
# Report of the CUNY Proficiency Examination Task Force

**The City University of New York**

**September 2010**

# Table of Contents

# Executive Summary

**The Charge to the Task Force**
In November 2009, the CPE Task Force was convened and charged by Executive Vice Chancellor Alexandra Logue to address the following questions:

1. *What goals do we wish to accomplish with a standard assessment tool? The Task Force should consider the question broadly, but also address two specific questions:*

    a. *How important is it that CUNY be able to measure the proficiency gains (in addition to the proficiency status) of its students?*
    b. *How important is it for CUNY to be able to benchmark the proficiency status or gains of its students against those of students at other institutions?*

2. *Given these goals, what are the strengths and limitations of the CPE in its current format? The Task Force should consider the question broadly, but also address the following:*

    a. *Is the test a valid and reliable measurement tool?*
    b. *To what extent has the test influenced instruction and learning at our colleges in desirable ways?*
    c. *Should the University modify the CPE instrument and/or significant aspects of CPE policy?*
    d. *How does the value contributed by the CPE compare to the costs of administering it?*

3. *Would any of the commercially available instruments better address the assessment goals articulated in #1? What value might these instruments contribute, compared to the costs of their administration?*

The Task Force divided its work into three parts, which are reflected in the organization of the report. The Task Force first delineated the purposes to which educational assessment can be put. Second, the Task Force examined the CUNY Proficiency Examination (CPE) as a measurement instrument and evaluated how well it has accomplished its intended purpose. The final section of the report is devoted to a review of commercially available examinations and their potential place in CUNY's assessment agenda.

**Goals of standardized assessment tools**
The Task Force identified three potential purposes for the system-wide implementation of standardized assessments at CUNY:

- *Certification* – the original purpose of the CPE in response to the Board of Trustees' resolution to certify the readiness of CUNY students for upper-division work. Certification examinations are primarily assessments of individuals, designed to determine whether standards of knowledge and ability have been met.

- *Program monitoring* –the assessment of the effectiveness of a college and its instructional components in meeting their goals, and the use of this information for improvement of instruction.

- *Public accountability* –an obligation to inform the public about a college's effectiveness in delivering its educational programs. The regional accrediting agencies, traditionally the primary vehicle for accountability, have focused more on measuring process than effectiveness. In recent years, there has been growing interest in the use of standardized measurements of learning.

**CUNY at a Crossroads**
As it contemplates the potential role for University-wide standardized assessments of proficiency, CUNY's leadership must weigh the relative priorities of certification, program monitoring, and public accountability. Three questions loom large:

1) Is there a continuing need for certification testing to insure that CUNY students completing general education are ready for upper division course work?

2) How potentially valuable is University-wide standardized testing for program monitoring at CUNY campuses? Although assessment depends on the development of an array of evidence to measure whether programmatic goals are being met, standardized instruments may be a useful component of this evidence.

3) How does the University wish to hold itself accountable to its publics?

In its deliberations, the Task Force has attempted to explore the implications of these questions for the future of system-wide testing at CUNY.

**Evaluation of the CPE**
*Overview.* Approved by the CUNY Board of Trustees in 1997 and implemented in 2001, the CPE is designed to certify that students who have reached the 45$^{th}$ credit are ready for upper division course work. Because every CUNY student must pass the test in order to graduate, it is a high-stakes examination. The CPE is a three-hour exam consisting of two prompts. The first is an academic writing task requiring a critical response to two

college-level texts, one 8-9 pages in length, read in advance of the test, and the other 1-1 ½ pages read during the test. The objective is to produce an essay that reflects a comprehension of the texts, an ability to synthesize the ideas in the two texts, and a critical analysis, all with reasonable control of language. The second task is intended to tap dimensions of quantitative reasoning. Students are provided with a short text of 150-200 words and two graphs. The task is to identify claims from the text and to state whether or not the graphs support the claims and why.

*Scoring.* Task 1 is scored on a rubric containing four dimensions of performance: 1) Organization, 2) Critical Reading, 3) Development of Ideas, and 4) Language. The Task 2 rubric consists of one dimension that evaluates ability to identify claims in a short text, read and interpret graphs, and relate the data in the graph to the text. Each of the dimensions is scored on a six-level performance scale. Total scores on the CPE range from 12 to 72, with Task 1 worth up to 48 points and Task 2 a maximum of 24. The minimum passing score has been 34 since the inception of the exam. The longitudinal pass rate of the most recent cohort of students tracked over three administrations is 93%.

Unlike the commercial alternatives, which are norm referenced, the CPE is criterion referenced--scored against a performance scale. Norm-referenced tests position the test taker in a population of test takers. Typically this is done by translating the individual's raw score into a percentile placement in the population. In contrast, criterion-referenced tests match the individual's performance against a fixed standard (in the case of the CPE, the rubrics used to score Task 1 and 2). A positive feature of the CPE is its intuitive rubric for Task 1, which has been embraced by many CUNY faculty members as a valid measure of critical reading and writing abilities.

*Reliability.* For the CPE, we can compute inter-rater reliability, a measure of the consistency of two independent readers. Agreement of readers is high, producing reliabilities ranging from .58 for the Language dimension of Task 1, to .74, .73 and .72 for the other three dimensions of Task 1, to .93 for Task 2.

*Validity.* The validity of the CPE is gauged by its ability to measure what it was designed to measure—readiness for upper division work. Task 1 appears to have face validity with faculty, in part because it was developed by a faculty task force and in part because it is a performance test based on an authentic task scored according to criteria that faculty use to grade writing in their classrooms.

Task 2 seems to have less validity. It measures only a few of the abilities that most faculty members would associate with quantitative reasoning, and the task itself is highly artificial. The scoring of Task 2 is prescriptive, and the task requires students to respond in a specific way. It appears that many students who score well on Task 1 but low on Task 2 do so simply because they do not understand the Task 2 prompt. However, once understood, Task 2 is seen as testing a skill that is not difficult. It certainly is not a prompt that deserves to be weighted as heavily as it is. Data show that Task 2 has an undue effect on CPE scaled scores that actually *reduces* CPE validity correlations with post-CPE academic outcomes.

Nevertheless, the CPE, taken as a whole, displays a measure of predictive and construct validity. We find consistent and significant relationships between CPE scores and grades earned in courses completed after taking the CPE. Additionally, CPE scores were found to be consistent with other measures of college-level ability.

*Impact of the CPE.* As a certification exam, the CPE has become fully integrated into the academic and administrative infrastructure of CUNY's undergraduate colleges. The test incentivized a sharper focus on writing skills and writing programs not only in English departments but also across the curriculum. The Task Force believes that the test, particularly Task 1, has the potential to provide valuable information about the ability of CUNY students to undertake academic writing as measured by the Task 1 dimensions—organization, critical reading, development of ideas, and command of language. Because Task 1 raw scores are distributed along the whole 8-48 scale in an approximately bell-shaped curve, the prompt might be used to measure gradations of writing ability both below and well above the minimum passing score. The test provides a performance scale and a standard that both faculty and students can use to monitor progress toward the levels of writing performance they are expected to achieve. Also, by comparing predicted to actual CPE scores, an analyst could focus attention on those colleges or programs that might be sources of promising practices. Although nine years of CPE test scores are available to the colleges, so far, little use has been made of the data for assessment purposes.

*Limitations of the CPE.* Despite the integration of the CPE into the CUNY landscape, there has been no discernible upward trend in writing proficiency as measured by mean scores on Task 1. Moreover, as a certification exam, the test is redundant. Almost every student who meets the prerequisites for taking the exam—basic skill proficiency in reading and writing and completion of 45 credits with a GPA of 2.0 or better—can pass the exam. Consequently, the test appears to add little information regarding readiness for upper division study. Further, for purposes of external accountability, the CPE has serious limitations. Because the test was designed by CUNY faculty and is administered only within CUNY, it is not possible to benchmark the achievements of CUNY students against those of comparable students at other institutions. Nor does the CPE, as currently administered, allow the University to measure improvement in analytic writing and quantitative reasoning. Because it is administered only once, at the 45$^{th}$ credit, the CPE does not measure gains, though if the funds were available it could conceivably be used for this purpose.

*Cost.* The CPE has become a very expensive exam to administer. The total annual cost of the exam is projected to increase from $3.34 million in 2009 to $4.92 million in 2010. (These figures include all costs of the exam—development, scoring, appeals, and administration, both centrally and at the campuses, as well as academic support, including CPE workshops.) In January 2010, the contract for development, production of test materials, and scoring moved from ACT to Pearson, and the cost of these services increased dramatically, from $1.66 million last year to a projected $3.2 million for this year. Because the test is high stakes, security requirements are rigorous, scoring must be conducted by two readers, and scores just below the cut point are automatically re-read so

that potential scoring errors do not prevent students from graduating. Every CUNY student takes the exam, further adding to the cost. Finally, the colleges must offer extensive support to students. The cost of the CPE will increase annually due to enrollment growth and price escalations built into the Pearson contract.

**A Review of Commercial Proficiency Tests**

In addition to the CPE, the Task Force conducted a review of three testing instruments: the Council for Aid to Education's (CAE's) Collegiate Learning Assessment (CLA), ACT's Collegiate Assessment of Academic Proficiency (CAAP), and the Educational Testing Service's Measure of Academic Proficiency and Progress (MAPP), recently renamed the ETS Proficiency Profile. The Task Force assessed the potential use of these instruments for certification, program monitoring, and public accountability.

*CLA.* The CLA is typically used for assessment purposes to measure gains in the cognitive abilities associated with general education, although in the wake of the Spellings Report, it has also become a popular tool for external accountability. It is not well suited for certification because it was designed to assess the performance of institutions or their subunits rather than individuals. The test is designed to measure the value added contribution of colleges and permits benchmarking with other institutions. Performance on the test is normatively scaled, facilitating comparison across colleges.

The CLA is administered entirely on computers linked to the internet, and consists of three constructed response prompts—The Performance Task (90 minutes) and two prompts designed to elicit analytical writing—Make-an-Argument (45 minutes) and Critique-an-Argument (30 minutes). In a typical administration, students take either the Performance Task or the two analytical writing tasks—totaling under two hours, including extra time allowed for the mechanics of test administration. In addition to these components of the CLA, students who have not taken the SAT or ACT are required to sit for the Scholastic Level Exam (SLE), a short-form 50-item measure of cognitive ability requiring just 12 minutes to administer. SAT, ACT and SLE scores are used to adjust CLA scores for initial ability, so that the institution can gain a more precise estimate of its contribution to any learning gains detected. The CAE is developing a shorter version of the CLA, requiring a maximum of 60 minutes, but as of August 2010 the CAE did not know when the shorter version would be ready.

The CLA's Performance Task is a complex performance prompt that requires students to employ critical thinking, analytic reasoning, problem-solving skills and written communication skills to answer a set of open-ended questions about a fictional but realistic scenario. The student is asked to read a narrative describing a situation, together with a set of instructions, and is given online access to a library of information sources such as letters, memos, summaries of research reports, newspaper articles, data displays and other documents. The students are expected to sift through these sources of evidence, weigh their value, and draw upon them in writing their responses to the questions in the prompt.

The CLA also contains two subtests of analytical writing, both of which measure the student's ability to express complicated ideas, weigh assertions and evidence, support ideas, create a coherent argument, and express thoughts in conventional English. The first, "Make-an-Argument," asks students to support or reject a position on an issue. Students are asked to take a clear position on either side of the question, and to support their position effectively. "Critique-an-Argument," presents a flawed argument and requires students to identify the logical flaws.

*CAAP*. Because it was designed for the assessment of individuals, the CAAP is the only test that could be a candidate to replace the CPE as a certification instrument. It could also be considered for program monitoring and public accountability efforts.
A strength of the CAAP is its topic coverage, with six subtests: math, reading, writing skills, writing essay, science, and critical thinking. The subtests can be administered in 40-minute modules. All but the essays consist entirely of multiple choice items, which makes them fast and inexpensive to score, but at the cost of face validity. All scores are reported on a norm-referenced scale. The essays are identical in format to the CUNY/ACT essay that the University has been employing to assess the writing ability of students when they first apply for admission.

*MAPP*. An advantage of the MAPP is that it can be administered and scored quickly. The instrument consists of 4 subtests: critical thinking, reading, writing, and math, each with 27 multiple choice items. The writing subtest multiple choice and focuses on usage and mechanics. The entire test takes 2 hours, and a 40 minute version is available. The intent of the MAPP is to obtain a sampling of ability at the institutional level. Scores can be used to counsel students, but it is not recommended that the results be used for high-stakes purposes. It is, however, suitable for program monitoring and accountability.

**Reliability and Validity**
In general, multiple choice tests have higher reliability correlations than the performance-based tests. The CAAP multiple choice subtests achieve reliabilities ranging from .87 to .92, while the MAPP reliabilities range from .91 to .94. ACT reports an inter-rater reliability of .75 for the CAAP essay test. For the CLA Performance Test, CAE reports an inter-rater reliability of .76 to .87, depending on the specific prompt. The reliabilities for Make-an-Argument are somewhat lower, ranging from .57 to .70 and for Critique-an-Argument, from .77 to .84. For the CPE, inter-rater reliabilities for Task 1 are .58 to .74, below those associated with the CLA's Performance Task, while the reliability of the Task 2 scoring is higher, at .93.

A study of the construct validity of the CAAP, MAPP and CLA suggests that these tests generally measure what they were intended to measure. The math and reading subtests correlate with the subtests in critical thinking, science and math in the expected pattern. However the subtests in writing and critical thinking are not quite so consistently more highly correlated with one another than they are with subtests of the other constructs. If we move to considerations of authenticity, the performance-based prompts of the CLA are regarded as actually measuring the abilities they were designed to measure, and these abilities are seen as important educational outcomes (Shavelson, 2010, pp 57-60).

**Implementation Logistics**
The administration of these instruments faces a common set of challenges. The first and most important is gaining buy in from faculty. If test results are to be taken seriously for improvement of teaching and learning, the faculty must embrace the test as a valid measure of key abilities that a college-educated person should acquire. A second challenge is sampling. If the test is to be a valid indicator of learning gains, the sample of test takers must be representative of the college. A third challenge is motivation to perform well on the test.

**Cost Analysis**
As noted previously, CUNY has been administering the CPE to all undergraduates reaching the 45th credit, at an estimated cost of $4.92 million in 2010. In addition to tallying the current cost of the CPE, the Task Force undertook a comparison of the CPE with the alternative tests, assuming for this exercise a much smaller number of test takers for purposes of assessment rather than certification. The cost of administering all the instruments considered here depends on a number of factors, including the base price of the test; the number of test takers; the cost of scoring the writing sample that is part of the CAAP and an option with the MAPP; the cost of incentives to motivate students to take the test; and investments in IT infrastructure. We conducted a cost comparison of the CPE, CLA, CAAP and MAPP assuming an administration of the test to 200 freshmen and 200 seniors per college. Projected expenditures range from $767,477 for the CPE to $929,783 for the CAAP. If the tests are to be used to assess learning gains for subgroups of students, the sample size for each college will be larger than the 200 freshmen and 200 seniors modeled here, and costs will be higher.

## Recommendations

The recommendations of the Task Force are the product of extensive and spirited discussions that often included diverse points of view. It was difficult to make specific recommendations because the issues are complex and nuanced. Yet the Task Force agreed unanimously about the importance of obtaining information for assessment and improving students' academic literacy, no matter which instruments are used.

**Recommendation 1. After much discussion, the Task Force reached consensus that CUNY should discontinue the use of the CPE as a high-stakes certification exam.**

As currently used, the CPE does not appear to add much additional information about readiness for upper division work once grades and credits have been considered. Virtually every student who can meet CUNY's basic proficiency requirements in reading and writing and can complete 45 credits with a GPA of 2.0 or better (the pre-requisites for taking the CPE) can pass the CPE. The university-wide longitudinal pass rate is now 93%, and the raw CPE score for Task 1 has remained basically flat. When the high cost of administering the CPE is considered as well, it seems impractical to continue

administering the CPE as a certification test.

If the University wishes to keep in place a certification test, only one other test considered in this report is a possible candidate to replace the CPE – the CAAP. It is the only test designed for student-level analysis that includes a performance writing task graded by rubric. The other two tests, the CLA and the MAAP, are not designed to be used for high stakes testing. Before entertaining the CAAP test seriously, however, the University together with its faculty should consider whether the reliance of the CAAP on multiple choice items, the broad domain of the test, and the basic level of the writing sample will meet its needs.

**Recommendation 2. Consideration should be given to retaining Task 1 of the CPE as an instrument for program monitoring. Because of the value of the prompt as a standard assessment tool, CUNY colleges should consider whether it is feasible to somehow embed the exam in the curriculum of some courses, thereby insuring that CUNY students meet or exceed standards of academic literacy.**

The CPE, particularly Task 1, has value as a tool for assessing mastery of the key elements of academic literacy: comprehension of collegiate texts, the ability to synthesize ideas in texts, critical analysis, and control of written language. Because the test is criterion referenced against a rubric, it is relatively easy to interpret the test scores in light of this fixed standard. The test appears to measure a range of abilities and consequently may be helpful in measuring progress toward goals of improving writing beyond the minimum level of proficiency signified by the current cut point on the CPE. Because the rubric is multidimensional, the prompt can also be a helpful diagnostic tool.

The Task Force is aware that the cost of the CPE is extremely high. If the test were no longer high stakes, however, the cost could be reduced substantially in a number of ways. CUNY could recycle its extensive library of test forms. The test could be read once rather than twice, with random quality control measures. (Inevitably, however, the use of a single reader would reduce the rigor of the scoring process.) Appeals would no longer be necessary. The cost of supplementary support and the administrative overhead associated with the exam would be reduced substantially.

To capitalize on its potential value as an assessment tool and to motivate students to continue to sit for the test and perform well on it once it were no longer a high-stakes test, the Task Force recommends that the test be embedded in the curriculum. Colleges and programs could then be held to agreed-upon standards, and test results could be used by colleges to monitor the achievement of CUNY students. Of course, widespread use of the CPE in this way could be hampered by the current fiscal constraints.

Up to now, the CPE has not been employed widely at CUNY as an assessment tool. Assessment offices have only recently begun to take a closer look at the available CPE data and to consider how it might be used to evaluate their college's learning outcomes. If the CPE is retained for assessment purposes, the University should institutionalize the

dissemination of CPE assessment data to institutional research and assessment offices at its colleges.

In addition to facilitating the continuing use of the CPE by colleges wishing to employ it as an assessment tool, the University must maintain its commitment to academic literacy, through such programs and activities as Writing Across the Curriculum (WAC) Writing in the Disciplines (WID), writing intensive courses, and appropriate academic support for students.

**Recommendation 3. If the CPE is to be retained for any purpose, Task 2 should be revisited.**

The validity of the CPE is compromised by the scoring and weighting of Task 2, as noted above. Given the growing recognition of the importance of quantitative reasoning skills across the University, it is clear that Task 2 should be rethought.

**Recommendation 4. There are a number of reasons why a nationally normed standardized test instrument might have a place in the University's assessment agenda. Assessment begins with a definition of the program, the program goals, and a plan of action, and assessment tools should be chosen to fit the purpose and goals of the program. The choice of a test must be made deliberatively and in consultation with faculty and program managers. Further, no one test can fulfill all purposes; effective assessment requires a battery of different types of tools, and those tools will differ for different campuses and disciplines.**

The University should promote and assist with the creation of a culture of evidence and continuous improvement. Of utmost importance is the conversation that takes place among stakeholders as they establish goals and progress indicators and analyze feedback on what is working and what is not, and use that information for future improvement. The emphasis placed on standardized tests in addition to locally developed ones will depend on the nature of the instructional activity and the usefulness of the standardized test results for making adjustments. Faculty and academic administrators can be expected to welcome data from standardized tests if the data are meant to guide program improvement.

**Recommendation 5. The Task Force advises experimentation with publicly benchmarking CUNY colleges if it can be done without compromising the primary function of enhancing students' learning, if the benchmarking methodology is sound, and if the cost is reasonable**

The Task Force recognizes the importance of public accountability, but urges caution if

the University decides to adopt an instrument for this purpose. CUNY must take care to educate members of the public about the distinction between level of performance and the "value added" by institutions serving less well prepared students. If the University adopts a new accountability test, it should consult with faculty, both to select the best instrument and to plan how to use the results not only for accountability but also for the improvement of teaching and learning.

# Introduction

In the fall of 1997, the City University of New York Board of Trustees passed a resolution that called for the development of an examination to certify that all CUNY undergraduates completing their first two years of study are ready to undertake upper-division study. This action by the Board ultimately led to the creation of the CUNY Proficiency Examination, or CPE, and its implementation beginning in the fall of 2001.

In view of its use over the past 10 years as a de facto degree requirement for the Associate and the Bachelor's degrees, and the considerable resources that have been expended in the development of test forms, their administration and scoring, Executive Vice Chancellor Alexandra Logue convened a Task Force in November 2009 and charged it with conducting a systematic review of the CPE. The Task Force was broadly constituted, including representation from the community and senior colleges, CUNY's Assessment Council, and the CPE Advisory Committee, as well as two provosts, two members of the University Faculty Senate, faculty members who are current and former CPE Liaisons, two members of the Office of Academic Affairs—the interim Dean for Undergraduate Studies and the Director of Assessment, Raymond Moy, who served as Chair. A list of Task Force members is available in Appendix A. Executive Vice Chancellor Logue requested that the Task Force address three questions:

1. *What goals do we wish to accomplish with a standard assessment tool? The Task Force should consider the question broadly, but also address two specific questions:*
    a. *How important is it that CUNY be able to measure the proficiency gains (in addition to the proficiency status) of its students?*
    b. *How important is it for CUNY to be able to benchmark the proficiency status or gains of its students against those of students at other institutions?*

2. *Given these goals, what are the strengths and limitations of the CPE in its current format? The Task Force should consider the question broadly, but also address the following:*
    a. *Is the test a valid and reliable measurement tool?*
    b. *To what extent has the test influenced instruction and learning at our colleges in desirable ways?*
    c. *Should the University modify the CPE instrument and/or significant aspects of CPE policy?*
    d. *How does the value contributed by the CPE compare to the costs of administering it?*

3. *Would any of the commercially available instruments better address the assessment goals articulated in #1? What value might these instruments contribute, compared to the costs of their administration?*

The first question extends the scope of the charge well beyond a review of the CPE by asking the panel to reconsider the fundamental purpose of standard assessments at CUNY. University-wide tests can potentially be used for several different, and not necessarily mutually exclusive, purposes-- to insure that all students have mastered a body of knowledge or skills, to assess learning, and to reassure the public that the University is meeting its responsibility to educate its students. In addition to a review of the CPE, the Task Force was charged with weighing these basic purposes and considering how the CPE and other instruments might accomplish them.

