genre conventions, according to the writer's purpose for the text. Through increasingly sophisticated use of sources, writers develop an ability to differentiate between their own ideas and the ideas of others, credit and build upon work already accomplished in the field or issue they are addressing, and provide meaningful examples to readers.
- Evidence: Source material that is used to extend, in purposeful ways, writers' ideas in a text.
- Genre conventions: Formal and informal rules for particular kinds of texts and/or media that guide formatting, organization, and stylistic choices, e.g. lab reports, academic papers, poetry, webpages, or personal essays.
- Sources: Texts (written, oral, behavioral, visual, or other) that writers draw on as they work for a variety of purposes -- to extend, argue with, develop, define, or shape their ideas, for example.

WRITTEN COMMUNICATION VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	**Milestones**		**Benchmark**
	4	3	2	1
Context of and Purpose for Writing *Includes considerations of audience, purpose, and the circumstances surrounding the writing task(s).*	Demonstrates a thorough understanding of context, audience, and purpose that is responsive to the assigned task(s) and focuses all elements of the work.	Demonstrates adequate consideration of context, audience, and purpose and a clear focus on the assigned task(s) (e.g., the task aligns with audience, purpose, and context).	Demonstrates awareness of context, audience, purpose, and to the assigned tasks(s) (e.g., begins to show awareness of audience's perceptions and assumptions).	Demonstrates minimal attention to context, audience, purpose, and to the assigned tasks(s) (e.g., expectation of instructor or self as audience).
Content Development	Uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.	Uses appropriate, relevant, and compelling content to explore ideas within the context of the discipline and shape the whole work.	Uses appropriate and relevant content to develop and explore ideas through most of the work.	Uses appropriate and relevant content to develop simple ideas in some parts of the work.
Genre and Disciplinary Conventions *Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields (please see glossary).*	Demonstrates detailed attention to and successful execution of a wide range of conventions particular to a specific discipline and/or writing task (s) including organization, content, presentation, formatting, and stylistic choices	Demonstrates consistent use of important conventions particular to a specific discipline and/or writing task(s), including organization, content, presentation, and stylistic choices	Follows expectations appropriate to a specific discipline and/or writing task(s) for basic organization, content, and presentation	Attempts to use a consistent system for basic organization and presentation.
Sources and Evidence	Demonstrates skillful use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	Demonstrates consistent use of credible, relevant sources to support ideas that are situated within the discipline and genre of the writing.	Demonstrates an attempt to use credible and/or relevant sources to support ideas that are appropriate for the discipline and genre of the writing.	Demonstrates an attempt to use sources to support ideas in the writing.
Control of Syntax and Mechanics	Uses graceful language that skillfully communicates meaning to readers with clarity and fluency, and is virtually error-free.	Uses straightforward language that generally conveys meaning to readers. The language in the portfolio has few errors.	Uses language that generally conveys meaning to readers with clarity, although writing may include some errors.	Uses language that sometimes impedes meaning because of errors in usage.

READING VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Reading is "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow et al., 2002). (From www.rand.org/pubs/research_briefs/RB8024/index1.html)

Framing Language

To paraphrase Phaedrus, texts do not explain, nor answer questions about, themselves. They must be located, approached, decoded, comprehended, analyzed, interpreted, and discussed, especially complex academic texts used in college and university classrooms for purposes of learning. Historically, college professors have not considered the teaching of reading necessary other than as a "basic skill" in which students may require "remediation." They have assumed that students come with the ability to read and have placed responsibility for its absence on teachers in elementary and secondary schools.

This absence of reading instruction in higher education must, can, and will change, and this rubric marks a direction for this change. Why the change? Even the strongest, most experienced readers making the transition from high school to college have not learned what they need to know and do to make sense of texts in the context of professional and academic scholarship--to say nothing about readers who are either not as strong or as experienced. Also, readers mature and develop their repertoire of reading performances naturally during the undergraduate years and beyond as a consequence of meeting textual challenges. This rubric provides some initial steps toward finding ways to measure undergraduate students' progress along the continuum. Our intention in creating this rubric is to support and promote the teaching of undergraduates as readers to take on increasingly higher levels of concerns with texts and to read as one of "those who comprehend."

Readers, as they move beyond their undergraduate experiences, should be motivated to approach texts and respond to them with a reflective level of curiosity and the ability to apply aspects of the texts they approach to a variety of aspects in their lives. This rubric provides the framework for evaluating both students' developing relationship to texts and their relative success with the range of texts their coursework introduces them to. It is likely that users of this rubric will detect that the cell boundaries are permeable, and the criteria of the rubric are, to a degree, interrelated.

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Analysis: The process of recognizing and using features of a text to build a more advanced understanding of the meaning of a text. (Might include evaluation of genre, language, tone, stated purpose, explicit or implicit logic (including flaws of reasoning), and historical context as they contribute to the meaning of a text.]
- Comprehension: The extent to which a reader "gets" the text, both literally and figuratively. Accomplished and sophisticated readers will have moved from being able to "get" the meaning that the language of the texte provides to being able to "get" the implications of the text, the questions it raises, and the counterarguments one might suggest in response to it. A helpful and accessible discussion of 'comprehension' is found in Chapter 2 of the RAND report, Reading for Understanding: www.rand.org/pubs/monograph_reports/MR1465/MR1465.ch2.pdf.
- Epistemological lens: The knowledge framework a reader develops in a specific discipline as s/he moves through an academic major (e.g., essays, textbook chapters, literary works, journal articles, lab reports, grant proposals, lectures, blogs, webpages, or literature reviews, for example). The depth and breadth of this knowledge provides the foundation for independent and self-regulated responses to the range of texts in any discipline or field that students will encounter.
- Genre: A particular kind of "text" defined by a set of disciplinary conventions or agreements learned through participation in academic discourse. Genre governs what texts can be about, how they are structured, what to expect from them, what can be done with them, how to use them
- Interpretation: Determining or construing the meaning of a text or part of a text in a particular way based on textual and contextual information.
- Interpretive Strategies: Purposeful approaches from different perspectives, which include, for example, asking clarifying questions, building knowledge of the context in which a text was written, visualizing and considering counterfactuals (asking questions that challenge the assumptions or claims of the text, e.g., What might our country be like if the Civil War had not happened? How would Hamlet be different if Hamlet had simply killed the King?).
- Multiple Perspectives: Consideration of how text-based meanings might differ depending on point of view.
- Parts: Titles, headings, meaning of vocabulary from context, structure of the text, important ideas and relationships among those ideas.
- Relationship to text: The set of expectations and intentions a reader brings to a particular text or set of texts.
- Searches intentionally for relationships: An active and highly-aware quality of thinking closely related to inquiry and research.
- Takes texts apart: Discerns the level of importance or abstraction of textual elements and sees big and small pieces as parts of the whole meaning (compare to Analysis above).
- Metacognition: This is not a word that appears explicitly anywhere in the rubric, but it is implicit in a number of the descriptors, and is certainly a term that we find frequently in discussions of successful and rich learning.. Metacognition, (a term typically attributed to the cognitive psychologist J.H. Flavell) applied to reading refers to the awareness, deliberateness, and reflexivity defining the activities and strategies that readers must control in order to work their ways effectively through different sorts of texts, from lab reports to sonnets, from math texts to historical narratives, or from grant applications to graphic novels, for example. Metacognition refers here as well to an accomplished reader's ability to consider the ethos reflected in any such text; to know that one is present and should be considered in any use of, or response to a text.

