CAE

Returning to Learning in an Age of Assessment

Introducing the Rationale of the Collegiate Learning Assessment

August 18, 2009

Roger Benjamin
Marc Chun
Chaitra Hardison
Esther Hong
Christopher Jackson
Heather Kugelmass
Alex Nemeth
Richard Shavelson

TABLE OF CONTENTS

PREFACE

Roger Benjamin

I have two goals in offering this monograph. First, my colleagues and I present the logic of the Collegiate Learning Assessment (CLA) and second, to show how this logic relates to the core of education: teaching and learning.

The CLA performance assessments grow out of a rich paper-and-pencil testing paradigm we can only now exploit because of the Internet. Performance assessments are authentic problems or simulations of real world issues that are comprised of an assortment of documents, such as tables, figures, graphs, newspaper reports and photographs. The Internet permits richer versions of the tasks to be developed and on-line delivery permits the performance assessments to be administered, scored, analyzed and reported to the students and their institutions more quickly, accurately, and inexpensively. The scoring of the performance tasks provide rich diagnostic information about the students' writing, analysis, and critical thinking skills.

The CLA is designed to permit comparisons within and between institutions to engage faculty and administrators in their efforts to improve the quality of teaching and learning. Without comparative benchmarks of student learning outcomes how do faculty and administrators at an institution benchmark the efficacy of their undergraduate programs? The counterintuitive element is that the CLA is not designed to provide these comparisons for strictly accountability purposes. Rather, CLA-based evidence to date shows that the differences in student learning outcomes between similar institutions suggests an abundance of best practices available for faculty at particular institutions to examine at counterpart schools.

Given these two factors, this monograph will explain how the CLA can help faculty improve teaching and learning. The board members of the Council for Aid to Education (CAE) – the institutional home of the CLA – did not come together to advance the accountability movement; rather, they saw the promise of the performance assessment paradigm as a means to improve education. Unlike standardized multiple choice tests, performance assessments are *tests worth teaching to*. Our fundamental goal, then, is to introduce the performance assessment paradigm to faculty throughout the higher education sector.

1. INTRODUCTION[1]

Roger Benjamin

BACKGROUND AND CONTEXT

This monograph spells out the logic of the Collegiate Learning Assessment's (CLA) role in postsecondary education. The principal goal of the CLA is to assist faculty in improving teaching and learning. Because arguments for and against assessment instruments often revolve around psychometric issues of reliability and validity, or larger policy debates about accountability, the question of the direct utility of assessment instruments for faculty in the classroom has remained marginal at best. Certainly, those associated with the CLA also spend considerable time thinking and writing about methodological and technical questions associated with the CLA[2]. Here, we place the question of the relevance of the CLA front and center to indicate how it is being (or can be) used to improve teaching and learning. We argue that the CLA strategy discussed here is necessary, if not sufficient, for meaningful progress to occur in assessment in higher education.

Summative vs. Formative Assessment

Summative assessment usually plays the role of determining what knowledge or skills students have attained at the individual, program or institutional level; assessment is separate from the learning process. In comparison, formative assessment is typically viewed as part of the teaching and learning process. (It may be useful to view formative and summative assessment as a continuum with all forms of assessment possessing some formative qualities.) In the case of the CLA, the performance tasks are designed to do both: to fulfill the traditional role of summative assessment for the institution, but more importantly, to also be used for formative purposes by faculty.

It is important to develop an integrated strategy that combines formative and summative dimensions in order to make sustained progress toward a more systematic approach to improve teaching and learning. This approach recognizes that faculty are the ultimate stakeholder of

[1] This chapter was prepared originally as the introduction to a forthcoming monograph comprised by a set of papers (many initially published in peer reviewed publications) on the aims, methods, uses and policy implications of the CLA. These papers can be consulted for detail on methodological and technical issues.
[2] See reference section at the end of this monograph.

assessment. Unless there is formative value, faculty will not take any assessment seriously. However, a point widely ignored or misunderstood is that **appropriate summative assessment is actually necessary in order to give faculty and administrators information they need to help frame a well grounded formative assessment program.**

Preferred assessment approaches should accomplish three things: first, satisfy the highest standards of reliability and validity so results can be meaningfully compared; second, be engaging to the student, constituting a significant learning as well as an assessment opportunity; and third, be recognized as authentic by faculty by assessing skills seen as critical for college graduates to master. Most existing formative and summative assessments typically meet one or two of these criteria, but rarely meet all three of them. The goal of the CLA is to satisfy all three criteria, because only in that way will these criteria be perceived as being of critical importance to faculty in higher education.