The task of choosing an appropriate test must begin with a definition of the testing purpose. This contextualized approach to test evaluation is consistent with current measurement best practice. In the latest edition of the American Council on Education's *Educational Measurement*, R. L. Brennan (Ed.), Westport, CT: Praeger, 2006, C.B. Schmeiser and C.J. Welch succinctly describe how contextualized use of test results is the ultimate basis for evaluating and choosing among tests:

> *Perhaps if test developers have learned anything at all in the last fifty years, it is that the practice of measurement in education is complex and almost wholly dependent on context. From test design through test score interpretation and application, developers must continuously be cognizant of the context(s) within which test results are used. Each context is unique, and tests must be based on a strong foundation of empirical validity evidence that addresses these varying contexts of use as effectively as possible.*

At the core of the Task Force's evaluation strategy are three closely-related best practices: 1) testing must have a purpose--tests should not be given for testing's sake; 2) test instruments should be designed and chosen with that specific purpose in mind; and 3) the quality of tests, their reliability and validity, should be evaluated in terms of how the test results are to be used. In its deliberations, the Task Force has taken into account all three criteria.

We begin with a delineation of the purposes to which educational assessment can be put, and their alignment with CUNY's objectives. It is this match of purpose and objectives that provides the context for comparing the usefulness of results from different tests. We address the second part of the charge by examining the CPE as a measurement instrument and evaluating how well it has accomplished its intended purpose. The final section of the report is devoted to a review of commercially available examinations and their potential place in CUNY's assessment agenda. The comparison is done with respect to purpose, test design, the logistics of test administration and scoring, and the use of test results.

## Purpose

The first question of the charge is the most difficult: *What goals do we wish to accomplish with a standard assessment tool?* In its discussions, the Task Force identified three broad alternatives. The first is certification that CUNY students have met one or more specified educational objectives. The second is program monitoring, which refers to the formative and summative assessment of teaching and learning for purposes of improvement. A third purpose is accountability, both to external publics and to stakeholders within postsecondary institutions. If we wish to hold ourselves accountable, we may choose to publish evidence that our students have mastered essential knowledge and skills and evidence that our students have gained this mastery because of their exposure to CUNY's educational programs—general education and the major.

### Certification

Certification examinations are primarily assessments of individuals, designed to determine whether standards of knowledge and ability have been met at a particular point in time, when the test is administered. Although certification tests are often given at the end of a program, they can also be used at the beginning as qualifying examinations. To design a certification exam, standards of performance are established, and an assessment protocol is developed to sample and score performance. When a certification test is implemented, the persons being certified must understand the performance expectations, the impact of the certification, and they must be given the opportunity to prepare. Examples are licensure examinations, the GED, and AP tests. The results of certification tests can be used as an indicator of institutional quality. However, their main purpose is to insure that students who have completed a curriculum have met its instructional goals. The onus of achievement is primarily on the student.

A key element of the development process is standard setting--establishing an appropriate level of performance and choosing a corresponding cut-score for use in making the certification decision. If set too low, unqualified candidates may be certified (with potentially negative consequences), while if set too high, the exam will unnecessarily serve a gate keeping purpose, denying qualified candidates a chance to participate. In setting a cut point, the test consumer must weigh the consequences of the two types of error, gather data on the performance of test takers, and if necessary adjust the standard.

### Program Monitoring

By "program monitoring" the Task Force means the assessment of the extent to which a college and its instructional components are meeting their goals, and the use of this information for improvement. When the testing purpose shifts from certification to program monitoring, the focus moves more toward the college and its instructional components, though learning is of course an enterprise that is shared between the

institution and the student. Programs can range in scope from an entire undergraduate curriculum, to its primary components-- general education and the major--to individual courses or workshops or other units of instructional intervention. Instructional programs typically have as their main objective a set of learning goals, which may vary in their complexity from a single skill or narrowly defined body of knowledge to a multi-dimensional and richly layered set of learning goals. Programs may also be designed to achieve goals in addition to learning, such as persistence, student engagement, or civic engagement for example. In this context, the purpose of a test is to measure the performance of students who have been exposed to a program in a way that can attribute gains to that program. The test can provide information that program managers may use to identify strengths and weaknesses, and to serve as a tool for internal accountability to monitor progress toward agreed upon goals.

Assessment is a process in which stakeholders come to agreement on the goals of the program, develop an array of appropriate measures, possibly but not necessarily including standardized measurements, gather and interpret the assessment data, and use the results for program improvement. An assessment program must be carefully aligned with the specific learning goals of the program being assessed. It should include multiple measures, including both direct and indirect evidence. It might employ quantitative and/or qualitative methods in a variety of formats. It might use locally designed instruments or standardized ones. The introduction of any standardized test to assess educational programs across a university system such as CUNY must be done with care and consultation. Because instructional programs vary in their learning objectives, a common instrument must tap core abilities that have broad currency.

### Public Accountability

Public accountability is an obligation to inform the public, including prospective students, about a college's effectiveness. Accountable institutions are committed to improvement and at the same time can persuade the public that they are effective through the use of compelling indicators of quality. Traditionally the regional accrediting agencies have been the primary vehicle for this assurance, but in recent years, their pre-eminence has been challenged. In its final report, the Spellings Commission on the Future of Higher Education (2006) was highly critical of the accrediting agencies and called for much greater transparency and accountability, making use of standardized measurements of value added. With the release of the report, the need to incorporate measures of learning gains into the public accountability agenda gained significant momentum (Ewell, 2009). The report of the Commission made the case eloquently (U.S. Department of Education, 2006):

> *There are …disturbing signs that many students who do earn degrees have not actually mastered the reading, writing, and thinking skills we expect of college graduates. Over the past decade, literacy among college graduates has actually declined. Unacceptable numbers of college graduates enter the*

> *workforce without the skills employers say they need in an economy in which, as the truism holds correctly, knowledge matters more than ever.*
>
> *Compounding all of these difficulties is a lack of clear, reliable information about the cost and quality of postsecondary institutions, along with a remarkable absence of accountability mechanisms to ensure that colleges succeed in educating students. The result is that students, parents, and policymakers are often left scratching their heads over the answers to basic questions, from the true cost of private colleges (where most students don't pay the official sticker price) to which institutions do a better job than others not only of graduating students but of teaching them what they need to learn.*

Of course a single testing instrument cannot replace the nuanced assessment process based on a wide range of measures and evidence that accrediting agencies typically require. Such a process is necessary to insure institutional quality. However, by adding a standardized instrument to their quiver of assessment tools, institutions of higher education can more easily compare their performances.

Standardized tests for purposes of public accountability may measure achievement or learning. An institution may wish to signal that its students have achieved a set standard of performance and it may simultaneously want to provide evidence that its students have acquired specified knowledge and abilities in the course of their engagement with that institution (value added).

Inevitably there is a tension between the goals of assessment and the goals of accountability, but they are not necessarily incompatible. Accountability instruments typically are devised by individuals or organizations external to the institution and its programs. In order to generate benchmark data, these instruments must measure broadly defined abilities (e.g. critical thinking, analytic reasoning, communication skills) in a standard way, or an agreed upon body of knowledge (e.g. American history, nursing skills, arithmetic), again with a common metric. Instruments that have been designed to measure these constructs will be useful to institutions and their programs for assessment purposes to the extent that they measure abilities and knowledge that are part of the learning goals of those institutions and programs. Because these goals are designed by faculty, they may not be consistent with the goals as externally defined and embodied in instruments used for accountability. That said, there may well be significant overlap, making standardized instruments a valuable tool for purposes of assessment, though one of many.

## CUNY at a Crossroads

For much of its history, CUNY has employed system-wide standardized testing primarily for purposes of certification, though with an eye to public accountability as well. When admissions criteria were relaxed as part of Open Admissions, the University introduced open admission tests to assess the basic skills of incoming students in reading, writing and mathematics. Students who could not reach the standards established at the time were placed into remedial course work. Faculty determined when remedial students were ready for credit-bearing course work. Over time, these placement tests acquired a certification function, as the University policy required passage of the same tests before a student could be admitted to the upper division of a bachelors program. The CPE continued this tradition of certification testing. In its 1997 resolution, the CUNY Board of Trustees required passage of the CPE as a degree requirement for associate degree programs, and as a de facto requirement for the baccalaureate degree, because passage of the test is a requirement for entry to the upper division. The CPE replaced the skills assessment tests in reading and writing for this purpose.

The CPE was born out of concern by the Board that CUNY students had been graduating without sufficient command of important academic skills, including the ability to read and interpret college-level texts, evaluate them, and to write clearly and effectively. The test was meant to reinforce academic standards, which were widely perceived, both inside and outside the University, to have slipped, and the test arguably achieved this goal. After the CPE was implemented in 2001, general education curricula across the University were revised to place more emphasis on the kind of writing tested in the CPE. The test was also meant to reassure the public that CUNY graduates had met standards of academic literacy, and quantitative reasoning. This emphasis on certification has meant that CUNY's testing program has rested on high-stakes testing of all students, and the use of common instruments and uniform standards across the system.

In charting the future of standardized testing instruments at CUNY, its leadership must address three fundamental questions:

> 1) Is there a continuing need for certification testing to insure that CUNY students completing general education are ready for upper division work and that CUNY graduates possess the general abilities now tested by the CPE?
>
> 2) How potentially valuable is University-wide standardized testing for program monitoring at CUNY campuses? Although assessment depends on the development of an array of evidence to measure whether programmatic goals are being met, standardized instruments may be a useful component of this evidence.
>
> 3) How does the University wish to hold itself accountable to its publics? In the wake of the release in 1999 of *CUNY Adrift*, *The Report of the Mayor's Task Force on CUNY*, the University employed the CPE as one signal of the increased rigor of its curriculum. CUNY has done so by publicizing the

implementation of the CPE. Perhaps the key vehicle for accountability, however, has been the accreditation process and the assessment that it has entailed. Over the past decade most CUNY colleges have substantially increased their investment in assessment to meet the growing demands of professional accrediting bodies as well as the Middle States Commission on Higher Education. The question for CUNY is whether this approach to accountability is adequate in the post-Spellings era, in which demands have grown for the use of common standardized instruments to measure and benchmark learning gains.

In its deliberations, the Task Force has attempted to explore the implications of these questions for a University-wide testing program. The Task Force has reviewed the CPE and three other higher education proficiency assessments [namely, the Council for Aid to Education's Collegiate Learning Assessment (CLA), ACT's Collegiate Assessment of Academic Proficiency (CAAP), and the Educational Testing Service's Measure of Academic Proficiency and Progress (MAPP), recently renamed the ETS Proficiency Profile]. Best practice in measurement requires that a test be evaluated against each purpose separately (American Education Research Association, American Psychological Association, National Council on Measurement in Education, 1999). The Task Force has attempted to assess the suitability of these instruments for each of the three purposes.

## Evaluation of the CPE

### Design

The CPE was originally designed as a mandatory certification that CUNY students are prepared for upper division work. For community college students intending to enter the labor force immediately after graduation, an additional goal of the CPE was to signal that the student had acquired abilities deemed important to employers. After the Board passed the resolution that mandated this "rising junior" examination, several years of development were required. The University made a false start, designing and piloting an instrument that proved to be impractical to administer and score. Next, the University formed a CPE Task Force comprised mostly of CUNY faculty (chaired by Dr. Bonne August, then Chair of the English Department at Kingsborough Community College), which formulated the specifications for Task 1 of the current CPE. Task 2 was added later on the advice of consultants as a means of improving the validity and reliability of the test.

The result of this development effort is a two-part assessment that asks students to complete tasks that, for all intents and purposes, could be asked of students completing a general education program. The first part is an academic writing task that requires a critical response to two college-level texts, one 8-9 pages in length, read in advance of

the test, and the other 1-1 ½ pages read during the test. The objective is to produce an essay that reflects a comprehension of the texts, an ability to synthesize the ideas in the two texts, and a critical analysis, all with reasonable control of language. A sample of Task 1 appears in Appendix C.

The second task was designed to tap some dimensions of quantitative reasoning. Students are provided with a short text of 300-350 words and two graphs, all on the same topic, but considered to be independent of one another. The task is to identify claims from the text and to relate the claims to the data presented in the graphs. The task assesses the ability of students to read and understand text containing quantitative information, identify assertions in the text, read and interpret graphs, and integrate that information with textual statements. The prompt does not require an essay, but rather simple summaries of each claim and statements about the degree to which the data in the appropriate graph supports or does not support the claim. A sample of Task 2 appears in Appendix D.

Both Tasks are scored by humans, with the use of rubrics that describe performance levels in four dimensions, or skill areas, for Task 1 and one dimension for Task 2. The Task 1 rubric, which appears in Appendix E, consists of the following dimensions, each measured on a six-point scale: 1) Organization, 2) Critical Reading, 3) Development of Ideas, and 4) Language. These dimensions are conceptualized as the basis on which faculty typically assign grades to analytic writing assignments. The rubric references the ability to 1) organize ideas into focused points for the reading audience, 2) show evidence that one grasps the main ideas and argument of the reading texts, 3) develop ideas adequately and effectively, and 4) demonstrate a command and flexibility in the use of the language.

The Task 2 rubric consists of one dimension measured on a six-point scale designed to measure the ability of students to identify claims in a short text and determine whether or not those claims are consistent with data in the graphs. Students receive points incrementally based on the number of claims they are able to identify and match correctly with the data in the graphs. The Task 2 rubric appears in Appendix F.

To create the scale, each of the two readers' scores (1-6) on each of the four dimensions of Task 1 is totaled. Across the four dimensions, the range of the total score for each reader is 4 to 24, and 8 to 48 for the sum of the two readers' scores. Similarly for Task 2 each reader scores on a 1 to 6 scale, creating a range for the two readers of 2 to 12. Because the Task 2 scores are double weighted, the total scale score ranges from 12 to 72. Finally, to adjust for variation in the difficulty levels of the forms within an administration and across administrations, the scale score totals are post-equated. The minimum passing score is 34 on this 12-72 scale. This cut point was set by a faculty panel at the inception of the exam, and has remained constant up to the present.

## Test Development and Scoring

CUNY has contracted with vendors (currently Pearson, and formerly ACT 2003-2009) to develop new test forms, field test them, produce test materials, and score the tests in accordance with CUNY specifications. Faculty panels review and approve the selection of passages and graphs to be used in the test prompts, as well as all exemplar anchor papers to be used in the scoring process. Through this process, the University can monitor the vendor's adherence to the test specifications and scoring rubric while insuring that the test development and scoring are conducted impartially and objectively.

To insure consistency in the scoring process, the vendors have implemented stringent quality control procedures. Only trained and certified readers may score the test. Each paper is scored independently by two readers, with a third expert reader brought in to resolve discrepancies of more than 1 performance level on any dimension. Among the quality control measures are random re-reads by the third reader, and the use of calibration papers (pre-scored examinations that are placed into each reader's workflow to insure their continuing adherence to scoring benchmarks).

## Score Performance

Table 1 (Appendix B) reports descriptive statistics for the dimension raw scores, the Task 1 and Task 2 total raw scores, the Task 1 + Task 2 total raw score, and a rescaled CPE total score. The dimension scores for Task 1 (Organization (O), Critical Reading (CR), Development of Ideas (DI), and Language (L)) have a range of 2-12, which corresponds to the sum of the two reader scores, each of which has a range of 1-6. The first three dimensions, pertaining to critical analysis, have means of 5.67-5.74 which, divided by 2, corresponds to a performance level of about 2.8. This score characterizes critical writing with non-minor flaws. In contrast, the language dimension shows a performance level mean of 3.85. Writing scored at this level may have lapses, but shows basic control. The dimension scores are added together to produce a raw Task 1 total score of 24.77.

Task 2 consists of a single dimension with an average of 7.49, or a performance level of about 3.75, corresponding to a basic ability to do the task, but not completely in accordance with the instructions.

## Score Scaling

Unlike the commercial alternatives, which are norm referenced, the CPE is criterion referenced--scored against a performance scale. Norm-referenced tests position the test taker in a population of test takers. Typically this is done by translating the individual's raw score into a percentile placement in the population. By contrast, criterion referenced tests match the individual's performance against defined criteria. Rather than placing the test taker in a distribution of scores, criterion referencing measures performance against a constant standard. The performance of students on the CPE is measured against the

criteria embedded in the Task 1 and 2 rubrics and a minimum passing score of 34.

It should be noted, however, that the CPE is a multidimensional test, measuring four dimensions in Task 1 and an additional dimension in Task 2. The scores for each dimension are combined and weighted to create a single score. When interpreting total scale scores for the CPE, one must keep in mind that the Task 1 subscores are weighted equally to one another, but the Task 2 scores are double weighted, and therefore influence the student's total score twice as heavily as any single Task 1 dimension. To interpret the CPE total scores, one should examine the relationship of each of the component dimensions to the total. There are two CPE total scores, a raw total, which is the sum of the dimension scores across tasks (O+CR+DI+L+Task 2), and the CPE total scaled score, which double weights Task 2 (O+CR+DI+L+2(Task 2)) and is equated to other forms within and across administrations.

Comparing the correlations of each of the dimensions with the two CPE total scores (See Table 2), one can clearly see that Task 2 performance is favored in the scaled total (with a correlation of .837 versus the other dimensions which range from .426 to .651). On the other hand, the Organization dimension has the highest correlation with the CPE Total Raw score with a correlation of .840. Here, the Task 2 correlation drops to .671, and the correlations of the other dimensions range from .558 to .831.

## Pass Rates

As a preliminary to this discussion, it may be helpful to summarize the key elements of the University's policy regulating eligibility for the exam and restricting the number of times a student can sit for the test. All students who have completed their 45$^{th}$ credit are required to take the CPE. In order to sit for the exam, a student must have earned a cumulative GPA of 2.0 or better and have demonstrated basic proficiency in reading and writing. Board policy states that students must pass the test by the time they have reached the 60$^{th}$ credit. However, under the policy as actually implemented, students may attempt the test three times. After the third unsuccessful attempt, students can no longer continue at CUNY on a matriculated basis, though an appeals process is available whereby students may obtain permission to attempt the test a fourth time. No student may continue in a degree program as a matriculated student after the fourth attempt. As a consequence of this approach to implementing Board policy, some students progress beyond the 60$^{th}$ credit without having passed the CPE.

Because students may attempt the exam more than once, it is necessary to compute pass rates using longitudinal methodology, measuring success over several administrations of the test. Table 3 reports a longitudinal pass rate over three potential attempts for cohorts of students who were first required to take the CPE in October and who sat for the test in the fall administration or in the subsequent January or March administrations. University-wide, this pass rate has ranged over the past several years from 91.1% to 92.7%. Most recently, the fall 2009 cohort of test takers posted a pass rate of 92.5%. But because this number is calculated only for those invitees who actually test, it is also important to know

how many students were required to take the test in the first place, including those who complied and those who did not sit for the test. For the fall 2009 cohort, the University-wide show rate was about 83%, capping an upward trend over the past several years (Table 4). If calculated against the original cohort of students who were required to test, the pass rate falls to about 77%.

The pass rates reported in Table 3 under-estimate the ultimate pass rate. This is the case because CUNY students do not progress regularly through college, often interrupting their studies with periods out of school. For this reason, and because some students attempt the test several times, a three-semester tracking period is not sufficient. It does appear, however, that the ultimate pass rate stabilizes after about three years. Table 5 reports the three-year pass rate for all students who were required to take the CPE for the first time in the fall 2005 administration. The aggregate three-year pass rate is shown in the lower right hand cell of the table—94%. The remaining 6% consists of students in several categories, including some who were still enrolled after three years but had not yet passed the test despite multiple attempts and others who had left the University before passing.

This analysis provides valuable additional information about the eventual pass rates of students who do not succeed initially. The first row of the table shows the outcomes for the total group of students who were first required to test in fall 2005. About one-quarter of the group did not show up to take the test at that administration, another 4% were allowed to defer the exam, 11% took the test and failed, and the remaining 62% passed on their first attempt. The ultimate pass rate for those who initially fail is quite high—77%, a testament to the support services that CUNY colleges provide for these students and their own persistence. For those who were deferred initially, the success rate is 82% and for those who did not sit when they were first required to and did not receive a deferral, the rate is quite high as well—89%.

Perhaps the most important point in this discussion is that most students pass the CPE, bolstering the general impression on the campuses, according to some members of the Task Force, that the CPE is not a very difficult test to pass. However, this finding must be interpreted in light of the fact that students must successfully complete 45 credits in good academic standing in order to be eligible to sit for the exam. The high pass rate on the CPE confirms what the University has already determined on the basis of the students' academic progress (45 credits) and standing (GPA $\geq$ 2.0). Almost every student who can meet these standards can pass the test, certifying the student as ready for graduation from an associate program and ready for upper division work. The CPE does not add much new information about readiness. Of course the high pass rate is also partly an artifact of the performance level at which the cut score has been set.

## Reliability

The reliability of a test refers to the consistency of its scores. Variability among raters,

forms and test items generates measurement error and must be taken into consideration in a test used for high-stakes decisions. Reliabilities are generally measured on a scale of 0 to 1.0, where 0 indicates a perfect lack of consistency and 1.0 indicates perfect agreement.

For all practical purposes the reliability of a long performance-based writing test such as the CPE can be measured in just a few ways. We can assess the consistency with which it is scored—its *inter-rater reliability*— and its consistency across forms of the same test—its *equivalent forms reliability.* *Internal consistency* reliability is the extent to which an individual would be expected to obtain the same score from one item to the next on the same test. Because the CPE consists of only one prompt to assess academic writing (Task 1) and one prompt to measure quantitative reasoning, it is not possible to compute this measure of reliability, which is applicable when several test items measure the same concept. In addition, it is not practical to determine the *test-retest reliability* of the CPE because of its length. This is the consistency of scores across administrations that are given within a short interval, with no intervention between sittings. To measure test-retest reliability we would have to recruit students who are willing to sit for the same 3-hour CPE exam twice within a short period of time.