READING VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Reading is "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow et al., 2002). (From www.rand.org/pubs/research_briefs/RB8024/index1.html)

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	**Milestones**		**Benchmark**
	4	3	2	1
Comprehension	Recognizes possible implications of the text for contexts, perspectives, or issues beyond the assigned task within the classroom or beyond the author's explicit message (e.g., might recognize broader issues at play, or might pose challenges to the author's message and presentation).	Uses the text, general background knowledge, and/or specific knowledge of the author's context to draw more complex inferences about the author's message and attitude.	Evaluates how textual features (e.g., sentence and paragraph structure or tone) contribute to the author's message; draws basic inferences about context and purpose of text.	Apprehends vocabulary appropriately to paraphrase or summarize the information the text communicates.
Genres	Uses ability to identify texts within and across genres, monitoring and adjusting reading strategies and expectations based on generic nuances of particular texts.	Articulates distinctions among genres and their characteristic conventions.	Reflects on reading experiences across a variety of genres, reading both with and against the grain experimentally and intentionally.	Applies tacit genre knowledge to a variety of classroom reading assignments in productive, if unreflective, ways.
Relationship to Text *Making meanings with texts in their contexts*	Evaluates texts for scholarly significance and relevance within and across the various disciplines, evaluating them according to their contributions and consequences.	Uses texts in the context of scholarship to develop a foundation of disciplinary knowledge and to raise and explore important questions.	Engages texts with the intention and expectation of building topical and world knowledge.	Approaches texts in the context of assignments with the intention and expectation of finding right answers and learning facts and concepts to display for credit.
Analysis *Interacting with texts in parts and as wholes*	Evaluates strategies for relating ideas, text structure, or other textual features in order to build knowledge or insight within and across texts and disciplines.	Identifies relations among ideas, text structure, or other textual features, to evaluate how they support an advanced understanding of the text as a whole.	Recognizes relations among parts or aspects of a text, such as effective or ineffective arguments or literary features, in considering how these contribute to a basic understanding of the text as a whole.	Identifies aspects of a text (e.g., content, structure, or relations among ideas) as needed to respond to questions posed in assigned tasks.
Interpretation *Making sense with texts as blueprints for meaning*	Provides evidence not only that s/he can read by using an appropriate epistemological lens but that s/he can also engage in reading as part of a continuing dialogue within and beyond a discipline or a community of readers.	Articulates an understanding of the multiple ways of reading and the range of interpretive strategies particular to one's discipline(s) or in a given community of readers.	Demonstrates that s/he can read purposefully, choosing among interpretive strategies depending on the purpose of the reading.	Can identify purpose(s) for reading, relying on an external authority such as an instructor for clarification of the task.
Reader's Voice *Participating in academic discourse about texts*	Discusses texts with an independent intellectual and ethical disposition so as to further or maintain disciplinary conversations.	Elaborates on the texts (through interpretation or questioning) so as to deepen or enhance an ongoing discussion.	Discusses texts in structured conversations (such as in a classroom) in ways that contribute to a basic, shared understanding of the text.	Comments about texts in ways that preserve the author's meanings and link them to the assignment.

QUANTITATIVE LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Quantitative Literacy (QL) – also known as Numeracy or Quantitative Reasoning (QR) – is a "habit of mind," competency, and comfort in working with numerical data. Individuals with strong QL skills possess the ability to reason and solve quantitative problems from a wide array of authentic contexts and everyday life situations. They understand and can create sophisticated arguments supported by quantitative evidence and they can clearly communicate those arguments in a variety of formats (using words, tables, graphs, mathematical equations, etc., as appropriate).

Quantitative Literacy Across the Disciplines

Current trends in general education reform demonstrate that faculty are recognizing the steadily growing importance of Quantitative Literacy (QL) in an increasingly quantitative and data-dense world. AAC&U's recent survey showed that concerns about QL skills are shared by employers, who recognize that many of today's students will need a wide range of high level quantitative skills to complete their work responsibilities. Virtually all of today's students, regardless of career choice, will need basic QL skills such as the ability to draw information from charts, graphs, and geometric figures, and the ability to accurately complete straightforward estimations and calculations.

Preliminary efforts to find student work products which demonstrate QL skills proved a challenge in this rubric creation process. It's possible to find pages of mathematical problems, but what those problem sets don't demonstrate is whether the student was able to think about and understand the meaning of her work. It's possible to find research papers that include quantitative information, but those papers often don't provide evidence that allows the evaluator to see how much of the thinking was done by the original source (often carefully cited in the paper) and how much was done by the student herself, or whether conclusions drawn from analysis of the source material are even accurate.

Given widespread agreement about the importance of QL, it becomes incumbent on faculty to develop new kinds of assignments which give students substantive, contextualized experience in using such skills as analyzing quantitative information, representing quantitative information in appropriate forms, completing calculations to answer meaningful questions, making judgments based on quantitative data and communicating the results of that work for various purposes and audiences. As students gain experience with those skills, faculty must develop assignments that require students to create work products which reveal their thought processes and demonstrate the range of their QL skills.

This rubric provides for faculty a definition for QL and a rubric describing four levels of QL achievement which might be observed in work products within work samples or collections of work. Members of AAC&U's rubric development team for QL hope that these materials will aid in the assessment of QL – but, equally important, we hope that they will help institutions and individuals in the effort to more thoroughly embed QL across the curriculum of colleges and universities.

Framing Language

This rubric has been designed for the evaluation of work that addresses quantitative literacy (QL) in a substantive way. QL is not just computation, not just the citing of someone else's data. QL is a habit of mind, a way of thinking about the world that relies on data and on the mathematical analysis of data to make connections and draw conclusions. Teaching QL requires us to design assignments that address authentic, data-based problems. Such assignments may call for the traditional written paper, but we can imagine other alternatives: a video of a PowerPoint presentation, perhaps, or a well designed series of web pages. In any case, a successful demonstration of QL will place the mathematical work in the context of a full and robust discussion of the underlying issues addressed by the assignment.

Finally, QL skills can be applied to a wide array of problems of varying difficulty, confounding the use of this rubric. For example, the same student might demonstrate high levels of QL achievement when working on a simplistic problem and low levels of QL achievement when working on a very complex problem. Thus, to accurately assess a students QL achievement it may be necessary to measure QL achievement within the context of problem complexity, much as is done in diving competitions where two scores are given, one for the difficulty of the dive, and the other for the skill in accomplishing the dive. In this context, that would mean giving one score for the complexity of the problem and another score for the QL achievement in solving the problem.

AAC&U Association of American Colleges and Universities

QUANTITATIVE LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

Definition

Quantitative Literacy (QL) – also known as Numeracy or Quantitative Reasoning (QR) – is a "habit of mind," competency, and comfort in working with numerical data. Individuals with strong QL skills possess the ability to reason and solve quantitative problems from a wide array of authentic contexts and everyday life situations. They understand and can create sophisticated arguments supported by quantitative evidence and they can clearly communicate those arguments in a variety of formats (using words, tables, graphs, mathematical equations, etc., as appropriate).

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	**Milestones**		
	4	3	2	1
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*	Provides accurate explanations of information presented in mathematical forms. *For instance, accurately explains the trend data shown in a graph.*	Provides somewhat accurate explanations of information presented in mathematical forms, but occasionally makes minor errors related to computations or units. *For instance, accurately explains trend data shown in a graph, but may miscalculate the slope of the trend line.*	Attempts to explain information presented in mathematical forms, but draws incorrect conclusions about what the information means. *For example, attempts to explain the trend data shown in a graph, but will frequently misinterpret the nature of that trend, perhaps by confusing positive and negative trends.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.	Competently converts relevant information into an appropriate and desired mathematical portrayal.	Completes conversion of information but resulting mathematical portrayal is only partially appropriate or accurate.	Completes conversion of information but resulting mathematical portrayal is inappropriate or inaccurate.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem.	Calculations attempted are either unsuccessful or represent only a portion of the calculations required to comprehensively solve the problem.	Calculations are attempted but are both unsuccessful and are not comprehensive.
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for competent judgments, drawing reasonable and appropriately qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for workmanlike (without inspiration or nuance, ordinary) judgments, drawing plausible conclusions from this work.	Uses the quantitative analysis of data as the basis for tentative, basic judgments, although is hesitant or uncertain about drawing conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.	Explicitly describes assumptions and provides compelling rationale for why assumptions are appropriate.	Explicitly describes assumptions.	Attempts to describe assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.	Uses quantitative information in connection with the argument or purpose of the work, though data may be presented in a less than completely effective format or some parts of the explication may be uneven.	Uses quantitative information, but does not effectively connect it to the argument or purpose of the work.	Presents an argument for which quantitative evidence is pertinent, but does not provide adequate explicit numerical support. (May use quasi-quantitative words such as "many," "few," "increasing," "small," and the like in place of actual quantities.)

INFORMATION LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

The ability to know when there is a need for information, to be able to identify, locate, evaluate, and effectively and responsibly use and share that information for the problem at hand. - Adopted from the National Forum on Information Literacy

Framing Language

This rubric is recommended for use evaluating a collection of work, rather than a single work sample in order to fully gauge students' information skills. Ideally, a collection of work would contain a wide variety of different types of work and might include: research papers, editorials, speeches, grant proposals, marketing or business plans, PowerPoint presentations, posters, literature reviews, position papers, and argument critiques to name a few. In addition, a description of the assignments with the instructions that initiated the student work would be vital in providing the complete context for the work. Although a student's final work must stand on its own, evidence of a student's research and information gathering processes, such as a research journal/diary, could provide further demonstration of a student's information proficiency and for some criteria on this rubric would be required.

Information Literacy VALUE Rubric

for more information, please contact value@aacu.org

Definition

The ability to know when there is a need for information, to be able to identify, locate, evaluate, and effectively and responsibly use and share that information for the problem at hand. - The National Forum on Information Literacy

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone	Milestones		Benchmark
	4	3	2	1
Determine the Extent of Information Needed	Effectively defines the scope of the research question or thesis. Effectively determines key concepts. Types of information (sources) selected directly relate to concepts or answer research question.	Defines the scope of the research question or thesis completely. Can determine key concepts. Types of information (sources) selected relate to concepts or answer research question.	Defines the scope of the research question or thesis incompletely (parts are missing, remains too broad or too narrow, etc.). Can determine key concepts. Types of information (sources) selected partially relate to concepts or answer research question.	Has difficulty defining the scope of the research question or thesis. Has difficulty determining key concepts. Types of information (sources) selected do not relate to concepts or answer research question.
Access the Needed Information	Accesses information using effective, well-designed search strategies and most appropriate information sources.	Accesses information using variety of search strategies and some relevant information sources. Demonstrates ability to refine search.	Accesses information using simple search strategies, retrieves information from limited and similar sources.	Accesses information randomly, retrieves information that lacks relevance and quality.
Evaluate Information and its Sources Critically	Thoroughly (systematically and methodically) analyzes own and others' assumptions and carefully evaluates the relevance of contexts when presenting a position.	Identifies own and others' assumptions and several relevant contexts when presenting a position.	Questions some assumptions. Identifies several relevant contexts when presenting a position. May be more aware of others' assumptions than one's own (or vice versa).	Shows an emerging awareness of present assumptions (sometimes labels assertions as assumptions). Begins to identify some contexts when presenting a position.
Use Information Effectively to Accomplish a Specific Purpose	Communicates, organizes and synthesizes information from sources to fully achieve a specific purpose, with clarity and depth	Communicates, organizes and synthesizes information from sources. Intended purpose is achieved.	Communicates and organizes information from sources. The information is not yet synthesized, so the intended purpose is not fully achieved.	Communicates information from sources. The information is fragmented and/or used inappropriately (misquoted, taken out of context, or incorrectly paraphrased, etc.), so the intended purpose is not achieved.
Access and Use Information Ethically and Legally	Students use correctly all of the following information use strategies (use of citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; distinguishing between common knowledge and ideas requiring attribution) and demonstrate a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Students use correctly three of the following information use strategies (use of citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; distinguishing between common knowledge and ideas requiring attribution) and demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Students use correctly two of the following information use strategies (use of citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; distinguishing between common knowledge and ideas requiring attribution) and demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Students use correctly one of the following information use strategies (use of citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; distinguishing between common knowledge and ideas requiring attribution) and demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.

Appendix D

List of Test Evaluation Materials

Test	Review materials	URL link
CAT	Critical Thinking Assessment Test - Overview	http://www.tntech.edu/cat/home/
	Critical Thinking Assessment Test – Features	http://www.tntech.edu/cat/overview/
	Critical Thinking Assessment Test – Technical	http://www.tntech.edu/cat/technical/
	Critical Thinking Assessment Test – Skills	http://www.tntech.edu/cat/skills/
	Critical Thinking Assessment Test – Development	http://www.tntech.edu/cat/development/
	General features of the CAT Test – Test specifications	http://commons.gc.cuny.edu/groups/task-force-on-assessment/documents/CAT Info - Test Specs
	Sample CAT Institutional Report	http://commons.gc.cuny.edu/groups/task-force-on-assessment/documents/CAT Info - Sample Results Report
	Faculty Driven Assessment of Critical Thinking: National Dissemination of the CAT Instrument (Barry Stein, 2010)	http://www2.tntech.edu/cat/presentations/CISSE2010.pdf
	Assessing Critical Thinking in STEM and Beyond (Barry Stein A. H., 2007)	http://www.tntech.edu/images/stories/cp/cat/reports/Innovationschapter.pdf
	Project CAT: Assessing Critical Thinking Skills (Barry Stein A. H., 2006)	http://www.tntech.edu/images/stories/cp/cat/reports/ProjectCat_NSF_NationalSTEMAssessmentConference.pdf
CAAP	CAAP Technical Handbook 2008-2009	http://www.act.org/caap/resources.html
	Test Validity Study (TVS) of the CAAP, MAAP and CLA	http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf
	CAAP Student Guide	http://www.act.org/caap/pdf/userguide.pdf
	CAAP Guide to Successful General Education Outcomes Assessment	http://www.act.org/caap/resources.html
	List of CAAP users broken down by college type	http://www.act.org/caap/resources.html
	Use of CAAP for the VSA	http://www.act.org/caap/pdf/10_11VSAGuidelines.pdf