*Inter-rater Reliability.* The grading of the CPE has attained a relatively high level of inter-rater reliability because of the rigor of the scoring process. All readers must be trained and certified on a qualification test. Once the readers are certified, their grading is monitored for discrepancies of two points or more. All such papers are read by a third reader, who resolves the discrepancy. Tables 6, 7, 8 and 9 report the degree of inter-rater agreement for each of the four dimensions used to score Task 1: Organization, Critical Reading, Development of Ideas, and Language. For the four years of data collected from fall 2005 to summer 2009, Rater 1 scores are cross tabulated with Rater 2 scores. By convention, readers are said to agree on a score if they have assigned exactly the same score (perfect agreement) or scores that differ by only 1 point (adjacent agreement). The shaded cells in each table contain discrepant scores of a difference of two rubric levels or more. From these analyses, we can see that differences of more than two levels are rare in the scoring of all four Task 1 dimensions. When we tabulated the results for the 164,460 papers graded from fall 2005 to summer 2009, we found that the discrepancy rate was only 2.2% for Organization and Critical Reading, 2.0% for Development of Ideas and .4% for Language.

This high level of agreement is due in part to the narrow range of scores assigned by the readers. Most readers assigned scores between 2 and 5 on each of the Task 1 dimensions, with comparatively few (<2%) papers scored as 1's or 6's. The language dimension had an even more restricted range, with most papers in the 3-5 range. Despite the restricted ranges, the Pearson r correlations were .74, .73 and .72 for the first three dimensions and .58 for the language dimension. These correlations are slightly lower than those associated with other human-scored standardized tests of writing (see Table A for a comparison).

The scoring patterns for Task 2 are quite different. As shown in Table 10, the percentage

of readers who assign discrepant scores is much higher than that for each of the Task 1 dimensions: 8.5% of the scores are discrepant. However, because readers who are not discrepant assign identical scores at a high rate, the reliability coefficient is quite high, at .93.

Table 10 reveals some serious problems with the Task 2 prompt. A review of the marginal distributions for Task 2 shows a high proportion of students receiving extreme scores of 1 or 6—about 45%. The high percentage of 1s indicates that a large number of test takers do not know how to respond to the Task 2 prompt. As a result, students whose true quantitative reasoning ability is higher than a 1 are being assigned that score because they have not understood the directions for the prompt. At the same time, the larger than expected number of 6s suggests that Task 2 is relatively easy compared to Task 1, and the scoring process seems to unduly reward individuals whose true ability should be assigned a score from the middle part of the scale. In short, although inter-rater reliability on Task 2 is quite high because of the high rate of perfect agreement among readers, Task 2 test scores appear to be influenced heavily by the test taker's ability to understand the directions for the prompt.

*Equivalent forms reliability.* To maximize the equivalence of test forms, the test developers chose prompts that were comparable in their level of difficulty and field tested each form on a population similar to CUNY students. These procedures seem to have produced a high degree of consistency across forms. When we examined the effect of form on the total CPE raw score by estimating variance due to form, we found that the form effect is small, explaining just over 3% of the variation in scores. Simply put, the particular form a student receives has very little impact on the total score that student receives. The small amount of variance due to multiple forms is further reduced by the post equating process.

## Validity

The CPE, particularly Task 1, appears to have a high degree of face validity as a measure of readiness for upper division course work, in part because the test was developed by a faculty task force and in part because it is a performance test based on a task deemed authentic. The test was never intended, however, as a comprehensive assessment of general education at CUNY. Task 1 consists of a prompt that might be assigned in an upper division class and is non-trivial. The rubric corresponds well with the criteria that faculty actually use to grade analytical writing in the classroom. Task 2 seems to have less face validity. It measures a few abilities that most faculty members would associate with quantitative reasoning—ability to identify assertions in text, ability to draw information from graphs, and the ability to assess the consistency of the two. But most observers would probably agree that these abilities do not adequately represent the domain of QR abilities. If the University wishes to improve the validity of this part of the CPE, it should convene a faculty panel and charge it with defining the domain and developing new test specifications.

To the general impression that the test nevertheless does have some face validity, we can

add statistical evidence that total CPE test scores correlate with other variables in an expected pattern (construct validity). If, for example, the test is in fact an indicator of readiness for upper division coursework, one would expect to find a correlation between CPE test scores on the one hand and grades earned after the test was taken on the other. To demonstrate construct validity it is not necessary to predict grades with high accuracy, but rather to verify the presence of a significant positive relationship. As our measure of post-CPE grades, we chose GPA one year after taking the CPE. Presumably, the higher the CPE score the higher the GPA.

Table 11[1] shows for students who took the test between fall 2005 and summer 2009 the relationship between success on the CPE and GPA earned in course work completed within one year after the CPE was taken. GPAs were categorized as below 2.0, between 2.0 and 2.99, 3.0 and 3.49, and 3.5-4.0. Reading the column percents, we see that of those who failed the CPE, 10.2% subsequently earned a GPA lower than 2.0, compared to just 5.3% of those who passed the CPE. Of those who failed the CPE, just 5.5% later compiled a GPA between 3.5 and 4.0, compared to 21.5% of those who passed the CPE. The data in this analysis clearly indicate that success on the CPE is correlated with later success in upper division course work.

Table 12 shows the relationship between CPE test results and subsequent grades even more clearly by disaggregating CPE test scores into categories of Low, Border Low, Border High, and High. Students who are grouped in the Low category received scores in the bottom two levels of each component dimension (total raw score of 20 or less), while those who have been classified as High scored in the top three levels of the rubric (total raw score of 40 or more). Border Low is 21-29, and Border High is 30-39. As we read the column percentages from left to right across the columns denoting performance levels on the CPE, we see that the likelihood of earning low grades falls as performance on the CPE rises (e.g., 13.7% of Low CPE performers have a GPA of less than 2.0, compared to 2.3% of High CPE performers). Conversely, the probability of earning high grades in courses taken post-CPE rises with each increment of improvement in CPE test results (e.g., 3.5% of Low CPE performers have GPA higher than 3.5, compared to 41.6% of High CPE performers). This analysis further supports the construct of the CPE as a measure of readiness for upper division work.

The final evidence of construct validity is shown in a correlation matrix containing four indicators of college readiness: SAT Critical Reading and Math, the New York State Regents English examination and the NYS Math A Regents examination, as well as two post-CPE variables, GPA earned in course work completed during the year after taking the CPE, and GPA at graduation (See Table 13). Also included in the matrix are CPE subscores for Task 1, Task 2, the total raw score, and the total scaled score. As one would expect, Task 1, a measure of the ability to write analytically, is correlated more highly with the Regents English and SAT Critical Reading than with the Regents math and SAT math, while Task 2, a measure of quantitative reasoning, is correlated more strongly with the SAT math than with the critical reading. (However, Task 2 correlates equally well with both Regents exams.) Finally, all CPE subscores and total scores (raw

[1] Chi Squares performed on the cross-tabulations in Tables 11, 12 and 13 are significant at $p < .001$.

and scaled) are significantly correlated with both post-CPE GPA indicators, as one would expect if the CPE were a valid indicator of readiness for upper division work. However, the magnitude of these correlations is not impressive. For example, both SAT and Regents test scores are stronger predictors of cumulative GPA at graduation than is the CPE raw score.

The matrix contains some additional data bolstering the case against Task 2. Compared to the total CPE scaled score, the total raw score is more highly correlated with the readiness and GPA indicators. Furthermore, Task 1 subscores are more strongly related to both GPA indicators, SAT and Regents scores than is the Task 2 score. Both of these patterns are evidence that the double weighting of Task 2 in the computation of the total scale score is compromising the validity of the test.

## Impact of the CPE

The CPE has been implemented primarily for purposes of certification, but to some extent as well for program monitoring and accountability. We address each of these three uses below.

*Certification.* Because the CPE was introduced as a degree requirement, it has inevitably had a substantial impact on the University and its students. The test incentivized a sharper focus on writing skills and writing programs not only in English departments but also across the curriculum. Soon after the CPE was first implemented, CUNY put in place the Writing Across the Curriculum (WAC) program, which encouraged the integration of writing instruction into all parts of the curriculum—not just freshman composition. Moreover, all but three CUNY colleges now require for graduation one or more writing intensive courses, which incorporate a significant writing component into their syllabus. Because the CPE is a degree requirement, it has been fully integrated into the administrative machinery of the colleges and University, including its administrative software. At each CUNY campus, systems have been created to notify students who must take the test, schedule them, and to identify students who are struggling with the test and to refer them to appropriate interventions—either courses or workshops. At each campus a CPE Liaison, who is funded by the central Office of Academic Affairs, handles appeals and requests for deferrals. Over time, the test has become widely accepted by faculty, students and administrators as a feature of the academic landscape and as a valuable tool for the improvement of analytic writing. In their deliberations, Task Force members noted that preparation for the CPE has provided an incentive for faculty to do more class work and make more assignments involving analytic writing. To integrate the CPE into the academic and administrative life of 17 colleges has been no small achievement.

This is not to say that the test is without its critics. Many high-achieving students resent having to sit for a test in order to demonstrate abilities that they feel they have already demonstrated amply in the classroom. This complaint may stem partly from the fact that the minimum passing score of 34 is uniform throughout the CUNY system, and is set low

enough that almost all students can reach it, though many students require more than one attempt. Hence, we have a standard that may be challenging for community college students, and easy to meet for students at the more selective senior colleges. Some writing instructors have complained that the resources of their writing labs and their instructional activities have become narrowly focused on CPE preparation. And as we will see, the campuses must spend large sums to administer the exam and to provide student support. Finally, we could find no evidence that the quality of writing by CUNY students as measured by the CPE has actually improved in recent years. When we reviewed student performance on Task 1 over the past 5 years, we found that the mean total raw score has fluctuated between 26.50 and 26.42 on a scale ranging from 8 to 48. There has been no discernible upward trend.

*Program monitoring.* Because the CPE has been administered primarily as a certification exam, all eyes have been focused on the pass rate, the ability of students to meet the University standard. This tendency is most evident in the University's Performance Management Process (PMP), in which the progress of the colleges toward meeting University-wide goals is tracked in an annual review. Because CPE show rates (the percentage of students who are required to sit for the exam who actually do so) and pass rates are PMP metrics, the PMP has created an incentive for colleges to review their programs, especially when the indicators are below the average for their college sectors. There is also anecdotal evidence that some colleges have instituted campaigns to encourage higher show rates and have directed more resources to writing courses and support activities in response to lower than average pass rates. A review of PMP data shows that over time show rates have been improving substantially, while pass rates have edged upward at the community colleges.

Despite this focus on pass rates, the test has the potential to provide valuable information about the ability of CUNY students to undertake academic writing as measured by the Task 1 dimensions—organization, critical reading, development of ideas, and command of language. Figure 1 (Appendix B, p. 61) shows that Task 1 raw scores are distributed along the 8-48 scale in roughly a bell-shaped curve, albeit with spikes at a few score points. Because scores are distributed across the whole scale, it can be used to measure gradations of writing ability both below and well above the minimum passing score. The test provides a performance scale and a standard that both faculty and students can use to monitor progress toward the levels of writing performance they are expected to achieve. In short, the test has the potential to provide valuable information for monitoring the level of academic literacy being achieved by CUNY students in all tiers of the University.

The mandatory nature of the CPE has had some additional positive side effects beside those already noted in the previous section. Because of the high stakes involved, students are strongly motivated to do well on the test, making the test scores an accurate representation of students' ability to perform the tasks required by the test. Another byproduct is the wide availability of data for assessment purposes. The institutional research and assessment offices at each campus have access to CPE test score data for almost all students who have reached the 45$^{th}$ credit for use in the analysis and

monitoring of programs. When CPE test scores are matched with transcript data, CUNY colleges gain even more power to evaluate the effectiveness of instructional programs. However, unless the CPE is administered to the same students more than once, it is not possible to measure performance gains directly. Multiple administrations of the test to the same students probably are not feasible because of the cost of the exam. However, it would be possible to estimate which students are gaining more than others by using initial writing proficiency as measured by the CUNY assessment test in writing and other correlates of analytical writing to compute a predicted CPE score. By comparing predicted to actual CPE scores an analyst could focus attention on those colleges or programs that might be sources of promising practices.

Given that none of the assessments described above have been conducted, the Task Force believes that the full potential of the CPE for use in program monitoring has not been tapped. CPE scores (rather than pass rates) could be compared across majors, programs and subprograms to monitor progress toward standards defined in terms of the test. The CPE might also be used as a means of identifying which colleges or units within colleges are "beating the odds" in their ability to improve analytical writing, estimating initial proficiency as described above. Unfortunately, although nine years of CPE test scores are available to the colleges, so far, little use has been made of the data for assessment purposes.

*Public accountability*. The CPE has been used primarily as an internal accountability tool. As noted above, both show rates and pass rates are metrics included in the University's PMP, and colleges are held accountable for making progress toward University targets calling for improving compliance with the CPE requirement and success on the test.

For purposes of external accountability, however, the CPE has serious limitations. Because the test was designed by CUNY faculty and is administered only within CUNY, it is not possible to benchmark the achievements of CUNY students against those of comparable students at other institutions. The CPE does not provide a common yardstick against which the achievements of CUNY and non-CUNY students can be compared. Nor does the CPE, as currently administered, allow the University to measure improvements in analytic writing and quantitative reasoning. Because students take it only once, at the 45$^{th}$ credit, the CPE does not measure gains, though it could be used for this purpose, if the substantial funds needed to test twice or more were available. Still, even if the CPE were given more than once during a student's academic career, because the exam is administered only within CUNY, the public (and the CUNY community) would have no way of knowing how the gains of our students compare with those of students attending other institutions. The Board of Trustees' introduction of the test contributed to the University's subsequent success in addressing public concerns at the time about the rigor of CUNY's academic programs, but the CPE does not satisfy the demand for public accountability as it has evolved in recent years.

## Cost of the CPE

Because the CPE is a high-stakes test developed specifically for CUNY, administered only at CUNY, and administered to all CUNY students, it is expensive. The following factors contribute to its cost:

*Development.* Each prompt is field tested on a population of non-CUNY students similar to CUNY students and subjected to rigorous reviews for fairness. Because the test is constructed specifically for CUNY, development costs cannot be distributed across a larger number of institutions. Furthermore, to protect the security of the exam, multiple forms of the same test must be created for each administration.

*Scoring.* The vendor can allocate any fixed costs associated with scoring only to CUNY. Such costs might include the computer programming necessary to conduct the scoring, for example. Additional cost components can be attributed, once again, to the high-stakes nature of the test. Two readers score each portion of the test. When there is a discrepancy, a 3rd reader resolves it. Stringent scoring procedures are put in place to insure high inter-rater reliability, which is essential to any test, but especially to a high-stakes test. In addition, all papers scored 32 or 33, the two score points below the minimum passing score, are automatically rescored. About 5.6% of all test papers are subject to automatic appeal. Finally, a post-equating process is necessary so as not to disadvantage students who by chance may have received a comparatively difficult combination of long and short reading on Task 1 or a more difficult Task 2, and not to advantage students who received a less challenging combination of prompts.

*Administration.* The CPE poses a series of logistical challenges. Because the test is high stakes, test security is important. Each testing center employs a proctoring staff for each administration, with a recommended ratio of test takers to proctors of 20 to 1. Because the test is paper and pencil rather than online, record keeping for boxing and shipping the exams to be graded is labor intensive. Packing and shipping must be done with great care to protect the security of the exam and also to avoid lost examinations, which necessitate a retake. It should also be noted that at this time, because of the limitations of Pearson's scanning technology, the test is literally a paper and pencil test. Students must write their essays using pencils, an inconvenience for students and for testing offices, which must maintain a large supply of sharpened pencils. Colleges also devote substantial, though hard-to-quantify, resources to publicizing the test and urging students to register for and take the test. A record keeping system is necessary to schedule students for the test, and to keep track of no-shows, walk-ins, deferrals, and compliance with agreements by students to take an intervention if they have failed the test twice or more. These record-keeping requirements would have to continue to be met when the new CUNYfirst administrative software is introduced, requiring significant additional programming.

*Student support.* Because students must pass the test to graduate, CUNY has put in place a substantial support structure at each campus. Each campus has a CPE Liaison, who assists students with deferrals and appeals, and at some campuses, coordinates the development of preparatory workshops. Colleges provide an array of support, including

classes and workshops designed for students who are taking the test for the first time and support for students who have encountered difficulty with it. Students who fail the test a second time are required to take a writing intensive course.

Another factor contributing to the cost of the test is CUNY's policy of allowing students to attempt the test repeatedly if they fail it the first time. Students may sit for the test up to three times without an appeal and more if they receive permission to do so from the appeals committee in place at each campus.

An estimate of the total cost of administering the CPE in calendar year 2009 appears in Appendix G, Table G-1, column 1. In 2009, the University spent about $3.34 million on the CPE. Of this, $1.66 million went for development and production of the exam, including copyrights, and regular and appeals scoring**,** $1.56 million for central and campus-based staff costs, and $112,000 for OTPS.

Beginning with the January 2010 administration, the contract for development and scoring of the exam moved from ACT to Pearson, whose prices are significantly higher than historical costs. The projected expenditures for 2010 are shown in Appendix G, Table G-1, column 2. The total annual cost for calendar year 2010 compared to calendar year 2009 will be $4.92 million, as compared to $3.34 million. The development, production and scoring costs will almost double, from $1.66 million to $3.2 million. In future years these costs will continue to rise, both because of escalations in the Pearson contract and because of rising enrollments at CUNY. These estimates include the cost of student support, including CPE workshops.

If the use of the CPE for certification were discontinued, the cost of administering it would drop dramatically. Development costs would fall to $0 if forms were re-used from the University's library of 80 Task 1 forms and 54 Task 2 forms, and production costs would be reduced. If it were decided that the CPE would be useful as a writing assessment tool, the elimination of Task 2 would further reduce costs. If the test were no longer a degree requirement, scoring by two readers might no longer be necessary, although scoring by a single reader would reduce the rigor of the scoring process. The student support services now oriented to passing the CPE could be refocused as deemed most suitable. Table G-1, column 3, projects the costs for such a scenario, assuming that the test would be administered to all students at the 45$^{th}$ credit. Total costs for this option are estimated at $1.55 million, a 68% decrease from the 2010 estimated total cost.

## Review of Alternative College Proficiency Exams

Several standardized test instruments are now available in the marketplace for use in measuring general education outcomes. The Task Force reviewed them as possible options for the separate purposes of certification, program monitoring, and public accountability. The tests included in this comparison are the Council for Aid to Education's (CAE's) Collegiate Learning Assessment (CLA), ACT's Collegiate Assessment of Academic Proficiency (CAAP), and the Educational Testing Service's

Measure of Academic Proficiency and Progress (MAPP), recently renamed the ETS Proficiency Profile. These three instruments are the most widely adopted instruments of their kind, and all three instruments have been incorporated into the Voluntary System of Accountability (VSA). The VSA was launched in 2007 by the Association of Public and Land-grant Universities (APLU) and the Association of State Colleges and Universities (AASCU) as a mechanism for public four-year universities to provide comparable information to the public through an online report—the College Portrait. Participating institutions may select from among the CLA, CAAP or MAPP to demonstrate comparative levels of achievement and learning gains for their students. Table A (which appears below on page 40) contains a side-by-side comparison of the CPE, the CLA, CAAP and MAPP. Readers should not assume from this format that the tests are interchangeable. On the contrary, the table has been constructed to highlight design and implementation differences. Before proceeding with the comparison, it may be useful to introduce some key distinctions.

*Assessment of individuals versus institutions.* Assessments of individuals are designed to compare the performance of individuals to one another in order to assess each test taker's performance against a criterion or the population of test takers. Assessments of institutions aggregate the test scores of test takers attending those institutions to create institutional means. Institutional assessments compare institutions to one another on the basis of these mean scores in order to draw conclusions about the performance of the institution relative to a criterion or the population of institutions.

*Performance (constructed response) prompts versus multiple choice.* Richard Shavelson (2010) provides a clear discussion of the difference in approach between constructed response (performance) prompts and measurements comprised of multiple choice items. The latter arise from an empiricist psychometric tradition in which "everyday complex tasks are divided into component parts, and each is analyzed to identify the abilities required for successful performance" (p. 47). In the test development process, separate sets of multiple choice test questions are devised to measure each ability. When the finalized test instrument is administered, scores for each ability cluster are summed to create a total score representing the test taker's overall performance level. As Shavelson notes, "This approach, then, assumes that the sum of component part test scores equals holistic performance" (p. 47). The MAPP and CAAP adhere to this approach. In contrast, the CPE and CLA follow a criterion-sampling approach, in which tasks are sampled from the domain of behaviors that a population is expected to perform. The CLA and CPE prompt students to perform writing and analytical tasks that are expected of them in real life (or at least in upper division course work).

Typically, performance tests sacrifice some reliability for validity. Performance tasks are complex in nature, usually rely on human judgment for a score, and consist of a small number of prompts to sample behavior. The logistics of controlling variations in judgments (through rubrics) are complex, labor intensive, and expensive. Conversely, the scoring of multiple choice tests is objective, with human judgment reduced to a minimum and inexpensive. Constructs are measured not with one or two complex prompts but instead with multiple closed-response test items.

We now turn to a comparison of the testing instruments.

## Stated Purpose and Level of Assessment

The CPE and the CAAP have been designed primarily for student-level assessment, although test scores can be aggregated to assess academic units and the institution as a whole. Unlike the CPE and the CAAP, the CLA and MAPP are intended primarily for institutional assessment, although test results are reported to students and can be used for advisement and counseling. The CLA and MAPP vendors (CAE and ETS) caution that these tests are not intended for high-stakes decision making, arguing that their reliability is not high enough to support such a use. ACT also advises caution when using the CAAP results for high stakes. As noted earlier, because the CPE is used for high-stakes testing, the University has instituted two-reader scoring and automatic appeals, provided extensive support to students, and has afforded a minimum of three attempts to pass.