Test	Review materials	URL link
	User Norms 2008-2009	http://www.act.org/caap/resources.html
	ACT College Learning Outcomes Assessment Planning Guide	http://www.act.org/caap/pdf/CAAP_Booklet.pdf
	Sample Insitutional Summary Report	http://www.act.org/caap/report_summary.html
CLA	Architecture of the CLA Tasks	http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf
	CLA Scoring Criteria	http://www.collegiatelearningassessment.org/files/CLAScoringCriteria.pdf
	CAE Board Statement onAppropriate Uses of the CLA	http://www2.ed.gov/about/bdscomm/list/hiedfuture/4th-meeting/benjamin.pdf
	Sample 2009-2010 CLA Institutional Report	http://www.collegiatelearningassessment.org/files/CLA_0910_Report_University_College2.pdf
	CLA Frequently Asked Technical Questions	http://www.collegiatelearningassessment.org/files/CLA_Technical_FAQs.pdf
	Incentives, Motivation, and Performance on a Low-Stakes Test of College Learning (Steedle, 2010)	http://www.collegiatelearningassessment.org/files/Steedle_2010_Incentives_Motivation_and_Performance_on_a_Low-Stakes_Test_of_College_Learning.pdf
	Improving the Reliability and Interpretability of Value-Added Scores for Post-Secondary Institutional Assessment Programs (Steedle, 2010)	http://www.collegiatelearningassessment.org/files/Steedle_2010_Improving_the_Reliability_and_Interpretability_of_Value-Added_Scores_for_Post-Secondary_Institutional_Assessment_Programs.pdf
	Sample 2009-2010 CCLA Institutional	http://www.collegiatelearningassessment.org/ - Request directly from CAE.
	Sample score report showing breakdown by dimension	http://www.collegiatelearningassessment.org/ - Request directly from CAE.
ETS PP	ETS Proficiency Profile Overview	http://www.ets.org/proficiencyprofile/about
	ETS Proficiency Profile User's Guide, 2010	http://www.ets.org/s/proficiencyprofile/pdf/Users_Guide.pdf

Test	Review materials	URL link
	Validity of the Measure of Academic Proficiency and Progress (MAPP) Test	http://www.ets.org/s/mapp/pdf/5018.pdf
	Validity of the Academic Profile	http://www.ets.org/s/proficiencyprofile/pdf/Validity_Academic_Profile.pdf
	ETS Proficiency Profile Content	http://www.ets.org/proficiencyprofile/about/content/
	ETS Proficiency Profile Sample Questions	http://www.ets.org/s/proficiencyprofile/pdf/sampleques.pdf
	ETS Proficiency Profile Format	http://www.ets.org/proficiencyprofile/test_administration/format/
	Procedures to Administer the ETS Proficiency Profile	http://www.ets.org/proficiencyprofile/test_administration/procedures/
	ETS Proficiency Profile Proctor Administrator Manual	http://www.ets.org/s/proficiencyprofile/pdf/Proctor_Manual.pdf
	ETS Proficiency Profile Supervisor’s Manual	http://www.ets.org/s/proficiencyprofile/pdf/Supervisor_Manual_Paper_Pencil.pdf
	ETS Proficiency Profile Scores	http://www.ets.org/proficiencyprofile/scores/
	Standard Reports for the Abbreviated Form of the ETS Proficiency Profile Tests	http://www.ets.org/s/proficiencyprofile/pdf/Users_Guide_Abbreviated_Reports.pdf
	Standard Reports for the Standard Form of the ETS Proficiency Score	http://www.ets.org/s/proficiencyprofile/pdf/Users_Guide_Standard_Reports.pdf
	ETS Proficiency Profile Score Usage	http://www.ets.org/proficiencyprofile/scores/usage/
	ETS Proficiency Profile Score Reports	http://www.ets.org/proficiencyprofile/scores/reports/
	ETS Proficiency Profile Comparative Data	http://www.ets.org/proficiencyprofile/scores/compare_data/
	Sophomore – Doctoral/Research Universities	http://www.ets.org/s/proficiencyprofile/pdf/CredH_Carn1_AllTabs.pdf
	Sophomore – Master’s	http://www.ets.org/s/proficiencyprofile/pdf/CredH_Carn2_AllTabs.pdf

Test	Review materials	URL link
	Sophomore – Baccalaurate	http://www.ets.org/s/proficiencyprofile/pdf/CredH_Carn3_AllTabs.pdf
	Sophomore – Associates	http://www.ets.org/s/proficiencyprofile/pdf/CredJ_Carn4_AllTabs.pdf
	Sophomore – All Instutions	http://www.ets.org/s/proficiencyprofile/pdf/CredJ_CarnA_AllTabs.pdf
	EETS Proficiency Profile Case Studies	http://www.ets.org/proficiencyprofile/case_studies/
	ETS PP Case Study, Albertus Magnus College	http://www.ets.org/s/proficiencyprofile/pdf/AlbertusMagnus_casestudy.pdf
	Pricing & Ordering	http://www.ets.org/proficiencyprofile/pricing/
	ETS Proficiency Profile test – Review copy	Requested directly from ETS - mailto:highered@ets.org

Appendix E

Team Presentation Summaries

Name of Test: CAT **Sample presentation to Task Force**

Specification	**Basis of evaluation**	**Test information**
Test purpose and design are consistent with the Task Force charge and with CUNY learning objectives	Tests are to be used to: • Measure learning gains • Benchmark college performance against that of comparable institutions outside CUNY • Improve teaching and learning throughout CUNY Core learning outcomes across CUNY: • Reading • Critical thinking • Written communication • Quantitative literacy • Information literacy	CAT is designed to assess and promote the improvement of critical thinking and real-world problem solving skills. • Sensitive to class and course effects • Suitable for value-added analyses • National norms CAT is scored by the institution's own faculty. Faculty are encouraged to use the CAT as a model for developing authentic assessments and learning activities that improve students' critical thinking. From Bloom's Taxonomy of Cognitive Skills, the CAT is focused on those higher than the knowledge level: • Knowledge (rote retention) • Comprehension • Application • Analysis • Synthesis • Evaluation **Skills assessed by CAT:** *Evaluating Information* • Separate factual information from inferences. • Interpret numerical relationships in graphs. • Understand the limitations of correlational data. • Evaluate evidence and identify inappropriate conclusions. *Creative Thinking*

Specification	**Basis of evaluation**	**Test information**
		• Identify alternative interpretations for data or observations. • Identify new information that might support or contradict a hypothesis. • Explain how new information can change a problem. *Learning and problem Solving* • Separate relevant from irrelevant information. • Integrate information to solve problems. • Learn and apply new information. • Use mathematical skills to solve real-world problems. *Communication* • Communicate ideas effectively.
Psychometric quality	**Content Validity** Do the test tasks require the test-taker to use the skills and competencies described in the relevant LEAP VALUE rubrics? Does the scoring of the tasks reflect the progression of rubric skill levels?	General Features of the CAT Test Sample Disclosed Question There are 15 questions on the CAT that ask the test taker to: 1. Summarize the pattern of results in a graph without making inappropriate inferences. 2. Evaluate how strongly correlational-type data supports a hypothesis. 3. Provide alternative explanations for a pattern of results that has many possible causes. 4. Identify additional information needed to evaluate a hypothesis. 5. Evaluate whether spurious information strongly supports a hypothesis. 6. Provide alternative explanations for spurious associations. 7. Identify additional information needed to evaluate a hypothesis. 8. Determine whether an invited inference is supported by specific information. 9. Provide relevant alternative interpretations for a specific set of results. 10. Separate relevant from irrelevant information when solving a real-word problem. 11. Use and apply relevant information to evaluate a problem. 12. Use basic mathematical skills to help solve a

Specification	Basis of evaluation	Test information
		real-world problem. 13. Identify suitable solutions for a real-world problem using relevant information. 14. Identify and explain the best solution for a real-world problem using relevant information. 15. Explain how changes in a real-world problem situation might affect the solution.
	External Criterion Validity What evidence is there that the test detects learning gains at the institutional level?	CAT with ACT .501 SAT .516 Academic Profile .562 GPA .295 CCTST .645 CAAP .691 NESSE memorizing -.341 # of books .277 Thinking Crit .244 Capstone .231

Specification	Basis of evaluation	Test information
	Validity generalization Does the test developer clearly set forth how test scores are intended to be interpreted and used? Are there other participating colleges in its database of results that are comparable to those of CUNY and can serve in a benchmarking function?	Scores reported in raw score scale. Breakdown of sample by gender, college year, age, English proficiency, race. Point distribution by item. Mean performance by item – raw score & % of total Comparison of item means with national sample Comparison of pre and post test scores.
	Score accuracy for institution-level comparisons **Reliability** What evidence is there for stability of scores over different items or forms of the test? If tests are scored by humans, what is the inter-rater reliability of scores? Does the test developer provide guidance for sampling covariates, e.g., ESL status, gender, race?	Test-retest .80 Inter-rater .82 Internal constituency .695 No differential item functioning by culture. Controlling for SAT, GPA, and ESL status showed no gender or race effects on CAT performance.

Specification	Basis of evaluation	Test information
Test Development & Logistics	Is there a technical manual that describes the test development process, test specifications, scoring rubrics, field testing, and availability of multiple parallel forms? Is there a test administration manual that describes the testing protocol and any special testing requirements, e.g., online administration, administrator certification, test-taker preparation materials, scoring protocols How are test results communicated to the colleges? What guidance is there for score interpretation with respect to benchmarking and learning gains?	Detailed scoring guide. Faculty scoring of tests. Score Report

Task Force on Learning Outcomes Assessment
Test Evaluation Worksheet

Name of Test: CAAP **Evaluator(s): Ray, Mosen, David**

Specification	Basis of evaluation	Test information
Test purpose and design are consistent with the Task Force charge and with CUNY learning objectives	Tests are to be used to: • Measure learning gains • Benchmark college performance against that of comparable institutions outside CUNY • Improve teaching and learning throughout CUNY Core learning outcomes across CUNY: • Reading • Critical thinking • Written communication • Quantitative literacy • Information literacy	The Collegiate Assessment of Academic Proficiency (CAAP) was designed to assess academic achievement in Reading, Writing, Mathematics, Science, and Critical Thinking. The tests can be used modularly to test each area separately. Purpose: CAAP tests are used by both 2 and 4-year institutions to measure the academic progress of students and to help determine the educational development of individual students. • Group Basis – 1) to help institutions improve their instructional programs by measuring student progress in acquiring core academic skills; 2) to provide evidence that gen ed objectives are being met, document change in students' performance levels from one educational point to another; 3) provide differential performance comparisons in gen ed instructional programs within an institution; 4) compare local performance with that of other populations (e.g., similar insitutions across the nation). • Individual Basis – 1) to indicate a student's readiness for further education; 2) to identify interventions needed for subsequent student success; and 3) to assure some specified level of skill mastery prior to graduation or program completion. Note: Care should be taken when using CAAP results for these purposes. Local research should be conducted on the specific application of the CAAP program whenever possible. In addition, CAAP results should be used in a manner that will benefit students as well as institutions. Aside from information literacy the match with CUNY objectives is acceptable.

Specification	Basis of evaluation	Test information
Psychometric quality	**Content Validity** Do the test tasks require the test-taker to use the skills and competencies described in the relevant LEAP VALUE rubrics? Does the scoring of the tasks reflect the progression of rubric skill levels?	Reading test (Item D, p.50) • Referring skills • Reasoning skills • Sample booklet (Commons, doc 6o) Writing Skills (Item D, p.48) • Punctuation • Grammar • Sentence structure • Organization • Strategy • Style • Sample booklet (Commons, doc 6p) Writing Essay (Item D, p.53) • Formulating an assertion • Supporting the assertion • Organizing major ideas • Clear effective language • Sample booklet (Commons, doc 6q) Mathematics • Prealgebra • Elementary Algebra • Intermediate Algebra • Coordinate Geometry • Sample booklet (Commons, doc 6r) Science • Understanding • Analyzing • Generalizing • Sample booklet (Commons, doc, 6s) Critical Thinking • Analysis of Elements of Arguments • Evaluation of Arguments • Extension of Arguments • Sample booklet (Commons, doc. 6t) **Content Validity Assessment:** • All subtests involve the reading skill. • All VALUE rubrics require a higher level of student thought or production than is required by the CAAP test items. Except for the writing essay, and most parts of the mathematics sections, the test items are basically

Specification	Basis of evaluation	Test information
	External Criterion Validity What evidence is there that the test detects learning gains at the institutional level?	**See Chapter 7, Tech Handbook** **Criterion-Related Validity Evidence for CAAP Scores** *****CAAP as a measure of students' academic knowledge & skills:** **If sophomore CAAP scores and college GPA are both considered reliable and valid measures of academic skills acquired during the first two years of college, then there should be a statistical relationship between these variables. To test this assumption, sophomore CAAP scores were used to model sophomore GPAs.** **The median (cross-institutions) correlation between:** **Cumulative English GPA and CAAP was 0.37 (range of 0.26 to 0.57)** **Cumulative Math GPA and CAAP was 0.34 (range 0.11 to 0.40)** **Overall GPA with CAAP writing skills was 0.36** **Overall GPA with CAAP Math skills was 0.35** **Overall GPA with CAAP Reading skills was 0.38** **Overall GPA with CAAP Critical Thinking was 0.34** *****CAAP as a predictive measure:** **If junior course grades and GPAs are reliable and valid measures of junior-year academic performance, and if sophomore CAAP scores are valid measures of the skills needed to succeed in the junior year, then there should be a statistical relationship between sophomore CAAP score and junior-year grades and GPAs (use of regression model).** **The median (across institutions) correlation between junior GPAs and corresponding sophomore CAAP test scores were all moderately positive** **CAAP Critical Thinking with Junior English GPA was**

Specification	Basis of evaluation	Test information
		0.32 **CAAP Writing with Junior English GPA was 0.25** **CAAP Reading score with junior GPA was 0.25** **CAAP Math score with Junior Math GPA was 0.23** **Junior cumulative overall GPA, was somewhat more strongly associated with CAAP objective test scores that was junior non-cumulative overall GPA (e.e., median correlations between these GPA variables and CAAP Critical Thinking scores were 0.35 & 0.26, respectively.**
	Validity generalization Does the test developer clearly set forth how test scores are intended to be interpreted and used? Are there other participating colleges in its database of results that are comparable to those of CUNY and can serve in a benchmarking function?	*****CAAP as a measure of educational change:** **If CAAP scores are valid for measuring change over time, then CAAP score of sophomores should be greater than the CAAP scores of the freshmen.** **Note: Comparisons were made without any adjustment for academic skills or persistence. Using unadjusted cross-sectional data can tend to overestimate change.** **The CAAP scores were compared using ANCOVA (institution and educational level were the main effects, and the ACT Assessment Composite score was the covariate. The ANCOVA analyses were based on data for persisting students only, pooled across institutions.** **Results: Averaged scores on the CAAP objective tests increased from the freshmen to the sophomore year.**
	Score accuracy for institution-level comparisons **Reliability** What evidence is there for stability of scores over different items or forms of the test?	**Reliability is an estimate of the consistency of test scores across repeated measurements. The Kuder-Richardson Formula 20 (K-R 20) reliability estimates are reported in Table 4.5 for two forms of the CAAP examinations (Tech. handbook, page 15 & 16)** **Test Validity Study (TVS) Report:** **Three assessment of collegiate learning were**