## Test Components

*CLA.* The CLA is administered entirely on computers linked to the internet, and consists of three constructed response prompts—The Performance Task (90 minutes) and two prompts designed to elicit analytical writing—Make-an-Argument (45 minutes) and Critique-an-Argument (30 minutes). In a typical administration, students take either the Performance Task or the two analytical writing tasks—totaling about two hours, with extra time allowed for the mechanics of test administration. In addition to these components of the CLA, students who have not taken the SAT or ACT are required to sit for the Scholastic Level Exam (SLE), a short-form 50-item measure of cognitive ability requiring just 12 minutes to administer. SAT, ACT and SLE scores are used to adjust CLA scores for initial ability, so that the institution can gain a more precise estimate of its contribution to any learning gains detected. The CAE is developing a shorter version of the CLA, requiring a maximum of 60 minutes, but as of August 2010 the CAE did not know when the shorter version would be ready.

The CLA's Performance Task is a rich and complex performance prompt that requires students to employ critical thinking, analytic reasoning, problem solving skills and written communication skills to answer a set of open-ended questions about a fictional but realistic scenario. The student is asked to read a narrative describing a situation, together with a set of instructions, and is given online access to a library of information sources such as letters, memos, summaries of research reports, newspaper articles, data displays and other documents. The students are expected to sift through these sources of evidence, weigh their value, and draw upon them in writing their responses to the questions in the prompt. Each Performance Task prompt tests a slightly different combination of skills, and consequently is graded with a variable combination of rubrics (See CAE, *Architecture of the CLA* Tasks). The CAE describes the mix of skills tested as follows:

*Performance Tasks require students to marshal evidence from different*

> *sources; distinguish rational from emotional arguments and fact from opinion; understand data in tables and figures; deal with inadequate, ambiguous, and/or conflicting information; spot deception and holes in the arguments made by others; recognize information that is and is not relevant to the task at hand; identify additional information that would help to resolve issues; and weigh, organize and synthesize information from several sources* (CAE, *Architecture of the CLA Tasks* p. 2.

The CLA also contains two subtests of analytical writing, both of which measure the student's ability to express complicated ideas, weigh assertions and evidence, support ideas, create a coherent argument, and express thoughts in conventional English. The first, "Make-an-Argument," asks students to support or reject a position on an issue. As an example, the CAE provides the following assertion: *Government funding would be better spent on preventing crime than in dealing with criminals after the fact.* Students are asked to take a clear position on either side of the question, and to support their position effectively. "Critique-an-Argument," presents a flawed argument and requires students to identify these logical flaws. Examples are confusion of correlation with causation and confusion of proportions with absolute numbers.

*CAAP.* The CAAP consists of five multiple choice modules: writing (usage) (72 items), math (35 items), critical thinking (32 items), reading (36 items) and science (45 items). The writing test measures students' command of punctuation, grammar, sentence structure, strategy, organization, and style. The test consists of six prose passages, each of which is accompanied by a set of 12 multiple-choice test questions. Separate subscores are provided for usage/mechanics and command of rhetoric. The math module tests students' command of mathematical operations in pre-algebra; elementary, intermediate, and advanced algebra; coordinate geometry; and trigonometry. The science test is designed to measure students' skills in scientific reasoning. Although contents are drawn from a range of sciences—biological sciences, chemistry, physics and the physical sciences—the test emphasizes scientific reasoning skills rather than scientific content knowledge. The critical thinking module measures students' skills in clarifying, analyzing, evaluating, and extending arguments. Students are presented with four passages containing a set of arguments supporting a conclusion and are asked a series of questions about these arguments. Finally, the reading test measures reading comprehension in terms of reasoning and referring skills.

The modules may be administered in any combination. In addition, students are required to write two 20-minute essays. The essays are identical in format to the CUNY/ACT essay that the University currently employs to assess the writing ability of students when they first apply. This is an assessment of basic writing skills and is scored holistically on a six-point rubric.

*MAPP.* The MAPP is entirely multiple choice, with 4 modules: writing (usage) (27 items), math (27 items), critical thinking (27 items) and reading (27 items). A short open-response essay is optional. The critical thinking and reading items are clustered around common reading selections or visual displays in groups of two to four questions.

Critical thinking and reading questions are distributed across the humanities, social sciences and natural science. The math modules test the ability to recognize and interpret mathematical terms, read tables and graphs, evaluate formulas, order and compare numbers, interpret ratios and percentages, read scientific instruments and recognize and use equivalent mathematical formulas and expressions.

## Scoring and Scaling

Subject matter specialists should pay close attention to any test's scoring and scaling. For tests that are graded by rubric, they should review the rubric dimensions and the definitions of the level of performance. For multiple choice tests, they should be interested in the domain tested and the representativeness of the items.

Scaling is the process of relating raw scores to an interpretable basis. There are three common ways to do this in educational measurement. One is to locate a raw score in terms of its percentile position in a population. CLA, CAAP and MAPP scores are scaled in this way-- expressed as the percentile of test-takers who achieved a particular raw score. The norms can be expressed on the basis of a specified population, which could be defined in terms of any number of attributes. A second approach to scaling is to compute the percentage of a domain of topics that the test taker answered correctly. For example the COMPASS pre-algebra placement test score represents the estimated percentage of the total pre-algebra domain that the test taker has mastered.

The third approach is placement on a rubric scale of performance. Rubric-based scoring is descriptive of a performance, e.g., "writes without grammatical errors," "writes with a few errors that do not impede meaning." The CPE is the only test of the four that employs such a rubric for both scoring and scaling. Like the CPE, the CLA employs grading rubrics, but several analytic and holistic rubrics are used to grade a single performance task and then summed to compute a raw score. The combination of rubrics changes depending on the prompt being scored. This score is then related to the norming population's SAT scores and expressed on the SAT scale. In contrast, the CPE is scored with a single four-dimension analytical rubric for Task 1 and a single rubric for Task 2 and reported without norming, on the same scale on which it was originally graded. Similar to the CLA, the CAAP essay is also rubric scored, but the reporting scale is expressed in terms of percentile performance of other CAAP test takers in the same demographic.

In addition to the scale scores for freshmen and seniors, the CAE provides participating institutions two valuable pieces of information based on the CLA scores. For each school, the CAE computes a discrepancy score separately for freshmen and for seniors. The discrepancy score is the difference between the CLA score that would have been predicted for the institution based on the SAT profile of its incoming students and the actual CLA scores observed for its students. A positive discrepancy score suggests that the institution has performed better than expected, given the quality of its students as measured by the SAT. In addition to the two discrepancy scores, the CAE also reports a

Value-Added score, which is computed as the difference between the senior discrepancy score and the freshman discrepancy score. This is intended as a measure of how much proficiency in critical thinking, analytic reasoning, problem solving, and written communication the institution has imparted to its students. These metrics allow the institution to benchmark its performance against that of other institutions. Both ACT and ETS have developed a similar methodology that colleges can use to estimate value added on the basis of the CAAP and MAPP, but these vendors do not provide value added scores as part of their standard reports.

## Reliability and Validity

*Reliability.* In general, multiple choice tests have higher reliability correlations than the performance-based tests. This is to be expected because of the greater number of items in multiple choice correlations, whereas the inter-rater reliabilities reported for performance-based tests are based on only two raters for each prompt. The CAAP multiple choice subtests achieve reliabilities ranging from .87 to .92, while the MAPP reliabilities range from .91 to .94. ACT reports an inter-rater reliability of .75 for the CAAP essay test. For the CLA Performance Test, CAE reports an inter-rater reliability of .76 to .87, depending on the specific prompt. The reliabilities for Make-an-Argument are somewhat lower, ranging from .57 to .70 and for Critique-an-Argument, from .77 to .84. For the CPE, inter-rater reliabilities for Task 1 are .58 to .74, below those associated with the CLA's Performance Task, while the reliability of the Task 2 scoring is much higher, at .93.

*Validity.* A crucial question in the evaluation of any test is the degree to which the test actually tests what it is expected to measure. One way to assess validity is to compute correlations between the test and other constructs that are similar and different from those measured by the test in question. By building a construct correlation matrix, one can determine whether test components correlate more highly with other tests of the same construct than with those of different constructs. Earlier, we showed that CPE Tasks 1 & 2 correlate consistently as expected with SAT, Regents, and GPA.

In the Test Validity Study (TVS), Klein and his colleagues (2009) created a construct correlation matrix for the CLA, CAAP, and MAPP, using the components of each test. Construction of such a matrix is possible because all three tests (13 subtests) were administered to the same students attending the 13 institutions participating in the study. This matrix is reproduced here in Table 15 (Appendix B). The top matrix shows the student-level correlations, separately for each construct: Critical Thinking, Writing, Mathematics, Reading, and Science, while the bottom matrix gives the school-level correlations. Because institutional means are more stable than individual scores, the correlations in the school-level matrix are much higher than those in the individual-level matrix. High school-level correlations are desirable if tests are to be used to assess institutions; high individual-level correlations are desirable if a test will be used to make assessments of individuals—for example certification or advisement.

If a test is valid, subtests of the same construct should correlate more highly with one

another than with subtests of different constructs. A review of both matrices reveals this pattern for the math and reading subtests, but the correlations among the writing subtests are not consistently higher than they are with subtests in the other constructs. The same can be said for critical thinking. Inspection of the school-level correlations for the key component of the CLA, the performance task, illustrates the pattern. The test correlates as follows with the other subtests in the critical thinking domain: .83 with the MAPP critical thinking subtest, .79 with the CAAP critical thinking subtest, and .73 with the CLA Critique-an-Argument subscores. However, the CLA performance task correlates even more highly with subtests outside of this domain: the MAPP writing test (.84), the MAPP mathematics test (.91), the CAAP mathematics test (.91) and the MAPP reading test (.90). Similarly the MAPP critical thinking subtest correlates more highly with the math modules of the MAPP and CAAP than it does with the CLA Performance Test, CLA Make-an-Argument, and the CAAP writing test.

Three additional patterns are noteworthy. First, the correlations involving the CAAP essay at both the student and college level are consistently lower than those of any of the other subtests. A review of the distribution of CAAP essay scores for all two- and four-year colleges, public and private (ACT, 2008) suggests a reason: only 3% of test takers scored in the top three score levels of a six-point score range, and only 8% scored at the bottom score point. Almost all test takers scored either a 2 or a 3, probably accounting for the low correlations throughout the construct matrices. Second, there is a strong test publisher/test format relationship. Tests that were developed by the same publisher and tests using same format (multiple choice versus constructed response) correlate more highly with one another than with the subtests of other publishers and with other formats. Third, subtests employing constructed response formats (CLA and CAAP essay) tend to have lower correlations than multiple choice tests.

If we move to considerations of authenticity, performance-based prompts are commonly regarded as better measures of critical thinking skills than are multiple choice tests of this construct. One study of faculty perceptions of the CLA documents their view that the performance test measures an important educational outcome and that the test measures what it is supposed to measure (Hardison and Valamovska, 2008). The CAE has gathered feedback from students as well, and finds that they perceive the CLA as measuring their ability to analyze and communicate (Shavelson, 2010).

These results by themselves do not dictate the adoption of any one of the three alternatives to the CPE. Although the reliability data for all three are reassuring, the validity results are not without ambiguity. Rather, such a choice must also consider a comparison of the tests with respect to their suitability for their intended purposes.

## Implementation logistics

When contemplating the use of any of these instruments, administrators must first decide whether to adopt a cross-sectional design, in which separate samples of freshmen and

upper classmen, typically seniors, are drawn and tested during the fall and spring semesters of the same academic year. The alternative is a longitudinal design, in which a sample drawn from a cohort of freshmen is given the test again once or twice during their academic career at the college. The cross sectional design is more practical and yields results much more quickly. A study by the CAE and funded by Lumina found no clear advantage to administering the CLA using the longitudinal design (Klein, 2009). To our knowledge no similar studies have been conducted for the MAPP or CAAP tests.

The administration of these instruments faces a common set of challenges. The first and most important is gaining buy in from faculty. If test results are to be taken seriously for improvement of teaching and learning, the faculty must embrace the test as a valid measure of key abilities that a college-educated person should acquire. Acceptance of the test begins with a discussion of the broad instructional goals of the college and the best means of assessing progress toward them. The Council of Independent Colleges (2008) noted with respect to its members' efforts at adopting the CLA as a standard measure of learning outcomes that faculty buy-in was a significant hurdle:

*Virtually all the institutions in the Consortium have had to address faculty resistance to the CLA and have struggled to get students to take the CLA. On campuses where the CLA has been introduced through administrative channels (such as the president or vice president for academic affairs), faculty members resisted it because they perceived it as a "top-down" initiative. On other campuses, some faculty members have initially found it too time consuming, a distraction from other work, or have resisted efforts they perceive as moving toward "teaching to the test." As with many campus discussions, greater success seems to come when there is a shared commitment and transparency about efforts to assess and improve student learning.*

Faculty acceptance is critical to the success of any standardized exam, whether it be the CLA, the CPE, or either of the alternatives.

A second challenge is sampling. If the test is to be a valid indicator of learning gains, the sample of test takers must be representative of the college. A simple request for volunteers would likely yield a more highly motivated, selected group of students than the average undergraduate. It is necessary to identify a random or representative sample, invite these students to take the test, and then motivate them to show up for the testing session. The urgency of converting invitees into test takers is all the more critical in a longitudinal design, because high attrition rates over time can be fatal to the assessment effort.

A third challenge is motivation to perform well on the test. However, the CLA Testing Manual (p. 44) points out that the goal, at least for the CLA, is to measure typical rather than maximal performance, since typical performance is a closer approximation to performance in real life situations such as the work place. A fourth challenge has to do with the length of the test. The CLA, CAAP and MAPP are divided into modules that can be administered separately or in clusters. By engaging in matrix sampling, it is possible to administer all modules to members of the sample without any member taking all

modules.

## Cost Estimates

The cost of administering all the instruments considered here depends on a number of decisions, including the following:

1. The number of students taking the test. The larger the number of students tested, the greater the ability of the institution to assess learning in subgroups of it students. The CAE recommends that institutions using the cross-sectional design sample a minimum of 100 freshmen and 100 seniors. ACT cautions users not to sample fewer than 25 students in a subgroup of interest, while ETS provides guidance on the sample size needed to obtain given confidence levels but does not recommend specific Ns.
2. Scoring of the essay. Colleges administering the CAAP may choose to score the writing sample locally or pay ACT $13.50 per essay to score it. For colleges using the MAPP the choice is whether or not to give the essay portion of the test, which is optional. ETS charges $5.00 to score each essay.
3. Incentives. Students are not inherently motivated to sit for a lengthy examination that is not a degree requirement. Colleges have tried a variety of strategies to motivate sampled students to keep their testing appointment and complete the tests, including payment of as much as $50, course extra credit, and embedding the test as a required element of a course or activity such as freshman orientation.
4. Investments in IT infrastructure. Some institutions have integrated the test into their registration systems.

To place our comparative discussion of costs in proper context, we must recall our earlier discussion of the current cost of administering the CPE to all CUNY undergraduates as they reach the 45$^{th}$ credit. As summarized in Table G-1, the University spent $3.34 million in 2009. Because of the higher cost of scoring and enrollment growth, total expenditures will increase to $4.92 million in 2010. These figures include all costs associated with the test, including development, scoring, academic support, test administration at the campuses, and expenditures by the central Office of Academic Affairs.

Table G-2 (Appendix G) shows the results of a cost comparison of the CPE, CLA, CAAP and MAPP assuming an administration of the test to 200 freshmen and 200 seniors. (This scenario assumes that the CPE would be used for assessment rather than for certification, as is the case now.) In addition to the direct cost of acquiring and scoring the tests, these estimates factor in CUNY central and campus expenses associated with administering the examination, as well as a rich incentive of $50 per student. The costs of such an administration are fairly similar for all four alternatives, ranging from $767,477 for the CPE to $929,783 for the CAAP. If the tests are to be used to assess learning gains for subgroups of students, the sample size for each

college will be larger than the 200 freshmen and 200 seniors modeled here, and costs will be higher.

**Table A**
**Summary Comparison of the CPE CLA, CAAP and MAPP**

| Test Characteristic | CPE | CLA | CAAP | MAPP |
|---|---|---|---|---|
| **Stated Purpose** | Certify readiness for upper division work | CLA Assessment Services provide a means for measuring an institution's contribution to the development of key higher order competencies, including the effects of changes to curriculum and pedagogy. To gauge summative performance authentically | Help determine if individual students are adequately prepared for upper-division coursework, and if they're not, what interventions may be appropriate | 1) demonstrate program effectiveness for accreditation and funding purposes. 2) conduct various studies, such as cross-sectional and longitudinal, using ETS Proficiency Profile data to determine how much their students are learning and how they can improve learning outcomes. |
| **Level of analysis** | Student,<br>Institution | Institution | Student,<br>Institution | Institution<br>Student |
| **High stakes** | Yes | No | Caution | No |
| **Test Components** | **Analytic Reading and Writing** – read an 8-9 page college level text in advance and a 1-1 ½ page selection at test. Write a focused essay drawing on the relationship between the two passages and extend it to your own experience, understanding or ideas.<br><br>**Analyzing and Integrating material from graphs and text –** Two graphs and a brief passage all on the same topic, but independent of one another, are given out at the test. Task is to identify claims from the passage and describe relationships with information | **Performance Task** – read a short scenario that poses a dilemma, along with a variety of documentation such as newspaper articles, emails, photographs, charts and reports. Task is to answer a series of open ended questions.<br><br>**Make-an-Argument** – Given a one sentence topic in the form of an argument – e.g., "Government funding would be better spent on preventing crime than in dealing with criminals after the fact." Task is to plan and write a critical essay.<br><br>**Critique-an-Argument** – Task is to read a short one paragraph | **Writing skills** – 72-item multiple choice<br>**Math** – 35 item multiple choice<br>**Critical thinking** – 32 item multiple choice based on reading passages<br>**Writing essay** – two 20 minute essays in response to a situational prompt<br>**Reading test** – 36 item multiple choice based on reading passages<br>**Science test** – 45 item multiple choice based on reading passages and graphs | **Reading** – 27 item multiple choice based on reading passages<br>**Critical thinking** – 27 item multiple choice based on reading passages<br>**Writing** – 27 item multiple choice<br>**Math** – 27 item multiple choice |

| Test Characteristic | CPE | CLA | CAAP | MAPP |
|---|---|---|---|---|
| | from the graphs. | passage posing an argument and to write an essay explaining what is wrong with the argument. | | |
| **Scoring** | Rubric, human | Rubric, human (PT)<br>Computer (CA, MA) | MC, computer<br>Rubric, human (essay) | MC, computer |
| **Scaling** | Rubric | Normed | Normed | Normed |
| **Inter-rater reliability** | Task 1 .58-.74<br>Task 2 .93[2] | PT .76-.87<br>MA .57-.70<br>CA .77-.84[3] | Essay .75 | NA |
| **MC reliability** | NA | NA | .87-.92 | .91-.94 |
| **Validity correlations with various external criteria** | .27-.50Task 1<br>.18-.28 Task 2<br>.29-.49 (CPE Total)[4] | .32-.58 (PT)<br>.40-.52 (CA)<br>.37-.47 (MA) | .32-.75(Critical thinking)<br>.44-.72(writing)<br>.28-.40 (essay)<br>.39-.76 (math)<br>.44-.76 (reading)<br>.28-.74 (science) | .34-.86 (critical thinking)<br>.33-.76 (writing)<br>.29-.76 (math)<br>.31-.86 (reading) |
| **Motivation for participation** | Degree requirement, curriculum aligned | Recommends incorporation into curriculum | Recommends incorporation into curriculum | No guidance |
| **Test time** | 3 hours | 90 minutes (PT)<br>75 minutes (CA+MA)<br>15 minutes (optional SLE for students without SAT or ACT scores) | 40 minutes per module | 2 hours<br>40 minute abbreviated form. |
| **Test mode** | Paper & pencil | Online | Paper & pencil or online | Paper & pencil or online |
| **Cost** | | | | |
| **CPE 2010** | $4.92 million | | | |
| **Cross-sectional assessment** | $767,477 | $812,594 | $929,783 | $790,397 |

[2] Tables 6-10, Inter-rater agreement crosstabulations (p. 55-57)
[3] Test Validity Study, reliability and validity statistics for CLA, CAAP, and MAPP (Klein, 2009)
[4] Table 12, Construct Validity Correlation Matrix (p. 59)

## Summary and Analysis

*CLA*. The CLA has been designed primarily as an assessment of a core set of cognitive abilities commonly associated with general education-- critical thinking, analytic reasoning, problem solving, and written communications skills. It also is well suited for use as an instrument for external accountability because of the ease with which institutional results can be adjusted for the ability of incoming students, learning gains measured, and these gains benchmarked against the performance of other institutions. Performance on the test is normatively scaled, facilitating comparison across colleges. The CLA is not well suited for certification testing because it was constructed to assess the performance of institutions or their subunits rather than individuals, according to the CAE. Used as an institutional assessment, the CLA is positioned as a standardized authentic measure of college-level skills and ability that can be assessed both longitudinally and cross-sectionally and controlled for entering ability (SAT or ACT scores).

Aside from the CPE, the CLA is the only test among the three that is a performance test—based on a sampling of real-life behaviors—rather than an indirect test that relies on multiple choice items. It may therefore have more face validity among faculty who object to the use of multiple choice tests and who agree that the CLA's prompts elicit the cognitive abilities they value in general education. The CAE recommends that the CLA be embedded as a requirement in a course or other activity as the best method for assuring motivated participation in the sampling tests, though monetary and other rewards are probably more commonly employed.

The CAE provides an array of support services for colleges that wish to use the CLA as an assessment tool. The Performance Task Academy trains faculty to develop performance tasks similar to those in the CLA using their own subject area materials and expertise. The objective of the Academy is to bring activities that develop critical thinking and analytical reasoning into the classroom. The CAE also maintains a library of performance tasks on which faculty can draw for this purpose. In addition to the Academy, the CAE is offering student and institutional diagnostic reporting as an aid to faculty seeking to provide feedback to students about their critical thinking skills.