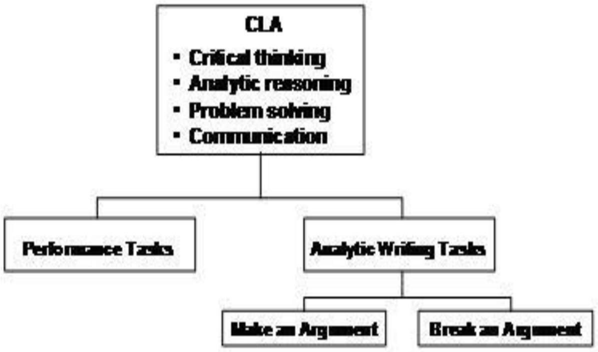
Specification	Basis of evaluation	Test information
	If tests are scored by humans, what is the inter-rater reliability of scores? Does the test developer provide guidance for sampling covariates, e.g., ESL status, gender, race?	**administered as part of the Test Validity Study (TVS): CAAP, CLA, & MAPP. A total of 13 tests administered at each of the study's 13 schools between Aug. 2008 and Nov. 2008. Each of the 13 campuses was responsible for recruiting a sample of 46 freshman and 46 seniors.** **Conducted analyses on student- and school-level data. Student-level data can be used to identify remediation needs, whereas school-level data may be used to inform policy, resource allocation, and programmatic decisions. In the report, the authors attempt to answer three questions:** **First, we asked whether scores from tests that purport to measure the same construct (critical thinking, reading, etc.) and employ the same response format (MC or constructed-response) are correlated higher with each other that with tests that measures different constructs and/or employ a different response format. A high positive correlation between two tests indicates that schools that obtain high score on one test also tend to obtain high scores on the other test. These correlations were computed separately using freshman class means and senior class means, and the two were averaged. See Table 2b (school-level matrix with standard correlation shown above the diagonal and reliabilities shown on the diagonal.** **To evaluate the simultaneous effects of construct and response format on the correlations, average correlations with other measures were computed and arranged in Table 3-B (school-level data). As expected, the highest correlations appear in the "same construct-same format" column, and the lowest correlations tend to appear in the "different construct, different format" column. Comparing the first and the third data columns provides an indicator or the effect of construct (holding response format constant).** **Second, is the difference in average scores between freshmen and seniors related to the construct tested, response format, or test's publisher?**

Specification	Basis of evaluation	Test information
		By creating an index (effect size), we could measure score gains between freshman & seniors in unit of SD. Seniors had higher mean scores than freshman on all the tests except for the math module. Note the effect sizes were not significantly different from zero. (TVS Report, Page 27) **Adjusted Effect Size: Controlling for differences in entering abilities** **Third, What are the reliabilities of school-level scores on different tests of college learning?** **School-level reliabilities reflect the consistency of school's mean score across theoretical repeated examinations with different samples of students. Table 5 provides a summary of school-level reliability coefficients for the measures. Score reliability is not a major concern when using school level results with sample sizes comparable to those obtained for this study – score reliability was high on all 13 tests (mean was 0.87 and the lowest value was 0.75)** **Overall, when the school was the unit of analysis, there were very high correlations among all the measures, very high score reliabilities, and consistent effect sizes.**
Test Development & Logistics	Is there a technical manual that describes the test development process, test specifications, scoring rubrics, field testing, and availability of multiple parallel forms? Is there a test administration manual that describes the testing protocol and any special testing requirements, e.g., online administration, administrator certification, test-taker preparation materials, scoring protocols How are test results communicated to the colleges? What guidance is there for score interpretation with respect to benchmarking and learning gains?	ACT provides a comprehensive CAAP supervisor's Manual with step-by-step instructions on how to administer and interpret tests. Standard CAAP Reporting package consists of five components: the Institutional Summary Report, Student Score reports, The Student Roster Report, Certificates of Achievement, and the Score Report Interpretive Guide (See Appendix B) In addition, additional fee-based reports include Data CD & Combined Institutional Summary Report. The institutional Summary Report provides faculty with a "snapshot" of students' learning on a group basis at one point in time. The average score or quartile groupings can be used as performance indicators for

Specification	Basis of evaluation	Test information
		the institutions. Furthermore, ACT offers research reports that provide additional utility from CAAP test results – CAAP Linkage Reports demonstrates student “performance gain” and Content Analysis Reports provide information about student performance within specific content areas of a given test module
	Cost Test publisher’s stated costs Costs to be borne by the institution (e.g., computer lab, proctors, test scorers, etc.)	

Task Force on Learning Outcomes Assessment
Test Evaluation Worksheet

Name of Test: CLA **Evaluator(s): R.Fox, H.Everson, K. Barker, M. Edlin**

Specification	**Basis of evaluation**	**Test information**
Test purpose and design are consistent with the Task Force charge and with CUNY learning objectives	Tests are to be used to: • Measure learning gains • Benchmark college performance against that of comparable institutions outside CUNY • Improve teaching and learning throughout CUNY Core learning outcomes across CUNY: • Reading • Critical thinking • Written communication • Quantitative literacy • Information literacy  CLA • Critical thinking • Analytic reasoning • Problem solving • Communication Performance Tasks Analytic Writing Tasks Make an Argument Break an Argument	The Collegiate Learning Assessment (CLA) was as a performance assessment to measure reading comprehension, critical thinking, written communication, quantitative literacy, and information literacy. It was designed to permit comparisons within and between institutions, and to engage faculty in meaningful discussions of the quality of teaching and learning. The format of the CLA is such that it focuses on higher order thinking and reasoning skills, and presents assessment tasks that require students to analyze and interpret complex stimulus materials. (Not a multiple choice format assessment). The CLA includes three types of prompts within two task types: Performance Tasks (PT) and Analytic Writing Tasks (AWT). Students are randomly assigned to a task type and then to a prompt within that tasks, and uses a matrix sampling design to reduce the testing burden on individual students, and provide the institution with the benefits from the full breadth of task types.

Specification	Basis of evaluation	Test information
Psychometric quality	**Content Validity** Do the test tasks require the test-taker to use the skills and competencies described in the relevant LEAP VALUE rubrics? Does the scoring of the tasks reflect the progression of rubric skill levels?	Refer to Rick Fox's matrix of the content coverage. Problem solving, quantitative reasoning and written communications are assessed. The CLA relies on an automated scoring system that incorporates both analytic and holistic scoring. More information needed on the mechanics of scoring.
	External Criterion Validity What evidence is there that the test detects learning gains at the institutional level?	**There is substantial evidence, largely correlational, about the relationship of students' performance on the CLA and performance on other measures of college admissions tests, and grades in college.**

Specification	**Basis of evaluation**	**Test information**
	Validity generalization Does the test developer clearly set forth how test scores are intended to be interpreted and used? Are there other participating colleges in its database of results that are comparable to those of CUNY and can serve in a benchmarking function?	**The primary unit of analysis for the CLA is the institutional level. The aggregate scores appear to be useful for measuring growth over time, and for making comparisons across institutions.**
	Score accuracy for institution-level comparisons **Reliability** What evidence is there for stability of scores over different items or forms of the test? If tests are scored by humans, what is the inter-rater reliability of scores? Does the test developer provide guidance for sampling covariates, e.g., ESL status, gender, race?	**School level correlations** **CLA** with SAT/ACT .87 to .88 (Analytic Writing Tasks) .78 to .92 (Performance Tasks) **Student level correlations** SAT/ACT .40 to .53 (Analytic Writing Tasks) .55 to .72 (Performance Tasks) Substitute for SAT/ACT: Scholastic Level Exam (SLE) with ACT .68 SAT Verbal/Critical Reading .68 SAT Math .66 SAT Composite Equivalent .77