*CAAP*. Because it was designed for the assessment of individuals, the CAAP is the only test that could be a candidate to replace the CPE as a certification instrument. It could also be considered for program monitoring and public accountability efforts.

The strong point of the CAAP is its topic coverage, with six subtests: math, reading, writing skills, writing essay, science, and critical thinking. The subtests can be administered in 40-minute modules, which makes them less intrusive than longer performance tests. All but the essays are multiple choice in format, which makes them fast and inexpensive to score, but with some sacrifice of face validity. All scores are reported on a norm-referenced scale. The writing task is not an analytic one. In fact, as

mentioned earlier, the CAAP essay prompt is identical to the prompt that CUNY has been using to assess basic proficiency in writing since 2001. That test is to be replaced with a new assessment, starting fall 2010. In a correlation matrix of CLA, CAAP, and MAPP subtest scores designed to explore their construct validity, the CAAP essay did not differentiate its relationship with non-writing versus writing subtests, and was at the low end of the correlation range with all subtests (Klein, Liu, & Sconing, 2009).

*MAPP*. One advantage of the MAPP is its brevity in administration and scoring. There are 4 subtests: critical thinking, reading, writing, and math, each with 27 multiple choice items. The entire test takes 2 hours, and a 40 minute version is available. The intent of the MAPP is to obtain a sampling of ability at the institutional level. The writing subtest is not an essay and focuses on usage and mechanics, though an optional essay is available. Because of its summative nature, the test is not recommended for use as a certification instrument, but may be suitable for program monitoring and accountability.

*Certification.* For certification, the viable options are the CPE or the CAAP. To go this route, the University would have to rework Task 2 in order to improve the prompt's validity as a measure of quantitative reasoning. The most urgent problem with the exam, however, is its high cost. While far less expensive, the CAAP is currently unknown to CUNY faculty, and measures a much broader range of abilities and knowledge primarily on the basis of multiple choice questions. Its ability to measure the qualities of academic writing that CUNY faculty value is questionable.

*Program monitoring*. If CUNY no longer regards certification as a primary need for standardized assessment and wishes to invest in a CUNY-wide instrument to bolster program monitoring, then the CPE and the three off-the-shelf tests can all be considered for this purpose. To proceed in this direction, faculty and program administrators must assess the degree of alignment between the goals of the course or program they wish to assess and the knowledge and abilities tested by the CPE, CAAP, MAPP and CLA. If a performance assessment is desired, the University faces a choice between the CPE and the CLA. The CLA tests a much broader and richer array of abilities, is less expensive, can measure learning gains, and can benchmark these gains against those of institutions external to CUNY.

*Public accountability.* The CLA, MAPP and CAAP are all potentially useful as instruments for documenting comparative learning gains for external audiences. For this purpose, the CPE could not be used, since it is not possible to benchmark outside CUNY. If the cost of the CPE could be reduced, it might be feasible to measure learning gains by administering the exam in the freshman year and again either at the 45$^{th}$ credit or in the senior year, depending on whether the goal is to assess general education or the entire undergraduate career. However, the University could not learn whether these gains are any greater or less than the learning gains of comparable students at other institutions.

The CLA appears to have an advantage for this purpose because of the higher face validity of its performance-based assessments and its growing acceptance among institutions of higher education. As more institutions adopt the CLA, the opportunities

for benchmarking increase. The CAE provides gain scores as part of its standard reporting package, unlike the MAPP and CAPP. Recently, the CAE introduced a version of the CLA designed for community colleges—the CCLA.

When administering a standardized assignment of collegiate learning for external accountability, the standard approach is to test a random sample of freshmen and seniors. This is not always easy. If the sampling is not random, the results may not be representative of the institution. It is usually necessary either to require students to test or to create an attractive incentive for them to do so, raising the cost of the test. Finally, because of attrition, the seniors are a selected group, in many ways dissimilar to entering freshmen. (This is an issue both in a longitudinal design and in a cross-sectional approach.) Although the major testing organizations now attempt to control for these dissimilarities with SAT or ACT scores, the correlations between these tests and the CLA are only moderate at the individual level and may not control for factors such as motivation or other correlates of performance on the CLA on which freshmen and seniors could differ.

## Recommendations for CUNY

The Task Force recommendations are the product of extensive discussions that often included diverse points of view. The discussions were stimulating and rich, and hopefully characteristic of the conversations about assessment that will take place in the wider CUNY community after release of this report. Although members of the Task Force sometimes initially took a stance on an issue, there was openness throughout our discussions to consideration of testing alternatives, including departures from the status quo. It was difficult to make specific recommendations because the issues are complex and nuanced. Yet the Task Force agreed unanimously about the importance of obtaining information for assessment and improving students' academic literacy, no matter which tests are used.

**Recommendation 1. After much discussion, the Task Force reached consensus that CUNY should discontinue the use of the CPE as a high-stakes certification exam.**

As currently used, the CPE does not appear to add much additional information about readiness for upper division work once grades and credits have been considered. In order to sit for the CPE, a student must have demonstrated basic proficiency in reading and writing and earned 45 credits with a GPA of 2.0 or better. Virtually every student who can meet this standard can pass the CPE. The university-wide longitudinal pass rate is now 93%, and the raw CPE score for Task 1 has remained basically flat. In addition, both the New York State English Regents and the SAT critical reading test are more highly correlated with grades earned in upper division course work than is the CPE. When the high cost of administering the CPE (projected to be $4.9 million in 2010) is considered as well, it seems impractical to continue administering the CPE as a certification test.

If the University wishes to keep in place a certification test, only one other test considered in this report is a possible candidate to replace the CPE – the CAAP. It is the only test designed for student-level analysis that includes a performance writing task graded by rubric. The other two tests, the CLA and the MAAP, are not designed or recommended by their publishers to be used for high stakes testing. Before entertaining the CAAP test seriously, however, the University and its faculty should consider whether its reliance on multiple choice items, the broad domain of the test, and the basic level of the writing sample will meet its needs.

**Recommendation 2. Consideration should be given to retaining Task 1 of the CPE as an instrument for program monitoring. Because of the value of the prompt as a standard assessment tool, CUNY colleges should consider whether it is feasible to somehow embed the exam in the curriculum of some courses, thereby insuring that CUNY students meet or exceed standards of academic literacy.**

The CPE, particularly Task 1, has value as a tool for assessing mastery of the key elements of academic literacy: comprehension of collegiate texts, the ability to synthesize ideas in texts, critical analysis, and control of written language. Because the test is criterion referenced against a rubric, it is relatively easy to interpret the test scores in light of this fixed standard. Moreover, because the test has become institutionalized over the past nine years as a degree requirement, this standard is familiar to many members of the faculty, and accepted by them as valid. Finally, because the rubric measures four distinct dimensions, the prompt can be useful as a tool for diagnosing students' strengths and weaknesses.

As Figure 1 (Appendix B) demonstrates, Task 1 raw scores are distributed more or less normally, with a large enough standard deviation to place students along a gradient of proficiency in academic literacy. Consequently, the instrument may be helpful in measuring progress toward goals of improving writing beyond the minimum level of proficiency signified by the current cut point on the CPE.

The Task Force is aware that the cost of the CPE is extremely high. If the test were no longer high stakes, however, the cost could be reduced substantially in a number of ways. CUNY could recycle its extensive library of test forms. The test could be read once rather than twice, with random quality control measures. Appeals would no longer be necessary. The cost of supplementary support and the administrative overhead associated with the exam would be reduced substantially.

To capitalize on its potential value as an assessment tool and to motivate students to continue to sit for the test and perform well on it once it were no longer a high stakes test, the Task Force recommends that the test be embedded in the curriculum. Colleges and programs could then be held to agreed-upon standards, and test results could be used by

colleges to monitor the achievement of CUNY students. Of course, widespread use of the CPE in this way could be hampered by the current fiscal constraints.

Up to now, the CPE has not been used widely at CUNY as an assessment tool. Assessment offices have only recently begun to take a closer look at the available CPE data and to consider how it might be used to evaluate their college's learning outcomes. If the CPE is retained for assessment purposes, the University should institutionalize the dissemination of CPE assessment data to institutional research and assessment offices at its colleges.

In addition to facilitating the continuing use of the CPE by colleges wishing to employ it as an assessment tool, the University must maintain its commitment to academic literacy, through such programs and activities as Writing Across the Curriculum (WAC) Writing in the Disciplines (WID), writing intensive courses, and appropriate academic support for students.

**Recommendation 3. If the CPE is to be retained for any purpose, Task 2 should be revisited.**

The validity of the CPE is compromised by the scoring and weighting of Task 2. The scoring of Task 2 is prescriptive, and the task requires students to respond in a specific way. It appears that many students who score well on Task 1 but low on Task 2 do so simply because they do not understand the prompt. However, once understood, Task 2 is seen as testing a skill that is not difficult. It certainly is not a prompt that deserves to be weighted as heavily as it is. Data show that Task 2 has an undue effect on CPE scaled scores that actually *reduces* CPE validity correlations with post-CPE academic outcomes. Moreover, the prompt lacks face validity with faculty. Given the growing recognition of the importance of quantitative reasoning skills across the University, it is clear that Task 2 should be rethought.

**Recommendation 4. There are a number of reasons why a nationally normed standardized test instrument might have a place in the University's assessment agenda. Assessment begins with a definition of the program, the program goals, and a plan of action, and assessment tools should be chosen to fit the purpose and goals of the program. The choice of a test must be made deliberatively and in consultation with faculty and program managers. Further, no one test can fulfill all purposes; effective assessment requires a battery of different types of tools, and those tools will differ for different campuses and disciplines.**

The University should promote and assist with the creation of a culture of evidence and continuous improvement. Of utmost importance is the conversation that takes place among stakeholders as they establish goals and progress indicators and analyze feedback

on what is working and what is not, and use that information for future improvement. The emphasis placed on standardized tests in addition to locally developed ones will depend on the nature of the instructional activity and the usefulness of the standardized test results for making adjustments. Faculty and academic administrators can be expected to welcome data from standardized tests if the data are meant to guide program improvement.

**Recommendation 5. The Task Force advises experimentation with publicly benchmarking CUNY colleges if it can be done without compromising the primary function of enhancing students' learning, if the benchmarking methodology is sound, and if the cost is reasonable.**

The Task Force recognizes the importance of public accountability, but urges caution if the University decides to adopt an instrument for this purpose. Because performance scores are highly correlated with the SAT and ACT, students attending less selective institutions tend to score lower on the MAPP, CAAP and CLA. CUNY must take care to educate members of the public about the distinction between level of performance and the "value added" by institutions serving less well prepared students. Another concern is with sampling methodology. The MAPP, CAAP and CLA are norm referenced, and one must take care to ensure that test-takers are representative of their colleges.

If the University adopts a new accountability test, it should consult with faculty, both to select the best instrument and to plan how to use the results not only for accountability but also for the improvement of teaching and learning.

## References and Works Cited


ACT (2008). *CAAP User Norms 2008-2009*. Iowa City: ACT.

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

Council for Aid to Education. *Architecture of the CLA Tasks.* Available at http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf

Council for Aid to Education. *2009-2010 CLA Administration User Manual*. New York: Council for Aid to Education.

Chun, M. (2009). Looking Where the Light Is Better: A Review of the Literature on Assessing Higher Education Quality. *Peer Review*, Winter/Spring 2002, 16-25.

Council of Independent Colleges (2008). *Evidence of Learning: Applying the Collegiate Learning Assessment to Improve Teaching and Learning in the Liberal Arts College Experience.* Washington, D.C.: Council of Independent Colleges.

Ewell, P. T. (2009). *Assessment, Accountability, and Improvement: Revisiting the Tension.* Urbana-Champaign: National Institute for Learning Outcomes Assessment.

Hardison, C.M. & Vilamovska, A. (2009). *The Collegiate Learning Assessment: Setting Standards for Performance at a College or University*. Santa Monica, CA: Rand Education.

Klein, S. (2009). *The Lumina Longitudinal Study Summary Findings*. New York: Council for Aid to Education.

Klein, S., Freedman, D., Shavelson, R., and Bolus, R. (2008). Assessing School Effectiveness. *Evaluation Review, 32,* 511-525.

Klein, S., Liu, O. L., & Sconing, J. (2009). *Test Validity Study (TVS) Report.* Washington, D.C.: FIPSE.

Schmeiser, C. B., & Welch, C. J. (2006). Test Development. In R. L. Brennan (Ed.), *Educational Measurement. Fourth Edition. ACE/Praeger Series on Higher Education* (pp. 307-354). Lanham, MD: Rowman & Littlefield Education.

Shavelson, R. (2010). *Measuring College Learning Responsibly: Accountability in a New Era.* Stanford, CA: Stanford University Press.

Shulenburger, D. & Keller, C. (2009) *The FIPSE Learning Outcomes Test Validity Study: Findings Bearing on Validity When the Institution is the Unit of Analysis*. Washington, D.C.: U.S. Department of Education.

U.S. Department of Education (2006). *A Test of Leadership: Charting the Future of American Higher Education (Report of the Commission appointed by Secretary of Education Margaret Spellings).* Washington, D.C.: U.S. Department of Education.

## Appendix A – CPE Task Force Members and Staff

| Member | Campus | Title |
|---|---|---|
| Raymond Moy, Chair | CUNY Office of Assessment | Director of Assessment |
| Michael Anderson | Brooklyn College | Director, Academic Assessment |
| Nancy Aries | CUNY Central Office, Academic Affairs | Interim University Dean for Undergraduate Education |
| Bonne August | New York City College of Technology | Provost & Vice President for Academic Affairs |
| Lenore Beaky | LaGuardia Community College | Professor of English, UFS Representative |
| Rex Butt | Bronx Community College | Professor of Communications, CPE Liaison |
| Cynthia Haller | York College | Professor of English, Chair CPE Advisory Committee |
| Eda Henao | Borough of Manhattan Community College | Professor of Modern Language |
| Hilary Klein | CUNY Central Office, Legal Affairs | Associate General Counsel |
| Howard Kleinmann | Queens College | Director, Academic Support |
| Sharona Levy | Brooklyn College | Professor, SEEK, UFS Representative |
| Keith Markus | John Jay College of Criminal Justice | Professor, Psychology |
| Sherri Ondrus | CUNY Central Office, Academic Affairs | Director, Performance Management Process |
| Kevin Sailor | Lehman College | Professor, Psychology |
| David Shimkin | Queensborough Community College | Professor of English, CPE Liaison |
| James Stellar | Queens College | Provost & Vice President for Academic Affairs |
| **Staff, CPE Task Force** | | |
| **Staff** | **Campus** | **Title** |
| Melissa Uber | CUNY Office of Assessment | Director of Testing |
| Eve Zarin | CUNY Office of Assessment | Faculty Liaison & CUNY Chief Reader |

# Appendix B – Tables

**Table 1**
**CPE Dimension Scores and Total Scores**
**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Organization | 148,850 | 2 | 12 | 5.67 | 1.616 |
| Critical Reading | 148,850 | 2 | 12 | 5.74 | 1.629 |
| Development of Ideas | 148,850 | 2 | 12 | 5.74 | 1.562 |
| Language | 148,850 | 2 | 12 | 7.61 | .929 |
| CPE Task 1 | 148,850 | 8 | 48 | 24.77 | 5.233 |
| CPE Task 2 | 148,702 | 2 | 12 | 7.49 | 3.582 |
| CPE Total Raw | 148,868 | 10 | 60 | 32.25 | 6.934 |
| CPE Total Scaled | 148,898 | 15 | 69 | 41.57 | 9.601 |
| Valid N (listwise) | 148,654 | | | | |

**Table 2**
**CPE Dimension Scores and Total Scores**
**Correlation Matrix**

| | Organization | Critical Reading | Development of Ideas | Language | CPE Task 1 | CPE Task2 | CPE Total Raw | CPE Total Scaled |
|---|---|---|---|---|---|---|---|---|
| Organization | 1 | .937 | .925 | .537 | .972 | .202 | .840 | .651 |
| Critical Reading | .937 | 1 | .882 | .524 | .957 | .205 | .831 | .647 |
| Development of Ideas | .925 | .882 | 1 | .526 | .952 | .193 | .820 | .630 |
| Language | .537 | .524 | .526 | 1 | .664 | .108 | .558 | .426 |
| CPE Task 1 | .972 | .957 | .952 | .664 | 1 | .203 | .862 | .666 |
| CPE Task 2 | .202 | .205 | .193 | .108 | .203 | 1 | .671 | .837 |
| CPE Total Raw | .840 | .831 | .820 | .558 | .862 | .671 | 1 | .937 |
| CPE Total Scaled | .651 | .647 | .630 | .426 | .666 | .837 | .937 | 1 |

**Table 3**
**Trends in the Longitudinal Pass Rate on the CPE**

| | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| Senior Colleges | 94.0 | 93.4 | 93.4 | 93.1 | 94.6 |
| Comprehensive Colleges | 91.5 | 90.2 | 88.4 | 89.4 | 89.8 |
| Community Colleges | 91.0 | 89.2 | 88.8 | 90.5 | 91.5 |
| **Total CUNY** | 92.7 | 91.5 | 91.1 | 91.5 | 92.5 |

A longitudinal pass rate is calculated over three administrations (October, January and March) for students required to sit for the exam for the first time in October.

**Table 4**
**Trends in the Percentage of Required Invitees Who Took the CUNY Proficiency Exam (CPE Show Rate)**

| | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| Senior Colleges | 77.9 | 77.3 | 81.7 | 84.4 | 85.0 |
| Comprehensive Colleges | 74.3 | 79.6 | 80.1 | 81.5 | 81.7 |
| Community Colleges | 75.3 | 78.2 | 81.2 | 82.1 | 81.8 |
| **Total CUNY** | 76.4 | 78.1 | 81.2 | 83.1 | 83.2 |

Percentaged on a cohort of students required to take the exam for the first time in October. A student who actually sat for the exam in the October administration or in either the subsequent January or March exams is counted as having taken the exam.

**Table 5**
**Three-year Outcomes by Status after the First Administration: October 2005***

| INITIAL STATUS | DEFERRED | | ABSENT | | FAILED | | PASSED | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|
| | N = 848 | | N = 5,108 | | N = 2,504 | | N = 1,3971 | N = 22,413 | |
| | 4% | | 23% | | 11% | | 62% | 100% | |
| | PASS RATE FOR INITIALLY DEFERRED | | PASS RATE FOR INITIALLY ABSENT | | PASS RATE FOR THOSE WHO INITIALLY FAIL | | PASSED | TOTAL INITIAL + EVENTUAL PASS RATE | |
| | N | % | N | % | N | % | N | N | % |
| AFTER 1 YEAR | 557 | 66% | 3,752 | 73% | 1,606 | 64% | 13,971 | 19,886 | 89% |
| AFTER 2 YEARS | 662 | 78% | 4,343 | 85% | 1,858 | 74% | 13,971 | 20,834 | 93% |
| AFTER 3 YEARS | 698 | 82% | 4,554 | 89% | 1,931 | 77% | 13,971 | 21,154 | 94% |

* Results are reported for a cohort of students who were required to take the test for the first time in October 2005.