Specification	Basis of evaluation	Test information
Test Development & Logistics	Is there a technical manual that describes the test development process, test specifications, scoring rubrics, field testing, and availability of multiple parallel forms? Is there a test administration manual that describes the testing protocol and any special testing requirements, e.g., online administration, administrator certification, test-taker preparation materials, scoring protocols How are test results communicated to the colleges? What guidance is there for score interpretation with respect to benchmarking and learning gains?	Detailed scoring guide programmed for the Pearson scoring process, as most tests are machine scored. A detailed guide is used to train faculty to be CAE certified. "The CLA provides CLA-like tasks to college instructors so they can "teach to the test." With the criterion-sampling approach, "cheating" by teaching to the test is not a bad thing. If a person "cheats" by learning and practicing to solve complex, holistic, real-world problems, she has demonstrated the knowledge and skills that educators seek to develop in students."* Schools presented with a report, "your results" – tables, graphs – see 2009-2010 CLA Institutional Report.
	Cost Test publisher's stated costs Costs to be borne by the institution (e.g., computer lab, proctors, test scorers, etc.)	TBD.

Test Evaluation Worksheet

Name of Test: ETS Proficiency Profile **Evaluators: Dahlia, Karrin, Lisa & Ellen**

Test Purpose and Design:

The ETS Proficiency Profile (defined “ETS Test”) has been offered since 1990 as an assessment of general education learning outcomes in 2 and 4-year colleges and universities. According to the materials provided by ETS, the ETS Test was designed to “assist in the assessment of the outcomes of general education programs in order to improve the quality of instruction and learning” (Document A*, p. 4). The test purports to measure four core skills as developed through the completion of general education courses: critical thinking, reading, writing and mathematics. It does not offer assessment in quantitative literacy and information literacy.

The ETS Test offers the flexibility to be administered in two forms (Standard or Abbreviated) either proctored paper-and-pencil or online versions. The Standard Form is intended to provide information about individual students as well as groups of student; it includes 108 multiple choice questions to be administered either in a single two-hour session or in separate one-hour sessions. The Abbreviated Form is not intended to provide information about individual students; it includes 36 multiple choice questions and can provide information about groups of at least 50 students. The test offers the option of adding 50 locally authored multiple choice questions as well as an essay, which is analyzed by the *e-rater* computer program.

The ETS Test data can enable institutions to:

- Assess student growth in the core skills at different stages in their academic careers and identify skill areas for improvement or recruitment (using Standard form only).
- Conduct studies, such as cross-sectional and longitudinal, to assess student proficiency in core academic areas to determine strengths, weaknesses and opportunities for improvement of curriculum (using Standard form only).
- Compare performance against approx. 400 academic institutions nationwide either based on Carnegie classification or a customized selection of peer institutions (using either Standard or Abbreviated forms).
- Conduct trend analysis to evaluate improvement and overall learning outcomes (using either Standard or Abbreviated forms).

Test Purpose and Design Assessment:

- The ETS Test is partly consistent with our purpose in the areas of reading and writing.
- The ETS Test is only minimally consistent with our purpose in the area of critical thinking in that it addresses this skill only as a small part of the reading component.
- The ETS Test is not consistent with our purpose in the areas of quantitative literacy and information literacy.

Psychometric quality: Content validity:

The test questions measure students’ abilities in four areas (Document A*, pp. 4, 9-13; Document A2).

Reading

- Interpret the meaning of key terms
- Recognize the primary purpose of a passage
- Recognize explicitly presented information
- Make appropriate inferences
- Recognize rhetorical devices

Comments: Skills tested focus on reading comprehension. VALUE rubrics require a higher level of thought and production than is addressed by the ETS test items.

Writing

- Recognize the most grammatically correct revision of a clause, sentence or group of sentences
- Organize units of language for coherence and rhetorical effect
- Recognize and reword figurative language
- Organize elements of writing into larger units of meaning

Comments: Skills tested emphasize sentence level proficiency. VALUE rubrics require a higher level of thought and production than is addressed by the ETS test items. ETS most nearly addresses VALUE writing rubrics in the area of "control of syntax and mechanics"; however, the ability to identify and correct errors in others' writing does not equate to the ability to avoid errors in one's own work. The test does not require the development or expression of ideas.

Critical Thinking

- Distinguish between rhetoric and argumentation in a piece of nonfiction prose
- Recognize assumptions
- Recognize the best hypothesis to account for information presented
- Infer and interpret a relationship between variables
- Draw valid conclusions based on information presented

Comments: VALUE rubrics require a higher level of thought and production than is addressed by the ETS test items. For example, the ability to identify and evaluate a hypothesis in a reading is not the same as the ability to create a hypothesis on one's own, especially if it involves selecting, evaluating and synthesizing a variety of texts, evidence or other materials.

Mathematics

- Recognize and interpret mathematical terms
- Read and interpret tables and graphs
- Evaluate formulas
- Order and compare large and small numbers
- Interpret ratios, proportions, and percentages
- Read scientific measuring instruments
- Recognize and use equivalent mathematical formulas or expressions

Comments: Skills tested focus on calculation (arithmetic), rather than quantitative literacy. VALUE rubrics require a higher level of thought and production than is addressed by the ETS test items.

Content Validity Assessment:

- All subtests involve reading comprehension.
- All VALUE rubrics require a higher level of thought and production than is addressed by the ETS test items.

Psychometric quality: External criterion validity:

External criterion validity is based on comparing the measure in question (i.e., ETS test) with a variety of other measures to see how highly correlated they are with measures that we believe should be correlated. One can also examine how uncorrelated they are with measures we believe should not be correlated (discriminant validity). A problem is that we don't have a good measure of what we want to measure; therefore, we want to see some correlation with GPA but not perfect correlation, because we are trying to measure something that is distinct in certain ways. We have two sources: the VTS Report which examined all the tests, comparing them to one another; an ETS report that examined relationships to a variety of measures, such as GPA, major, etc. (Marr).

The TVS report reported the following: ETS correlations at the school level with other candidates tests of same skill were (from Table 2b): .93 (critical thinking with CAPP), .83 critical thinking with CLA PT), .93 (critical thinking with CLA CA), .86 (writing with CLA MA), .97 (writing with CAAP), .70 (writing with CAAP essay), .98 (Math with CAAP), .86 (Reading with CAAP).

The Marr study in 1995 got detailed longitudinal data on students from students in several 4-year colleges to examine the relationship of the various ETS scores with: percentage of core curriculum, percentage of advanced electives, class level, GPA, major area (i.e., business, education, humanities/arts, natural sciences, social sciences, math/engineering). They also measured correlations between skill areas. The student was the unit of analysis and all quantitative variables were made into categorical measures (e.g., 5 GPA categories, 4 core curriculum completion categories). (Technical note: They did this analysis with MANOVA, meaning that they jointly estimated the several dependent variable relationships; this would be equivalent to seemingly unrelated regression with the independent variables as dummies.) Most relationships were statistically significant (% of core curriculum, GPA, major field) with the directions as expected, although practical significance seemed moderate (e.g., going from none of the core curriculum to 100% of the core curriculum resulted in a 5 point increase in humanities (relative to a possible 30)). They also found that no effect of completing advanced electives, after removing the effect of increasing core curriculum.