## Table 6: Agreement of Raters 1 and 2 in the Scoring of Task 1, Organization Dimension

| Rater 1 | | Rater 2: 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|---|
| 1 | Observed | 3415 | 1547 | 147 | 28 | 5 | 0 | 5142 |
| | Expected | 150.4 | 1648.2 | 2221.1 | 995.7 | 109.6 | 17 | |
| 2 | Observed | 1232 | 39517 | 11412 | 1096 | 75 | 3 | 53335 |
| | Expected | 1560.2 | 17095.7 | 23037.9 | 10328.1 | 1136.4 | 176.7 | |
| 3 | Observed | 135 | 10517 | 49644 | 9458 | 347 | 22 | 70123 |
| | Expected | 2051.3 | 22476.8 | 30289.4 | 13579 | 1494 | 232.4 | |
| 4 | Observed | 21 | 1045 | 9397 | 20153 | 1067 | 78 | 31761 |
| | Expected | 929.1 | 10180.5 | 13719.1 | 6150.4 | 676.7 | 105.3 | |
| 5 | Observed | 6 | 82 | 402 | 1021 | 1907 | 125 | 3543 |
| | Expected | 103.6 | 1135.7 | 1530.4 | 686.1 | 75.5 | 11.7 | |
| 6 | Observed | 2 | 7 | 36 | 91 | 103 | 317 | 556 |
| | Expected | 16.3 | 178.2 | 240.2 | 107.7 | 11.8 | 1.8 | |
| | Total | 4811 | 52715 | 71038 | 31847 | 3504 | 545 | 164460 |

*r* =.74

## Table 7: Agreement of Raters 1 and 2 in the Scoring of Task 1, Critical Reading Dimension

| Rater 1 | | Rater 2: 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|---|
| 1 | Observed | 3453 | 1624 | 154 | 41 | 6 | 1 | 5279 |
| | Expected | 159.1 | 1604.4 | 2292 | 1085.1 | 120.6 | 17.8 | |
| 2 | Observed | 1327 | 36059 | 12136 | 1043 | 95 | 4 | 50664 |
| | Expected | 1526.8 | 15398.2 | 21996.9 | 10414.4 | 1157.1 | 170.7 | |
| 3 | Observed | 140 | 11220 | 48587 | 10190 | 371 | 25 | 70533 |
| | Expected | 2125.5 | 21437 | 30623.5 | 14498.6 | 1610.9 | 237.6 | |
| 4 | Observed | 30 | 990 | 10096 | 21248 | 1178 | 85 | 33627 |
| | Expected | 1013.3 | 10220.2 | 14599.9 | 6912.3 | 768 | 113.3 | |
| 5 | Observed | 4 | 85 | 393 | 1192 | 1982 | 118 | 3774 |
| | Expected | 113.7 | 1147 | 1638.6 | 775.8 | 86.2 | 12.7 | |
| 6 | Observed | 2 | 6 | 38 | 92 | 124 | 321 | 583 |
| | Expected | 17.6 | 177.2 | 253.1 | 119.8 | 13.3 | 2 | |
| | Total | 4956 | 49984 | 71404 | 33806 | 3756 | 554 | 164460 |

*r* =.73

## Table 8: Agreement of Raters 1 and 2 in the Scoring of Task 1, Development of Ideas Dimension

| | | | Rater 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Rater 1 | 1 | Observed | 1311 | 288 | 29 | 13 | 0 | 0 | 1641 |
| | | Expected | 16.2 | 551 | 716.4 | 307.6 | 43.3 | 6.4 | |
| | 2 | Observed | 260 | 41617 | 12704 | 968 | 85 | 5 | 55639 |
| | | Expected | 549.1 | 18683.4 | 24291.6 | 10429.2 | 1468.3 | 217.5 | |
| | 3 | Observed | 34 | 12259 | 48886 | 9851 | 428 | 26 | 71484 |
| | | Expected | 705.5 | 24004 | 31209.4 | 13399.2 | 1886.4 | 279.5 | |
| | 4 | Observed | 13 | 947 | 9665 | 18559 | 1382 | 75 | 30641 |
| | | Expected | 302.4 | 10289.1 | 13377.6 | 5743.5 | 808.6 | 119.8 | |
| | 5 | Observed | 4 | 105 | 483 | 1339 | 2309 | 153 | 4393 |
| | | Expected | 43.4 | 1475.2 | 1918 | 823.4 | 115.9 | 17.2 | |
| | 6 | Observed | 1 | 9 | 35 | 97 | 136 | 384 | 662 |
| | | Expected | 6.5 | 222.3 | 289 | 124.1 | 17.5 | 2.6 | |
| | | Total | 1623 | 55225 | 71802 | 30827 | 4340 | 643 | 164460 |

*r* =.72

## Table 9: Agreement of Raters 1 and 2 in the Scoring of Task 1, Language Dimension

| | | | Rater 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Rater 1 | 1 | Observed | 16 | 4 | 2 | 1 | 0 | 0 | 23 |
| | | Expected | 0 | 0.2 | 4.3 | 17.6 | 0.7 | 0.1 | |
| | 2 | Observed | 7 | 624 | 970 | 146 | 1 | 0 | 1748 |
| | | Expected | 0.2 | 17.4 | 325.7 | 1339.9 | 56.9 | 7.9 | |
| | 3 | Observed | 0 | 876 | 17483 | 13585 | 81 | 5 | 32030 |
| | | Expected | 4.5 | 318.2 | 5967.2 | 24552.6 | 1042.9 | 144.5 | |
| | 4 | Observed | 0 | 128 | 12111 | 109783 | 2367 | 125 | 124514 |
| | | Expected | 17.4 | 1237.1 | 23197 | 95446.3 | 4054.3 | 561.8 | |
| | 5 | Observed | 0 | 2 | 71 | 2402 | 2733 | 169 | 5377 |
| | | Expected | 0.8 | 53.4 | 1001.7 | 4121.7 | 175.1 | 24.3 | |
| | 6 | Observed | 0 | 0 | 2 | 150 | 173 | 443 | 768 |
| | | Expected | 0.1 | 7.6 | 143.1 | 588.7 | 25 | 3.5 | |
| | | Total | 23 | 1634 | 30639 | 126067 | 5355 | 742 | 164460 |

*r* =.58

## Table 10: Agreement of Raters 1 and 2 in the Scoring of Task 2

| | | | Rater 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | **Total** |
| Rater 1 | 1 | Observed | 29749 | 51 | 3570 | 235 | 40 | 10 | 33655 |
| | | Expected | 6884.9 | 413 | 9195.6 | 3827.6 | 4871.3 | 8462.7 | |
| | 2 | Observed | 60 | 1735 | 72 | 132 | 18 | 4 | 2021 |
| | | Expected | 413.4 | 24.8 | 552.2 | 229.8 | 292.5 | 508.2 | |
| | 3 | Observed | 3510 | 62 | 38108 | 534 | 1901 | 944 | 45059 |
| | | Expected | 9217.9 | 552.9 | 12311.5 | 5124.5 | 6521.9 | 11330.2 | |
| | 4 | Observed | 235 | 133 | 526 | 16958 | 605 | 299 | 18756 |
| | | Expected | 3837 | 230.2 | 5124.7 | 2133.1 | 2714.8 | 4716.3 | |
| | 5 | Observed | 48 | 28 | 1730 | 539 | 19973 | 1256 | 23574 |
| | | Expected | 4822.6 | 289.3 | 6441.2 | 2681.1 | 3412.1 | 5927.8 | |
| | 6 | Observed | 8 | 7 | 884 | 287 | 1243 | 38799 | 41228 |
| | | Expected | 8434.2 | 505.9 | 11264.8 | 4688.8 | 5967.4 | 10366.9 | |
| | | Total | 33610 | 2016 | 44890 | 18685 | 23780 | 41312 | 164293 |

*r* =.93

**TABLE 11**
**GPA in Course Work Completed Within One YEAR AFTER CPE**
**BY FAIL/ PASS CROSSTABULATION***

| | | | Fail | Pass | Total |
|---|---|---|---|---|---|
| GPA 1 Year After CPE | = < 2.0 | N | 1569 | 3827 | 5396 |
| | | % of row | 29.1% | 70.9% | 100.0% |
| | | % of column | 10.2% | 5.3% | 6.1% |
| | = 2.0-2.9 | N | 9725 | 30218 | 39943 |
| | | % of row | 24.3% | 75.7% | 100.0% |
| | | % of column | 63.5% | 41.7% | 45.5% |
| | = 3.0-3.49 | N | 3191 | 22911 | 26102 |
| | | % of row | 12.2% | 87.8% | 100.0% |
| | | % of column | 20.8% | 31.6% | 29.7% |
| | = 3.5-4.0 | N | 838 | 15558 | 16396 |
| | | % of row | 5.1% | 94.9% | 100.0% |
| | | % of column | 5.5% | 21.5% | 18.7% |
| | **Total** | N | 15323 | 72514 | 87837 |
| | | % of row | 17.4% | 82.6% | 100.0% |
| | | % of column | 100.0% | 100.0% | 100.0% |

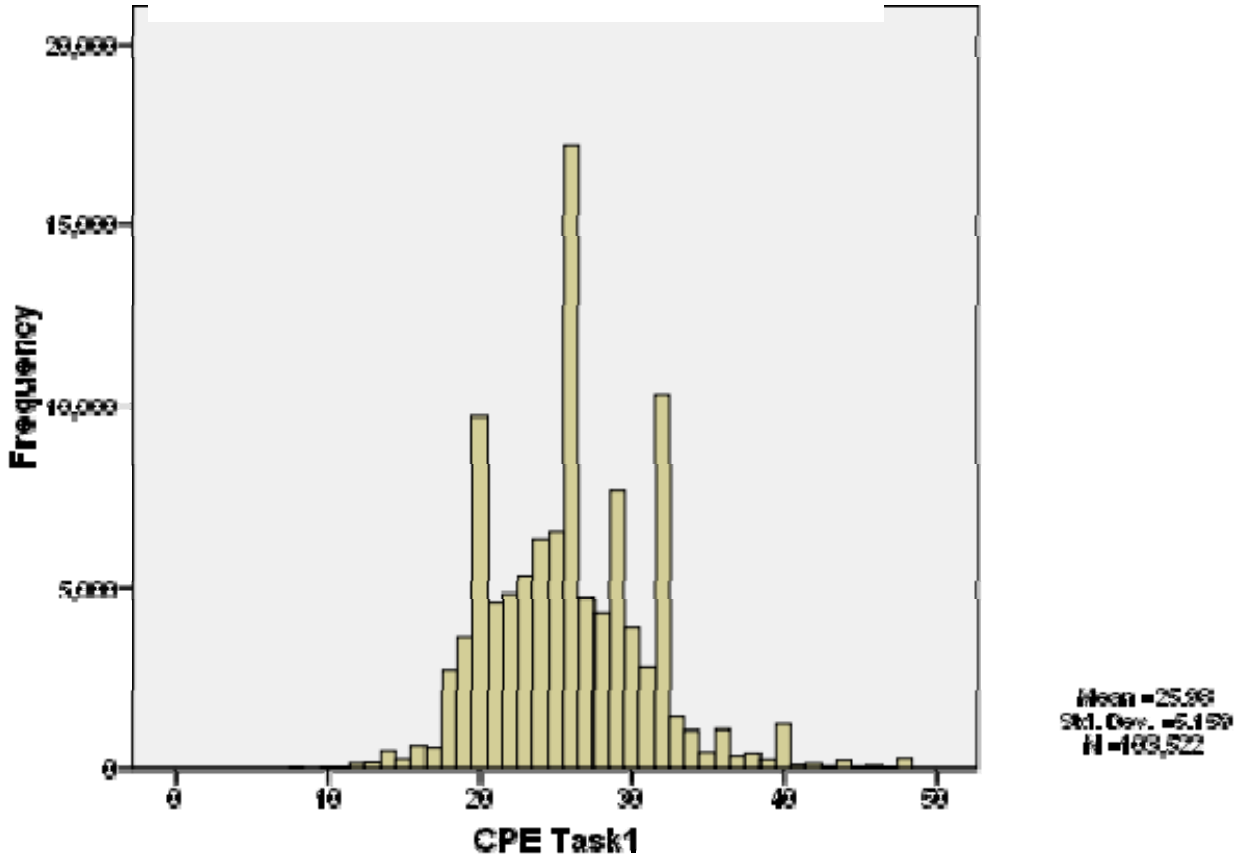
* Pearson chi-square significance <.000

**TABLE 12**
**GPA in Course Work Completed Within One AFTER CPE BY CPE PROFICIENCY LEVEL CROSSTABULATION***

| GPA 1 Year After CPE | | | Low | Border Low | Border High | High | Total |
|---|---|---|---|---|---|---|---|
| | < 2.0 | N | 382 | 2292 | 2391 | 327 | 5392 |
| | | % of row | 7.1 | 42.5 | 44.3 | 6.1 | 100.0 |
| | | % of column | 13.7 | 9.3 | 5.2 | 2.3 | 6.1 |
| | 2.0-2.9 | N | 1874 | 14742 | 19896 | 3403 | 39915 |
| | | % of row | 4.7 | 36.9 | 49.9 | 8.5 | 100.0 |
| | | % of column | 67.4 | 59.5 | 43.3 | 23.8 | 45.5 |
| | 3.0-3.49 | N | 429 | 5937 | 15106 | 4608 | 26080 |
| | | % of row | 1.6 | 22.8 | 57.9 | 17.7 | 100.0 |
| | | % of column | 15.4 | 24.0 | 32.9 | 32.3 | 29.7 |
| | 3.5-4.0 | N | 96 | 1811 | 8535 | 5944 | 16386 |
| | | % of row | 0.6 | 11.1 | 52.1 | 36.3 | 100.0 |
| | | % of column | 3.5 | 7.3 | 18.6 | 41.6 | 18.7 |
| | Total | Total N | 2781 | 24782 | 45928 | 14282 | 87773 |
| | | % of Total | 3.2 | 28.2 | 52.3 | 16.3 | 100.0 |

* Pearson chi-square significance <.000

**Figure 1**
**Distribution of CPE Task 1 Scores – Fall 2005 through Summer 2009 Test Administrations**



## TABLE 13
## Correlations Among Indicators of College Readiness, CPE Scores and CUNY Grades (Fall 2005 Cohort of Test Takers)

| | | CPE Task 1 | CPE Task 2 | CPE Total Raw | CPE Total Scaled | GPA 1 Year Later | SAT Verbal | Regents - English | Cum GPA at Grad | SAT Math | Regents Math |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CPE Task 1 | *r* | - | | | | | | | | | |
| | N | 16444 | | | | | | | | | |
| CPE Task 2 | *r* | .245** | - | | | | | | | | |
| | N | 16419 | 16432 | | | | | | | | |
| CPE Total Raw Score | *r* | .893 | .653 | - | | | | | | | |
| | N | 16444 | 16424 | 16449 | | | | | | | |
| CPE Total Scaled | *r* | .736 | .825 | .956 | - | | | | | | |
| | N | 16444 | 16432 | 16449 | 16457 | | | | | | |
| GPA 1 Year Later | *r* | .330 | .243 | .371 | .351 | - | | | | | |
| | N | 11821 | 11812 | 11824 | 11827 | 16456 | | | | | |
| SAT Verbal | *r* | .502 | .195 | .488 | .416 | .371 | - | | | | |
| | N | 5794 | 5792 | 5794 | 5795 | 6050 | 7629 | | | | |
| Regents - English | *r* | .456 | .222 | .463 | .410 | .367 | .625 | - | | | |
| | N | 6034 | 6029 | 6034 | 6035 | 6148 | 6683 | 7937 | | | |
| Cum GPA at Grad | *r* | .353 | .255 | .397 | .373 | .830 | .438 | .437 | - | | |
| | N | 12284 | 12279 | 12287 | 12291 | 12460 | 5550 | 5771 | 15936 | | |
| SAT Math | *r* | .346 | .282 | .408 | .393 | .388 | .625 | .520 | .460 | - | |
| | N | 5794 | 5792 | 5794 | 5795 | 6050 | 7629 | 6683 | 5550 | 7629 | |
| Regents Math | *r* | .271 | .222 | .322 | .309 | .340 | .379 | .468 | .415 | .672 | - |
| | N | 1803 | 1800 | 1803 | 1803 | 1771 | 1838 | 2101 | 1661 | 1838 | 2112 |

** All correlations are significant at the 0.01 level (2-tailed).

**Table 14**
**College-Level Construct Correlation Matrix for the CPE**

| | CPE Task1 | CPE Task2 | CPE Total Raw | CPE Total Scaled | GPA 1 Yr after CPE | Cum GPA Grad | SAT Verbal | SAT Math | Regents - English | Regents Math |
|---|---|---|---|---|---|---|---|---|---|---|
| CPE Task1 | 1 | .702 | .979 | .939 | .894 | .771 | .922 | .837 | .734 | .886 |
| CPE Task2 | .702 | 1 | .832 | .895 | .665 | .621 | .725 | .745 | .509 | .707 |
| CPE Total Raw | .979 | .832 | 1 | .987 | .887 | .778 | .924 | .864 | .716 | .892 |
| CPE Total Scaled | .939 | .895 | .987 | 1 | .886 | .802 | .881 | .827 | .637 | .888 |
| GPA 1 Yr after CPE | .894 | .665 | .887 | .886 | 1 | .924 | .751 | .649 | .483 | .884 |
| Cum GPA Grad | .771 | .621 | .778 | .802 | .924 | 1 | .661 | .587 | .364 | .856 |
| SAT Verbal | .922 | .725 | .924 | .881 | .751 | .661 | 1 | .968 | .910 | .791 |
| SAT Math | .837 | .745 | .864 | .827 | .649 | .587 | .968 | 1 | .917 | .760 |
| Regents - English | .734 | .509 | .716 | .637 | .483 | .364 | .910 | .917 | 1 | .530 |
| Regents Math | .886 | .707 | .892 | .888 | .884 | .856 | .791 | .760 | .530 | 1 |

## Table 15
## Correlation Matrix of CLA, CAAP, & MAPP Components

Table 2a.

*Student-level correlation matrix with standard correlations shown above the diagonal*

| Construct(s) | Test | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. | 9. | 10. | 11. | 12. | 13. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Critical Thinking | 1. MAPP | | 0.75 | 0.53 | 0.52 | 0.76 | 0.45 | 0.68 | 0.34 | 0.63 | 0.46 | 0.86 | 0.76 | 0.74 |
| | 2. CAAP | | | 0.58 | 0.47 | 0.66 | 0.39 | - | 0.32 | 0.57 | - | 0.71 | - | 0.74 |
| | 3. CLA PT | | | | - | 0.50 | - | 0.49 | 0.32 | 0.46 | 0.40 | 0.55 | 0.52 | 0.52 |
| | 4. CLA CA | | | | | 0.48 | 0.47 | 0.49 | 0.40 | 0.46 | 0.44 | 0.49 | 0.50 | 0.50 |
| Writing | 5. MAPP | | | | | | 0.44 | 0.72 | 0.33 | 0.60 | 0.51 | 0.73 | 0.70 | 0.63 |
| | 6. CLA MA | | | | | | | 0.44 | 0.37 | 0.40 | 0.39 | 0.43 | 0.46 | 0.39 |
| | 7. CAAP | | | | | | | | - | 0.58 | 0.48 | 0.70 | 0.71 | - |
| | 8. CAAP Ess. | | | | | | | | | 0.29 | - | 0.31 | - | 0.28 |
| Mathematics | 9. MAPP | | | | | | | | | | 0.76 | 0.60 | 0.55 | 0.71 |
| | 10. CAAP | | | | | | | | | | | 0.46 | 0.44 | - |
| Reading | 11. MAPP | | | | | | | | | | | | 0.76 | 0.70 |
| | 12. CAAP | | | | | | | | | | | | | - |
| Science | 13. CAAP | | | | | | | | | | | | | |

Table 2b.

*School-level correlation matrix with standard correlations shown above the diagonal and reliabilities shown on the diagonal*

| Construct(s) | Test | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. | 9. | 10. | 11. | 12. | 13. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Critical Thinking | 1. MAPP | 0.93 | 0.93 | 0.83 | 0.93 | 0.96 | 0.85 | 0.89 | 0.62 | 0.95 | 0.93 | 0.96 | 0.82 | 0.93 |
| | 2. CAAP | | 0.87 | 0.79 | 0.87 | 0.94 | 0.79 | 0.91 | 0.75 | 0.90 | 0.86 | 0.93 | 0.76 | 0.95 |
| | 3. CLA PT | | | 0.75 | 0.73 | 0.84 | 0.67 | 0.77 | 0.58 | 0.91 | 0.91 | 0.90 | 0.76 | 0.86 |
| | 4. CLA CA | | | | 0.85 | 0.92 | 0.90 | 0.90 | 0.61 | 0.82 | 0.77 | 0.91 | 0.91 | 0.79 |
| Writing | 5. MAPP | | | | | 0.91 | 0.86 | 0.97 | 0.70 | 0.92 | 0.90 | 0.96 | 0.87 | 0.90 |
| | 6. CLA MA | | | | | | 0.84 | 0.83 | 0.67 | 0.74 | 0.72 | 0.82 | 0.86 | 0.69 |
| | 7. CAAP | | | | | | | 0.88 | 0.74 | 0.83 | 0.78 | 0.93 | 0.89 | 0.81 |
| | 8. CAAP Ess. | | | | | | | | 0.75 | 0.57 | 0.56 | 0.62 | 0.71 | 0.61 |
| Mathematics | 9. MAPP | | | | | | | | | 0.94 | 0.98 | 0.94 | 0.71 | 0.98 |
| | 10. CAAP | | | | | | | | | | 0.92 | 0.91 | 0.70 | 0.96 |
| Reading | 11. MAPP | | | | | | | | | | | 0.91 | 0.86 | 0.91 |
| | 12. CAAP | | | | | | | | | | | | 0.88 | 0.65 |
| Science | 13. CAAP | | | | | | | | | | | | | 0.92 |

# Appendix C – Task 1: Analytical Reading and Writing

### Sample Writing Assignment and Student Essays

This task is based on reading selection A, , "Two Ways of Thinking about Money" by Jerome M. Segal, which you were given to read and study in advance and on Reading Selection B, "There's No Place Like Work" by Arlie Russell Hochschild. The readings are printed below. , "There's No Place Like Work" and review "Two Ways of Thinking about Money" in light of the writing assignment, which is printed following Reading Selection B below.

---

## Sample Reading Selection A

### Two Ways of Thinking about Money *

Jerome M. Segal

(From Segal, Jerome M. *Graceful Simplicity: The Philosophy and Politics of the Alternative American Dream.* Copyright © 1999. The University of California Press.)

In popular imagery, especially when seen from afar, America is often portrayed as if there were only one meaning to the American Dream. This is not so. The ambivalent response that many in the world have toward American life is mirrored in an ambivalence that many Americans have toward their own life, and this is an essential part of the American tradition, even when people are making it in America. There is always that nagging question, "Is this really the way to live?"

Long before there was an America, there were two American Dreams, and they reflect two ways of thinking about money. In Western thought, from the very beginning to the present day, people had doubts about the real value of riches and the things money can buy. There has always been a conflict between the view that "more is better" and the view that "just enough is plenty."

This divide is reflected in two very different visions of the good life. It is the underlying thesis of this essay that the Alternative Dream, the dream that rests upon the attainment of a simple life, is the sounder vision.

#### Aristotle's Challenge to our Way of Life

This essay is about contemporary life, but I want to start with Aristotle for two reasons. First, because his challenge to a money-oriented form of life remains as powerful today as it was 2,300 years ago. Second, because, for all his wisdom, Aristotle never had to wrestle with the problems we face. So many of the contemporary problems that prevent people in the middle class from enjoying the good life are the unanticipated consequences of three forms of genuine moral and social progress that Aristotle never envisioned: the elimination of slavery, the liberation of women, and the affirmation of the right of ordinary working people to self-fulfillment. Seeing both the strengths and weaknesses in Aristotle gives us a clearer perspective on our own situation.

Aristotle's *Politics* is surprising in that it opens with a discussion of the household. But this is exactly the right touchstone for both politics and economics. The household is a central ground of the good life, and all economic arrangements must be judged by whether they enable the household to perform its function as locus and support for the human good. This is one of the central messages of this essay: we must put the proper functioning of the household at the center of the way we think about economic life.

The core issue, as Aristotle puts it, is property and "the art of acquisition"—that is, how people make a living. He starts with the observation that there are a variety of different modes of subsistence and that this gives rise to a variety of different ways of life. This is as true among animals as it is of humans. Some animals live in herds, and others live in isolation. Some eat plants and others meat. Among human beings, Aristotle identifies five "natural" ways of life: pastoral, farming, fishing, hunting, and, interestingly, piracy. What he calls "true wealth" is acquired through these activities and consists of the amount of household property that suffices for the good life. This he regards as a limited amount. We can call this the perspective that "just enough is plenty."

In distinction to these modes of acquisition that supply the household with its needs, there is a second form of the art of acquisition, which Aristotle believes to be "unnatural":

> The other form is a matter only of retail trade, and it is concerned only with getting a fund of money, and that only by the method of conducting the exchange of commodities.

> The acquisition of wealth by the art of household management as contrasted with the art of acquisition in its retail form has a limit; and the object of that art is not an unlimited amount of wealth.[1]

The difference is between an approach to acquisition that views it as functional to the life of the household and one in which it takes on a life of its own, such that it reproduces unchecked without regard to the larger life of the organism, and ultimately undermines that life. It is the very description of what we now understand as cancer.

What Aristotle presents here isn't just an academic distinction, but a clash between two different ways of life, each captured by a different way of thinking about money. In the first, money and the things one can buy with it play an important but limited role. Life is not about money. It is not about getting rich. It is about something higher, whether it is philosophy, or art, or the pursuit of knowledge, or participation with one's fellow citizens in the ever-absorbing process of governing the democratic *polis* [city or city-state]. Every person lives within a household, and the household has its economic needs—but the point is to attain only what is sufficient to enable one to turn away from money-getting and undertake the real activities of life.

In this first vision of life, only some ways of making a living are viewed by Aristotle as acceptable. His list of farmer, hunter, fisherman, herdsman, or pirate has an arbitrary quality to it. What is important is what these choices are intended to rule out. The one thing you cannot do is spend your life grubbing for money. You do not become a businessman, a retail trader, a man of commerce. These all represent a kind of slavishness to money. Nor (one would hope) do you find yourself so destitute that you must work for someone else, for that, too, is a form of slavery. The good life requires some degree of good fortune. Ideally for Aristotle, you are born financially independent.

But how do people manage to go so wrong about money? How does it gain such control over their lives? Aristotle suggests that this emerges from a deep misconception about the nature of human happiness; it is this that leads to the focus on the pursuit of higher and higher levels of consumption and of the higher income necessary to sustain them.

Aristotle identifies what he terms "external goods"; these externals include wealth, property, power, and reputation. These are the elements that make up the standard vision of success both then and now. To these, Aristotle contrasts elements of character, what he terms the "goods of the soul": fortitude, temperance, justice, and wisdom.[2] This is a familiar distinction, between inner and outer, between matters of worldliness and matters of virtue. We continue to make these distinctions when we are reflective, not so much about our own lives, but when we think about what we want for our children—are we more concerned that our children be rich and successful or that they develop into good human beings? We tell them that these "externals" are not what is really important in life, and we hope that they will listen.

Aristotle tells us that happiness "belongs more to those who have cultivated their character and mind to the uttermost, and kept acquisition of external goods within moderate limits."[3] Those who lose in life are those "who have managed to acquire more external goods than they can possibly use and are lacking in the goods of the soul."[4] (For "soul" we might substitute "character" or "mental health.")

Of course, one might say, "Why the either/or? Why not have both?" But Aristotle, and many others, thought that we really do have to choose. In explaining the relationship between externals and the good life, Aristotle tells us: "External goods, like all other instruments, have a necessary limit of size… any excessive amount of such things must either cause its possessor some injury, or at any rate, bring him no benefit."[5]

Aristotle is saying that with all external goods, we find that the more we have, the less utility we receive from each additional amount, and that at some point "any excessive amount" does us no good and may even harm us. In other words, the pleasure from the first ice-cream cone is greater than from the second, and most of us can hardly eat a third.

Translated into a thesis about money, Aristotle's formulation tells us that, beyond a given level, additional increments of money are not only useless, but negative in their effect. Translated into a thesis about the society at large, it suggests that economic growth beyond a given point is actually harmful to human happiness. It is a straightforward rejection of the idea that "more is better."

Aristotle goes further in his account. For Aristotle the issue is even more serious than a life of wasted pursuit. The pursuit of higher and higher levels of income results in a distortion of the personality, such that we never come to be the persons that we most truly are; we are divorced from our truest selves. Instead, people are "led to occupy themselves wholly in the making of money… using each and every capacity in a way not consonant with its nature."[6]

It should be clear that Aristotle's critique is not merely about certain specific economic activities (e.g., retail sales as opposed to production). It is an indictment of a general outlook and form of life. When these become dominant in society, the object of criticism is then the entire form of social life or civilization. Such a civilization, and I believe Aristotle would include much of the modern world in this category, is to be condemned as representing a distortion of human nature and a general thwarting of the possibility of human fulfillment.

When every human capacity gets placed at the service of obtaining money, *we ourselves are transformed and distorted*.[7] That's why you can't have it all—why there is conflict between the two American Dreams—who "you" are changes through the choices you (and your household) make toward matters of acquisition, careers, "success." Within the Aristotelian framework, to say that our capacities—that is, our selves—are separated from their proper function is to say that we are thus denied self-actualization or human fulfillment. It is also to say that we are thus denied the possibility of living well; for to live well for Aristotle is to express one's richest potentials at high levels of excellence.

Thus, Aristotle, in his analysis of the limited place of money in the good life, and in his emphasis on how absorption in acquisition undermines both the healthy personality and the good life, can be seen as the intellectual father of a philosophy of simple living. But before leaving Aristotle, we must recognize the other side of the picture. Aristotle was not a believer in the general equality of all men and women. Though he did not believe that specific races of people deserve to be enslaved, he believed that there were some individuals who were "natural slaves" in that they lacked the capability of governing themselves. Of course, at some point in life—when we are children—we all lack this capability. But Aristotle believed that a significant class of adult males, and women generally, lacked the capability to govern themselves.

These views about the naturalness of slavery and the subservience of women turn out to have an intimate relationship to the question of simple living, and to graceful living in particular. Ultimately, I want to argue that most wealth resides in the ability to draw on the services of other people. We normally think of such wealth as residing in financial assets (e.g., money, stocks, bonds, real estate), but it can equally reside in relationships (e.g., friendship, parent-child relationships, marriage). It can also reside in institutionalized relations of unequal power such as slavery, rigid class distinctions, and the domination of women. When one has access to the services of others through such institutional structures, it is indeed easier to live well with less money; one has found nonmonetized ways of accessing valued services. The great challenge is thus to find a way to live simply and well, not only without excessive dependence upon money, but without reliance on unjust social institutions such as slavery, patriarchy, and rigid class systems.

For Aristotle, this never really clicked into place. While he recognized that not all who were in fact slaves were of a "slavish nature," he did not challenge slavery itself. Likewise, for Aristotle, the existence of mass poverty does not emerge as a problem. With his acceptance of the naturalness of slavery and the subservience of women, and his acquiescence to the limited servitude of workers, the socioeconomic frameworks of the Greek city-states in which he lived fit neatly into a theory of human development. The city-state or *polis* is the environment within which human fulfillment occurs. That the vast majority of persons simply fall by the wayside does not raise any pressing problems in Aristotle's worldview. Having limited potential, they reach their full development within subservient roles. Indeed, it is really not until the eighteenth century that the equality of ordinary people in their entitlement and potential for achieving the highest levels of human development finds support from political ideologies and groups. And it is not until the twentieth century that equality begins to be substantially extended to women.

In spite of these flaws, what Aristotle did do remains of enormous importance. He challenged the idea that acquiring more and more things was good for the individual. He set his critique of commercial and acquisitive forms of life within a theory of human development that stressed the exercise and perfection of distinctly human capacities, capacities that are distorted and stunted if we allow economic pursuits to dominate our lives. We have lost sight of much that Aristotle has to teach us with respect to the place of the economic within the good life: *The point of an economy, even a dynamic economy, is not to have more and more; it is to liberate us from the economic—to provide a material platform from which we may go forth to build the good life. That's the Alternative American Dream.*

## Simple Living and American Dreams

We entirely mistake our own history if we think of simple living as some recent fad. The idea of simple living has always been part of the American psyche—sometimes central, sometimes only a minor theme, but always present. From the earliest days of the American experience, advocates of simple living have challenged consumerism and materialism. Simple living, especially in America, has meant many things.[8] For Christians, the central inspiration for a life of simplicity has been the life of Jesus. In the hands of the Puritans, this emerged as a life of religious devotion, a lack of ostentation, and plenty of hard work. It was certainly not a leisure expansion movement as it is today. Nor was simple living a matter of individual choice; laws about consumption invoked the power of the state to restrict conspicuous display, and economic life was regulated to limit the role of greed in human affairs.

In the hands of the Quakers, the concept of the simple life underwent an evolution. For the Puritans, at least part of the motivation for sumptuary laws was to prevent those in the lower classes from putting on the manners of those above them; among Quakers, the restrictions on display and consumption became more widely applicable. Most important, the pursuit of luxurious consumption was linked to a broad range

of injustices and social problems, including alcoholism, poverty, slavery, and ill treatment of the Indians. Here, perhaps, are the origins of a radical politics of plain living—the belief that if people adopted the simple life, all of society would be transformed.

The key Quaker theorist of the simple life was John Woolman. Central to Woolman's thought was the recognition that people could be "necessitated to labor too hard." Thus, he maintained that "every degree of luxury of whatever kind and every demand for money inconsistent with divine order hath some connection with unnecessary labor." Woolman saw his listeners' desire for luxurious consumption as the core motive that resulted in slavery, the practice "of fetching men to help to labor from distant parts of the world, to spend the remainder of their lives in the uncomfortable conditions of slaves." He also identified selfishness as the cause of past wars, telling us to "look upon our treasures, and the furniture of our houses, and the garment in which we array ourselves, and try [to see] whether the seeds of war have nourishment in these our possessions, or not." Were Woolman alive today, it is likely that he would extend his critique, arguing that excessive consumption, and the desire for it, is at the root of both the drug and environmental problems we face. Indeed, Woolman would probably have been receptive to the idea that the harsh poverty of many Third World countries emerges from the excessive consumption of the rich nations.[9]

In the mid-1700s, in the years prior to the American Revolution, the ideas of simple living and democratic government were intertwined. For many of the leaders of the Revolution, however, the ideal was not the simple life of Jesus, but the simple life of the self-governing citizens of ancient Greece and Rome. Key figures in the revolutionary period, in particular Samuel Adams, were deeply concerned about the relationship between our political health and the individual pursuit of luxury. The rebirth of democracy in the world brought with it an interest in the ancient Greek and Roman experiments, and why they disappeared. There was a concern (as there is today) with the virtue of officeholders. Genuine democracy seemed incompatible with too great an absorption in getting rich. There was great fear of the corrupting influences of unbridled commercialism. When the colonists boycotted British goods, it was not just a tactic of the independence movement; Britain was viewed as the Great Satan, exporting the corruptions of capitalism.

Benjamin Franklin's views on these questions are also worth noting; they, too, have a contemporary echo. In Franklin we have an unusual mixture: the espousal of frugality, hard work, and restrained consumption as the vehicles for getting ahead, the central patterns of behavior that will lead to wealth. Franklin was concerned with how the average person might remain free in his own life, be his own master. He warns of the perils of spending and in particular of borrowing. The great thing is to save. Franklin also warned that the dangers of excessive consumption are easily missed. In this vein, Franklin rails against going into debt. Credit cards would have seemed to him the instruments of our undoing. "What Madness must it be to run in Debt for these Superfluities!… think what you do when you turn in Debt; you give to another Power over your Liberty.… Preserve your Freedom; and maintain your Independence: Be Industrious and free; be frugal and free."[10]

Filled with a sense of adventure and experiment, but less interested in accumulating wealth, was Henry David Thoreau. In *Walden*, he looked about himself and saw mostly foolishness—people not knowing how to grab hold of the gift of life. With words that had echoes of Aristotle, he told Americans that our necessities are few, yet we subject ourselves to endless labor. He described a world that had taken the wrong turn. "The twelve labors of Hercules were trifling in comparison with those which my neighbors have undertaken; for they were only twelve and had an end."[11] Wealth itself is a curse because it enslaves us. "I see young men, my townsmen, whose misfortune it is to have inherited farms, houses, barns, cattle and farming tools; for these are more easily acquired than got rid of." We miss that which is best in life. "Most men, even in this comparatively free country, through mere ignorance and mistake, are so occupied with the factitious cares and superfluously coarse labors of life that its finer fruits cannot be plucked by them."[12]

Yes, the necessities must be met, "for not till we have secured these are we prepared to entertain the true problems of life with freedom and a prospect of success."[13] But "most of the luxuries, and many of the so-called comforts of life are not only not indispensable, but positive hindrances to the elevation of mankind. With respect to luxuries and comforts, the wisest have ever [always] lived a more simple and meager life than the poor."[14] For Thoreau, it is not necessity that enslaves us. Rather we have become the "slave-drivers" of ourselves, "the slave and prisoner of our own opinion of ourselves." Once we have satisfied our necessities, rather than laboring for superfluities, it is time to "adventure on life." But few undertake this adventure. Instead, "the mass of men lead lives of quiet desperation."[15] It is from a disease of the spirit that Thoreau recoils.

Thus Thoreau called Americans away from their over-absorption with economic life, from their self-subjugation to a life of toil. Unlike earlier advocates of simple living, he was not calling people to religion or to civic engagement; rather he was calling us as individuals to find our own nature, to define ourselves at a higher level of experience. He called for simple living in order to enable the life of the mind, of art, literature, poetry, philosophy, and an almost reverential engagement with Nature.

Interest in simple living was harder to find in the post-Civil-War period, but it reemerged powerfully

toward the turn of the century. There was a reaction against materialism and the hectic pace of urban life. In those days it was *The Ladies' Home Journal* (of all things) that led the charge against the dominant materialist ethos. Under a crusading editor, Edward Bok, it served as a guide for those in the middle class seeking simplicity.

After World War II, as after World War I, the Civil War, and the American Revolution, there was a surge in consumption, and simple living receded into the background. But again in the 1960s there was a critique of the affluent lifestyle and a renewed interest in plain living. In the 1970s, with the energy crisis, this merged with a broad environmentalism. Many saw the energy crisis not as an economic or political problem to be overcome, but as an occasion for a spiritual renewal that would turn us away from the rampant materialism of modern life. One of these was President Jimmy Carter.

"We worship self-indulgence and consumption," Carter declared, taking his place in a great American tradition of social criticism. "Human identity is no longer defined by what one does but by what one owns." And, like earlier critics, Carter lamented the emptiness of such an existence. "We've discovered that owning things and consuming things does not satisfy our longing for meaning."

Carter saw the problem as residing in what he termed "a mistaken idea of freedom"—one in which we advocate "the right to grasp for ourselves some advantage over others." He called on Americans to unite together in a crusade of energy conservation:

> We often think of conservation only in terms of sacrifice… solutions to our energy crisis can also help us to conquer the crisis of spirit in our country. It can rekindle a sense of unity, our confidence in the future, and give our nation and all of us individually a new sense of purpose.[16]

This was his so-called "malaise" speech, and while it failed as an effort to transform the national spirit, and certainly failed Carter politically, it did capture well the link between environmental concerns and simple living that many Americans continue to feel today. Carter was followed by the Reagan and Bush administrations, during which no similar critique was heard. But now, at the turn of the millennium, there is renewed interest in simple living, if not in the White House, then at least in the heartland.

This quick historical survey reveals that "simple living" has meant many things. There is an anticonsumptionist core in much American thinking on this subject, but great diversity with respect to the human good and the place of work, religion, civic engagement, nature, literature, and the arts. Concern with simple living has been largely apolitical at some times, and at others the heart of a general political and social vision.

Today, when there is once again a great interest in simple living in America, it is mainly an apolitical enthusiasm. Most, though not all, of the literature is of a "how to" variety, offering advice on how to live more rewardingly with less money. The attainment of a simpler, more meaningful life is seen as an individual project, not as a matter of collective politics. This individualistic approach unfortunately has many limitations. It needs to be supplemented by a broader, more collective "politics of simplicity."

---

[1] Aristotle, *Politics*, trans. Ernest Barker (New York: Oxford University Press, 1961), p. 26.
[2] Ibid., p. 280.
[3] Ibid., p. 280.
[4] Ibid., p. 280.
[5] Ibid., p. 281.
[6] Ibid., p. 26.
[7] Contemporary economic thought, taken in formal terms, can accommodate almost anything. Thus, the distortion of personality can be viewed as "an externality" generated by market transactions, adding to the costs of every market interaction. But virtually no economists have expanded the idea of externalities to include distortions of personality. It remains a formal possibility, but consideration of such impacts, central to earlier eras, is largely outside of the way we think of the economic realm.
[8] This is wonderfully explicated in David Shi's study *The Simple Life?: Plain Thinking and High Living in American Culture* (Oxford: Oxford University Press, 1985), and I have drawn heavily upon Shi's account for this summary.
[9] John Woolman, "A Word of Remembrance and Caution to the Rich" in *Words That Made American History*, ed. Richard N. Current and John A. Garroty (Boston: Little, Brown and Co., 1965).
[10] "The Way to Wealth" in *Benjamin Franklin, Autobiography and Other Writings* (Cambridge, Mass.: The Riverside Press, 1958).
[11] Henry David Thoreau, *Walden* (Princeton: Princeton University Press, 1973), p. 4.
[12] Ibid., p. 6.
[13] Ibid., p. 11.
[14] Ibid., p. 14.
[15] Ibid., p. 8.
[16] Jimmy Carter, "Energy Problems: The Erosion of Confidence," *Vital Speeches* XLV (15 August 1979): 642, 643 as excerpted in David E. Shi, *In Search of the Simple Life* (Salt Lake City: Gibbs M. Smith, 1986).

## Sample Reading Selection B

*"There's No Place Like Work"*

***Arlie Russell Hochschild***

*(From* Hochschild, Arlie Russell. "There's No Place Like Work." *New York Times Magazine* 20 April 1997: 50-7.)

Nationwide, many working parents are struggling. More mothers of small children than ever now work outside the home. In 1993, 56 percent of women with children between 6 and 17 worked outside the home full-time year-round; 43 percent of women with children 6 and under did the same. Meanwhile, fathers of small children are not cutting back hours of work to help out at home. If anything, they have increased their hours at work. According to a 1993 national survey conducted by the families and Work Institute in New York, American men average 48.8 hours of work a week, and women 41.7 hours, including overtime and commuting.

Let's take look at a specific company, which I will call Amerco. Amerco has "family friendly" policies. If your division head and supervisor agree, you can work part-time, share a job with another worker, work some hours at home, take parental leave or use "flex time." But hardly anyone uses these policies. In seven years, only two Amerco fathers have taken formal parental leave. Fewer than 1 percent have taken advantage of the opportunity to work part-time. Of all such policies, only flex time—which rearranges but does not shorten work time—has had a significant number of takers (perhaps a third of working parents at Amerco.

Forgoing family friendly policies is not exclusive to Amerco workers. A 1991 study of 188 companies conducted by the Families and Work Institute found that while a majority offered part-time shifts, fewer than 5 percent of employees made use of them. Thirty-five percent offered "flex place"—work from home—and fewer than 3 percent of their employees took advantage of it. And an earlier Bureau of Labor Statistics survey asked workers whether they preferred a shorter workweek, a longer one, or their present schedule; 28 percent would have preferred longer hours. Fewer than 10 percent said they wanted a cut in hours.

To be sure, some parents have tried to shorten their work hours. Twenty-one percent of the nation's women voluntarily work part-time, as do 7 percent of men. But while working parents say they need more time at home, the main story of their lives does not center on a struggle to get it. Why? Given the hours parents are working these days, why aren't they taking advantage of an opportunity to reduce their time at work.

The most widely held explanation is that working parents cannot afford to work shorter hours. Certainly this is true for many. But if money is the whole explanation, why would it be that at places like Amerco, the best-paid employees—upper-level managers and professionals—were the least interested in part-time work or job sharing, while clerical workers who earned less were more interested? Similarly, if money were the answer, we would expect poorer new mothers to return to work more quickly after giving birth than rich mothers. But among working women nationwide, well-to-do new mothers are not much more likely to stay home after 13 weeks with a new baby than low-income new mothers,

A second explanation goes that workers don't dare ask for time off because they are afraid it would make them vulnerable to layoffs. With recent downsizings at many large corporations, and with well-paying, secure jobs being replaced by lower-paying, insecure ones, it occurred to me that perhaps employees are "working scared." But when I asked Amerco employees whether they worked long hours for fear of getting on a layoff list, virtually everyone said no.

Were workers uninformed about the company's family-friendly policies? No. Some even mentioned that they were proud to work for a company that offered such enlightened policies. The evidence, however counterintuitive, pointed to a paradox: workers at the company I studied weren't protesting the time bind. They were accommodating themselves to it.

Why? I did not anticipate the conclusion I found myself coming to: namely, that work has become a form of "home" and home has become "work." The worlds of home and work have not begun to blur, as the conventional wisdom goes, but to reverse places. We are used to thinking that home is where most people feel the most appreciated, the most truly "themselves," the most secure, the most relaxed. We are

used to thinking that work is where most people feel like "just a number" or "a cog in a machine." It is where they have to be "on," have to "act," where they are least secure and most harried.

But the new management techniques that are so pervasive in the corporate world have helped transform the workplace into a more appreciative, personal sort of social world. Meanwhile, at home the divorce rate has risen, and the emotional demands have become more baffling and complex. In addition to teething, tantrums, and the normal developments of growing children, the needs of elderly parents are creating more tasks for the modern family—as are the blending, unblending, reblending of new stepparents, stepchildren, ex-spouses, and former in-laws.

Current research suggests that, however hectic their lives, women who do paid work feel less depressed, think better of themselves, and are more satisfied than women who stay at home. One study reported that women who work outside the home feel more valued than housewives do. Meanwhile, work is where many women feel like "good mothers."

Many workers feel more confident they could "get the job done" at work than at home. One study found that only 59 percent of workers feel their "performance" in the family is "good or unusually good," while 86 percent rate their performance on the job this way. The reality is that, increasingly, Americans say they want more time with their families, but the truth is that they would rather be at the office.

---

### Sample Writing Assignment

With these reading selections by Jerome M. Segal and Arlie Russell Hochschild in mind, write an essay in which you discuss the role of work in a person's life. In your essay, summarize Segal's key points about the importance of work and money. Draw a relationship between Segal's thinking and what you have just read about American attitudes toward the workplace. In light of the reading selections, discuss your own knowledge or observations about the role of work in a person's life. Also discuss the degree to which your perspective reflects the ideas of either or both writers.