Psychometric quality: Validity generalization:

ETS scores are provided in both relative terms ("scaled scores") and absolute terms ("proficiency classifications"). The scaled scores are normed—relative to a particular population. Implicitly, that population does not change over time, so that the scaled scores can be compared over time, although the composition of the population was not made clear. The scaled scores are available for each of the skills subscores and have a fair amount of potential sensitivity, with a 100 point range for the total score and 30 point ranges for the subscores.

The proficiency classifications, are based on absolute performance standards. There are three levels (1, 2 and 3) and students are rated within the level as being proficient, marginal or not proficient. (Note that critical thinking is only available at level 3.) The proficiency classifications form a well-designed scale so that students proficient at one level must be proficient at the lower level. Obviously, there is less (potential) sensitivity in the proficiency classifications, but they have the advantage of being absolute measures.

For benchmarking levels , ETS provides reports for various populations of students composed of particular lists of named schools. There are populations separated by student stage (i.e., entering freshmen, finishing freshmen, sophomore, junior, senior) and Carnegie institution classification. The unit of analysis for the reports is the student, not the school—the students from the schools are pooled. The demographics of the student groups are provided and they look very different from those of CUNY. For example, in the population of Sophomores in Associate Colleges that ETS provided, only 7% of students had a language other than English as the best language, with a full 89% saying that English was their best language and 4% stating that both were equal; only 5% were Hispanic; a full 77% were White; only 22% were in school part-time. However, based on an e-mail, ETS states that they could construct reports using institutions of characteristics we requested that would be more comparable to CUNY.

For the various comparison populations of schools, ETS provides the complete distributions (histograms with 1-point width bins for most of the range) of scaled scores and the complete distributions of proficiency classifications. Note that the unit of analysis in the reports is the student, not the school.

For measuring gains, there are fewer choices. Within-school gains could be measured using changes in scaled scores, based on either longitudinal data or comparable cross-sections, although both have methodological difficulties. Benchmarking gains is quite difficult and the only game in town is the VAS which uses adjusted standard deviation growth (effect sizes). ETS considers this the only valid way to benchmark, since it adjusts for both differences in initial SAT/ACT and differences (in SAT/ACT scores) in attrition. It also means that school growth is compared to standard school growth. There are many issues but they are the same for all the tests, not just ETS.

Psychometric quality: Reliability:

ETS has the highest reliability of all the tests we are considering. Consistency across test versions, at the school level, is measured by the correlations of school averages of different test versions given in Table 2b of the TVS report. For ETS, the results are: .93 (critical thinking), .91 (writing), .94 (mathematics), .91 (reading). Consistency across items is measured by the mean of random split-half reliabilities in Table 5 of the TVS Report: .95 (freshman critical thinking), .91 (senior critical thinking), .94 (freshman writing), .88 (senior writing), .95 (freshman math), .93 (senior math), .94 (freshman reading), .88 (senior reading).

Test Development & Logistics

ETS provides a general procedure manual known as the “User’s Guide”, “Proctor Manual” for online administration and a “Supervisor’s Manual” for paper-and-pencil administration.

ETS offers two scoring conventions: Norm-referenced scores (scaled scores) and Criterion-referenced scores (proficiency classifications). Scaled scores compare the scores of one student or group of students to another, or the same student or group of students at different points in time. Proficiency classifications note the level of proficiency obtained on a certain skill set. There are three skill levels for writing, mathematics and reading, with level 3 reading equivalent to the attainment of critical thinking.

Both test forms, Standard and Abbreviated, are statistically equated to offer the same level of detail at the group level. While both test forms provide total scores, scaled subscores and proficiency classifications at the group level, only the Standard test provides subscores and proficiency classifications for individual students. Demographic data is provided in group percentages with the potential for subgroup statistics based on a list of characteristics.

Additional demographic data sets are available for a fee. A customized report to compare performance against a select group of peer institutions is also available upon request.

The optional essay is analyzed using a computer program, *e-rater* and reported as a total score on a six-point scale. The 50 locally answered questions are reported as percentages of students' responses to each question and are not included in the total score or subscores.

Appendix F

Test Evaluation Scoring Sheet – Final Tally

Number of Task Force Members Assigning Scores of "1", "2" or "3"

Task Force on Learning Outcomes Assessment
Test Evaluation Scoring Sheet

Directions: For each section in the "Basis of evaluation" column, provide a score 1-3 to each of the four tests we reviewed, where:

1 = serious lack or deficiency

2= acceptable

3= outstanding or highly desirable feature

For any score of 1 or 3, briefly indicate the deficiency or highly desirable feature.

Specification	Basis of evaluation	CAT			CAAP			CLA			ETS PP		
		1	2	3	1	2	3	1	2	3	1	2	3
Test purpose and design are consistent with the Task Force charge and with CUNY learning objectives	1. Tests are to be used to: • Measure learning gains • Benchmark college performance against that of comparable institutions outside CUNY • Improve teaching and learning throughout CUNY	4	7	-	2	8	-	-	4	7	2	9	-
	2. Core learning outcomes across CUNY: • Reading • Critical thinking • Written communication • Quantitative literacy • Information literacy	5	6	-	7	3	-	-	4	7	9	2	-
Psychometric quality	**Content Validity** 3. Do the test tasks require the test-taker to use the skills and competencies described in the relevant LEAP VALUE rubrics?	2	9	-	8	2	-	-	2	9	10	1	-
	4. Does the scoring of the tasks reflect the progression of rubric skill levels	2	7	-	7	2	-	-	6	5	6	4	-

Specification	Basis of evaluation	CAT			CAAP			CLA			ETS PP		
		1	2	3	1	2	3	1	2	3	1	2	3
	External Criterion Validity 5. Is there evidence that the test detects learning gains at the institutional level?	6	4	-	1	8	-	-	7	4	2	9	-
	Validity generalization 6. Does the test developer clearly set forth how test scores are to be interpreted and used?	4	7	-	2	7	-	1	5	5	2	8	1
	7. Are there other participating colleges in its database of results that are comparable to those of CUNY and can serve in a benchmarking function?	6	2	-	3	5	-	4	5	1	4	5	1
	Reliability (*Score accuracy for institution-level comparisons)* 8. Is there evidence for the stability of scores over different items or forms of the test?	7	3	-	-	8	1	-	10	1	1	5	5
	9. Is the reliability of scores across items or raters acceptable?	6	4	-	1	7	1	-	8	3	1	6	4
	10. Does the test developer provide guidance for controlling the effects of sampling covariates (e.g., ESL status, gender, race) on scores.	4	6	-	-	6	-	-	7	2	2	6	1
Test Development & Logistics	11. Is there a technical manual that describes the test development process, test specifications, scoring rubrics, field testing, and availability of multiple parallel forms?	5	6	-	-	8	2	-	8	3	1	6	4
	12. Is there a test administration manual that describes the testing protocol and any special testing requirements, e.g., online administration, administrator certification, test-taker preparation materials, scoring protocols	2	6	-	-	8	1	-	7	3	1	7	2

Specification	Basis of evaluation	CAT			CAAP			CLA			ETS PP		
		1	2	3	1	2	3	1	2	3	1	2	3
	13. How are test results communicated to the colleges? What guidance is there for score interpretation with respect to benchmarking and learning gains?	5	4	-	1	8	-	1	5	4	2	7	1

Consensus Unacceptable 🟥 Acceptable 🟩 No Consensus ⬜