# Appendix D – Task 2: Analyzing and Integrating Information Text and Graphs

*The responses on the following pages, printed with permission, were written by CUNY students at an earlier CPE administration. The examination question they responded to is reprinted here:*

*The Education Gender Gap*

*The following article was recently published in a magazine about education.*

The gender bias against girls that educational specialists identified in schools in the early 1980s appears to have been eliminated. Unlike earlier indications that girls were performing consistently lower than boys and consistently lower than their abilities, statistics now show that females are the highest achievers at every level of education—from grade school through advanced University degrees. Girls' improvement is especially marked at the college level. Thirty years ago less than 40% of college graduating classes were comprised of women. Today, women earn an average 60% of all bachelor's degrees and 58% of all master's degrees. The United States Department of Education predicts these rates will continue to increase.

Yet, even as girls perform better, boys have begun to perform worse. Boys' reading and writing skills are a full $1\frac{1}{2}$ years behind their female classmates'. This gap is enough to put them at a profound disadvantage, since all other learning relies on these basic skills. As a result, boys are more than twice as likely as girls to be placed in a remedial education class. In addition, they are four times as likely to be diagnosed with a learning disability. It is not surprising, then, that twice as many boys drop out of school.

The new gender gap in academic achievement is real, and it threatens the future of millions of American boys. It took a concerted national effort to improve academic performance for girls; no less should be required for boys.

**FIGURE 1**

*The following chart was printed in a university bulletin.*

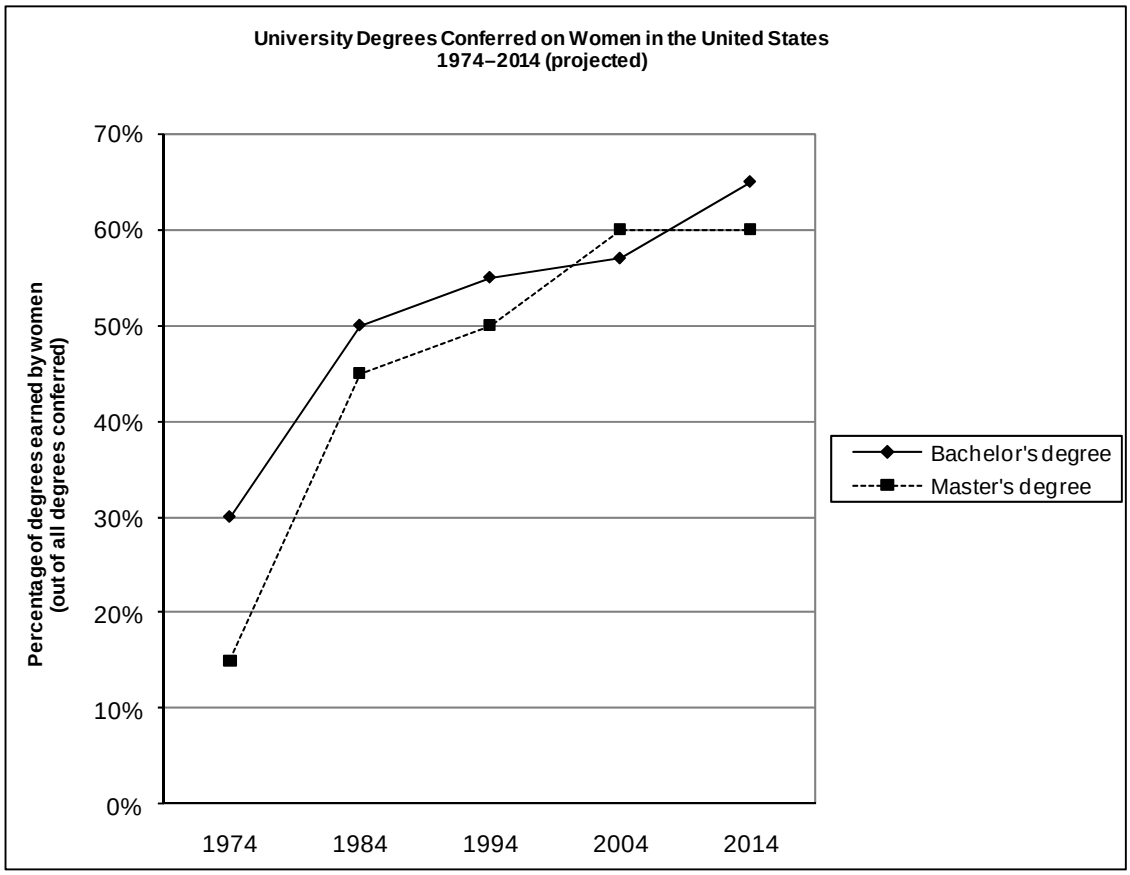
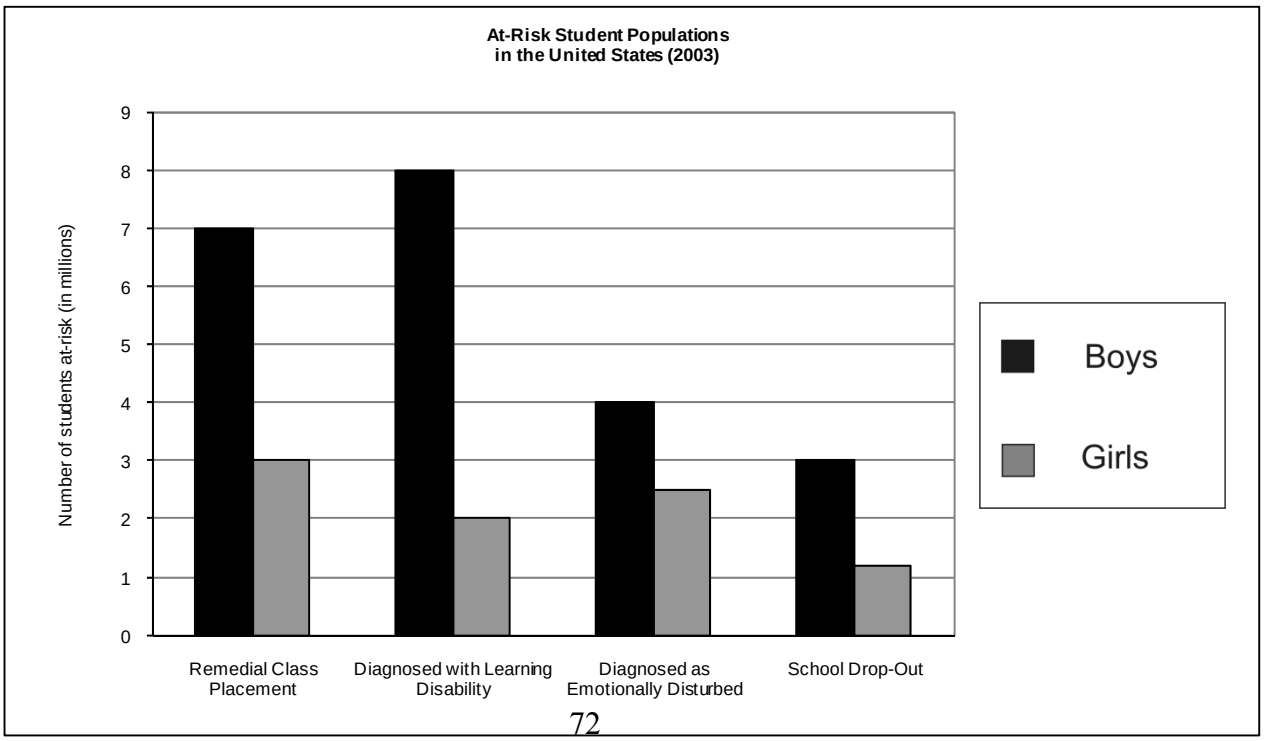



**FIGURE 2**

*The following graph was printed in school district report.*



# Appendix E – Task 1 Scoring Guide

| Organization | Critical Reading | Development of ideas | Language |
|---|---|---|---|
| **A.** Develops an essay that presents a focused response to the writing assignment, making appropriate and coherent connections among all parts of the assignment | **B**. Demonstrates understanding of the readings through summary and explanation of relevant material. | **C**. Incorporates, as support for own thoughts, references to the readings, identifying the sources formally or informally. | **D.** Communicates clearly and effectively, using appropriate conventions of language (e.g., grammar, spelling, punctuation). |
| **6** Addresses the writing assignment fully, analytically, and perhaps critically or imaginatively, with superior focus and coherence. | **6** Demonstrates superior and perhaps critical understanding of readings through accurate summary, full explanation, and insightful analysis of relevant sections. | **6** Makes insightful connections and distinctions between readings and own ideas; integrates references smoothly into own essay and identifies them consistently and correctly. | **6** Communicates with precision and enhanced expression through highly effective use of vocabulary and sentence variety; infrequent, if any, lapses in use of conventions. |
| **5** Addresses the writing assignment fully and analytically, with strong focus and coherence. | **5** Demonstrates strong understanding of readings through accurate summary, with appropriate explanation and analysis of relevant sections. | **5** Makes analytical connections and perhaps distinctions between readings and own ideas; integrates references into own essay and identifies them consistently and correctly. | **5** Communicates effectively throughout the essay, with few lapses in use of conventions. |
| **4** Addresses all parts of the writing assignment with adequate focus and coherence throughout. | **4** Demonstrates overall understanding of readings through appropriate summary and explanation, with some analysis. | **4** Makes and explains appropriate connections between readings and own ideas; identifies references consistently and correctly. | **4** Communicates clearly throughout the essay; sentences may contain some lapses in use of conventions, but these rarely impede comprehension. |
| **3** Addresses all or most parts of the writing assignment adequately, but focus may lapse briefly or connections may be missing. | **3** Demonstrates generally accurate understanding of readings although summary or explanation may be incomplete or not fully relevant. | **3** Makes some connections between readings and own ideas but they may not all be appropriate or adequately explained; identifies most references consistently and correctly. | **3** Generally communicates clearly throughout the essay although lapses in use of conventions may at times impede comprehension or prove distracting. |
| **2** Addresses some parts of the writing assignment **or** addresses all parts superficially; focus or coherence may break down at several points. | **2** Demonstrates partial understanding of the readings through summary or explanation, but understanding is flawed or explanation is incomplete. | **2** Makes few or unwarranted connections between readings and own ideas; may identify references inconsistently or incorrectly. | **2** Communicates clearly at times, showing some ability to use conventions, but whole sections are unclear or errors frequently impede comprehension. |
| **1** Shows little or no ability to address the writing assignment; may not link thoughts between paragraphs. | **1** Demonstrates little or no understanding of text. | **1** Makes no reference to background reading or makes no distinctions between background reading and own ideas. | **1** Communicates little because few sentences demonstrate appropriate use of conventions. |

# Appendix F – Task 2 Scoring Guide

## Score

**6** Accurately identifies two or more claims from the reading selection and explains the relationship of Figure 1 and 2 to these claims with accuracy, a high degree of complexity, and insight. Examinees can demonstrate a high degree of complexity or insight by successfully meeting the standards for a "5" AND by:

- introducing an additional claim and adequately discussing its relationship to one of the figures; or
- explaining an additional aspect of a figure's relationship to a claim; or
- discussing the relationship between one or more figures and the reading using perceptive analysis

**5** Accurately identifies two or more claims from the reading selection and explains the relationship of Figure 1 and Figure 2 to these claims with accuracy and a degree of complexity. Examinees can demonstrate a degree of complexity by successfully completing one of the following:

- introducing a third claim and adequately discussing its relationship to one of the figures
- explaining an additional aspect of a figure's relationship to a claim

**4** Accurately identifies two claims from the reading selection and adequately explains the relationship of Figure 1 to one of these claims and Figure 2 to the other.

**3** Accurately identifies two claims from the reading selection and adequately explains the relationship of one claim to one figure, but the connection between the other claim and a second figure is missing or inadequate (e.g., the connection is based on a misreading of the figure or simply repeats the language of the claim).

**2** Accurately identifies one claim from the reading selection and adequately establishes a relationship between that claim and one or both of the figures.

**1** Makes an attempt but does not accurately identify any of the claims or identifies one or more claims without establishing an adequate connection to either of the figures.

**0** Blank, completely off-topic, illegible, or written in a language other than English.

# Appendix G – Cost Analysis

This analysis presents the 2009 actual cost of administering the CPE, the projected 2010 expenditures under the new contract with Pearson, and the projected cost of ending the use of the CPE as a high-stakes certification exam and using it instead as a low-stakes assessment tool.

The 2009 actual expenditures for the CPE are detailed below in Table G-1, column 1. Development, copyrights, printing and delivery of the forms, and scoring were supplied by ACT. Campus and central university staff costs include the cost of faculty review of test prompts, administering the test at the campuses, workshops, CPE liaisons, and student support. The total cost was $3.34 million.

If the CPE continues to be conducted by Pearson, the total projected costs for the University will be approximately $4.9 million in 2010, assuming a 4% increase in the number of test takers due to rising enrollment (Table G-1, column 2). The first year of development costs are attributed to 2010, even though these expenditures are associated with the creation of new forms that would be put into service next year. Under the Pearson contract, the cost of development, production and scoring will approximately double, from $1.66 million to $3.20 million. Other cost components are projected to rise modestly due to inflation.

The low stakes alternative scenario, presented in Table G-1, column 3, makes the following assumptions:

1. The test would be taken by all students at the 45th credit, but would no longer be a degree requirement.
2. Only Task 1 would be administered.
3. CUNY would re-use CPE test forms and would not develop new ones.
4. Because the test would be low stakes, only one reader would be necessary and most of the expenditures for security and student support would be saved.
5. The test would continue to be scored by an external vendor.
6. Because the CPE is too long to be administered in most classes, we would need to continue to schedule test administrations outside of class.
7. Because Task 2 would be eliminated along with the 2nd reader, we project that scoring costs would decline by 60%.

The projected cost of this alternative is $1.55 million.

## Table G-1: Current and Projected Cost of the CPE

| | 2009 | Projected 2010 | Low Stakes Alternative (2010) |
|---|---|---|---|
| **Printing/Delivery of Materials** | $60,665 | $94,735 | $94,735 |
| **Copyright or Other** | $4,000 | $45,766 | $10,000 |
| **Test Development** | $170,000 | $1,076,067 | $0 |
| **Scoring** | | | |
| N tested | 39,195 | 40,763 | 40,763 |
| Cost per unit | $36.44 | $47.52 | $19.01 |
| Scoring cost | $1,240,828 | $1,937,048 | $774,901 |
| N appeals | 2,195 | 2,283 | 0 |
| Cost per unit | $72.87 | $15.63 | $0.00 |
| Appeals Scoring Cost | $139,748 | $35,679 | $0 |
| Reporting Test Scores | $46,962 | $14,043 | $5,617 |
| **Development, Production & Scoring Subtotal** | **$1,662,203** | **$3,203,338** | **$885,253** |
| **Staff** | | | |
| **Testing Full time Staff** | $515,036 | $527,912 | $316,747 |
| **Testing Hourly Staff** | $223,287 | $228,869 | $137,321 |
| **CPE Appeals Committee Staff** | $40,448 | $41,459 | $0 |
| **CPE Student Workshop Tutors** | $164,381 | $168,491 | $0 |
| **CPE Student Workshop Instructors** | $176,293 | $180,700 | $0 |
| **Writing/Learning Center Staff or other** (CPE Student Workshop related) | $104,039 | $106,640 | $0 |
| **Security Staff** (evenings and weekends) | $2,323 | $2,381 | $1,428 |
| **Maintenance Staff** (evening and weekends) | $378 | $387 | $232 |
| **A/V Staff** (set up in testing rooms) | $950 | $974 | $584 |
| **Mailroom Staff** process mail outgoing to students | $3,995 | $4,095 | $2,457 |
| **IT Staff** On-line registration set-up and maintenance, e-mail invites set-up and maintenance, batch program runs (103B) | $16,020 | $16,421 | $16,421 |
| **Other** | $10,100 | $10,353 | $6,212 |
| **Campus Total** | **$1,257,249** | **$1,288,680** | **$481,402** |
| **Central Office** | **$308,549** | **$316,263** | **$63,253** |
| **Staff Subtotal** | **$1,565,798** | **$1,604,943** | **$544,655** |

**Table continues next page**.

**Table G-1, continued**

| OTPS | 2009 | Projected 2010 | Low Stakes Alternative (2010) |
|---|---|---|---|
| **Supplies** | | | |
| Ink | $5,887 | $6,034 | $3,621 |
| Paper | $7,111 | $7,289 | $4,373 |
| Envelopes | $5,281 | $5,413 | $3,248 |
| Stock cards for postcards | $2,680 | $2,747 | $1,648 |
| **Mail (outgoing, to students)** | | | |
| letters | $33,301 | $34,133 | $20,480 |
| postcards | $3,207 | $3,287 | $1,972 |
| **Outside Vendor** | | | |
| Telephone tree for reminder notifications | $3,500 | $3,588 | $2,153 |
| **Facilities** | | | |
| Electricity on weekends (testing sessions and workshops) | $2,000 | $2,050 | $1,230 |
| Heat/AC on weekends (testing sessions and workshops) | $3,460 | $3,547 | $2,128 |
| **Other** | | | |
| e.g.. CPE Posters, Flyers, Attendance Cards | $13,265 | $13,597 | $8,158 |
| Incentives | | | |
| **Campus Total** | **$79,692** | **$81,684** | **$49,010** |
| **Central Office** | **$32,196** | **$33,001** | **$19,801** |
| **OTPS Subtotal** | **$111,888** | **$114,685** | **$117,821** |
| **Grand Total** | **$3,339,888** | **$4,922,966** | **$1,547,729** |

## Table G-2
## Estimated Annual Cost of Administering CPE, CLA, MAPP and CAAP to a Sample of 200 Freshmen and 200 Seniors per College

| | CPE | CLA | MAPP | CAAP |
|---|---|---|---|---|
| **Printing/Delivery of Materials** | $16,733 | $0 | $0 | $0 |
| **Copyright or Other** | $1,766 | $0 | $0 | $0 |
| **Test Development** | $0 | $0 | $0 | $0 |
| **Scoring** | | | | |
| N tested | 7,200 | 7,200 | 7,200 | 7,200 |
| Set up per college | | $6,500 | | |
| Cost per unit | $19.01 | $25.00 | $14.80 | $19.20 |
| Scoring of essay | | NA | $5.00 | $13.50 |
| Total cost of scoring | $136,872 | $207,000 | $142,560 | $235,440 |
| Reporting Test Scores | $2,247 | $0 | $0 | $0 |
| **Development, Production & Scoring Subtotal** | **$157,618** | **$207,000** | **$142,560** | **$235,440** |
| **Staff** | | | | |
| **Testing Full time Staff** | $55,948 | $55,948 | $55,948 | $55,948 |
| **Testing Hourly Staff** | $24,255 | $18,191 | $18,191 | $24,255 |
| **CPE Appeals Committee Staff** | $0 | $0 | $0 | $0 |
| **CPE Student Workshop Tutors** | $0 | $0 | $0 | $0 |
| **CPE Student Workshop Instructors** | $0 | $0 | $0 | $0 |
| **Writing/Learning Center Staff or other** (CPE Student Workshop related) | | $0 | $0 | $0 |
| **Security Staff** (evenings and weekends) | $0 | $0 | $0 | $0 |
| **Maintenance Staff** (evening and weekends) | $0 | $0 | $0 | $0 |
| **A/V Staff** (set up in testing rooms) | $0 | $1,800 | $1,800 | $0 |
| **Mailroom Staff** process mail outgoing to students | $1,000 | $1,000 | $1,000 | $1,000 |
| **IT Staff** On-line registration set-up and maintenance, e-mail invites set-up and maintenance, batch program runs | $16,421 | $16,421 | $16,421 | $16,421 |
| **Other** | $1,000 | $1,000 | $1,000 | $1,000 |
| **Campus Total** | **$98,623** | **$94,360** | **$94,360** | **$98,623** |
| **Central Office** | **$63,253** | **$63,253** | **$63,253** | **$63,253** |
| **Staff Subtotal** | **$161,876** | **$157,612** | **$157,612** | **$161,876** |

**Table continues on next page.**

**Table G-2, continued**

| OTPS | CPE | CLA | MAPP | CAAP |
|---|---|---|---|---|
| **Supplies** | | | | |
| Ink | $639 | $639 | $639 | $639 |
| Paper | $772 | $772 | $772 | $772 |
| Envelopes | $574 | $574 | $574 | $574 |
| Stock cards for postcards | $291 | $291 | $291 | $291 |
| **Mail (outgoing, to students)** | | | | |
| letters | $3,617 | $3,617 | $3,617 | $3,617 |
| postcards | $348 | $348 | $348 | $348 |
| **Other** | | | | |
| logistics to recruit students and encourage show | $36,000 | $36,000 | $36,000 | $36,000 |
| Incentives | $360,000 | $360,000 | $360,000 | $360,000 |
| **Campus Total** | **$42,243** | **$42,243** | **$84,485** | **$126,728** |
| **Central Office** | **$3,497** | **$3,497** | **$3,497** | **$3,497** |
| **OTPS Subtotal** | **$447,982** | **$447,982** | **$490,225** | **$532,467** |
| **Grand Total** | **$767,477** | **$812,594** | **$790,397** | **$929,783** |

*Technical Notes to Table G-2:*

- *Note that the CLA, CAAP, and MAPP have development costs rolled into the overall cost,whereas the CPE-Pearson figures separate these costs. For CLA, MAPP and CAAP, Reporting Test Scores, Print/Delivery/Materials, Form Development, Copyright or Other are included in the price of Scoring.*
- *For the Pearson-CPE, the cost per unit is $47.52 and re-scoring appeals is $15.63 per unit.*
- *For MAPP, it is assumed that students will take 2 or more objective modules online from the long version. Other test combinations are available. For both MAPP and CAAP it is assumed that students will take the optional essay.*
- *The CLA base cost is $6,500 for the first 100 students; additional student cost is $25 per student on-line only.*
- *The campus and central university staff and OTPS costs were estimated by pro-rating current staff costs associated with administering the CPE.*
- *Estimates for all exams assume payments of $50 per student.*