Evaluation Criteria For Assessment Instruments

To satisfy the needs of the formative component of an assessment program, the answers to three key questions become central.[3] First, **what components of student learning outcomes should be focused on?** Higher order skills (critical thinking, analytic reasoning, problem solving and written communication) are a promising focus because of a number of reasons that have been corroborated in five years of national testing with the CLA: (a) these skills have been and are emphasized in liberal arts or what is called common learning; (b) higher order skills are thought to be critical in the knowledge economy; and (c) these same higher order skills are reflected as core to virtually all mission statements of colleges and universities.

Second, **should comparative-based (summative) assessment at the institution-level be the initial focus?** The case for comparison is itself strong when there is little or no consensus about whether existing theories provide minimum guidance for research and development in a field.[4] If we had consensus about how to improve teaching and learning, we would not need such comparisons. That is not where we find ourselves in the field of teaching and learning.

[3] In other words, the creators of the CLA, like any group attempting to contribute to the field of learning outcomes in higher education, went through a number of decisions that make up a set of responses to the need for any assessment protocol to be reliable and valid, fair to all takers, respectful of time and resource limits for testing, and understand that decisions about what to test and how to test influences curriculum and instruction (Klein, 2002).
[4] Ragin, 1989.

However, some standardized comparative-based assessment instruments should be used to answer the question, "Is what the institution doing good enough?"[5] In other words, the right kind of comparative assessments (the ones that meet all three criteria for evaluation of assessments) present important signaling tools and checks for faculty and administrators to use as guidance at their own institutions.

Why focus initially on the institution-level of analysis? Higher-order skills are cumulative and not the province of individual courses or majors. And the whole of the institution should be greater than the sum of its departmental parts. Does comparison between institutions yield important information? The answer is yes with at least one assessment instrument, the CLA; up to at least two standard deviations of differences between similarly situated institutions on CLA value-added outcomes occurs. This means there is a large playing field of best practices to investigate by faculty to consider using to improve their teaching and learning (Benjamin, 2008).

Third, **how should the selected outcomes be measured?** Three plausible choices fall short:

- *College grades* are subject to inflation, which means there is too little discrimination among students. Additionally, grades are not used on the same scale across institutions (Klein et al., 2005).
- *Tests* given by academic departments may seem attractive; however, attempts to develop measures of the core outcomes of majors have proven problematic. Too few majors have agreed-upon definitions about what is central to them. And there are so many majors the question of which ones to consider central becomes controversial.

[5] See Graff and Birkenstein (2008) for a defense of standardization. They point out that many understandably equate all standardized tests with the highly questionable multiple choice tests that characterize No Child Left Behind. Other points they make include 1) the position that colleges are too diverse to be measured by any common standard can be turned into the argument that the basic skills students need to succeed in college are too complex and heterogeneous to be clearly explained, and 2) attacks on all educational standardization reinforce the current fragmented, non cumulative, disconnected undergraduate curriculum which masks the underlying importance of critical think skills that "…underlie effective academic work in any course or discipline and that it is not correct to hold that colleges are so diverse that they do not share common standards. Virtually all faculty agree that critical thinking skills are essential for graduating undergraduate students to possess.

- *Portfolios* [6] Many faculty propose using "portfolios" of student work, including term papers, descriptions of experiments they conducted or collaboration with others in performing a community service, extra curricular activities in which they engaged. Some portfolio programs make an effort to be structured, such as by specifying the number and types of products students are allowed to include in their portfolios. Other portfolio programs allow students to include almost anything they believe illustrates their accomplishments and growth, including videos of themselves.

  Although portfolios are a seemingly "authentic" means of presenting student work, they lack the critical feature of standardization that is essential to have in a measurement instrument. For instance, how can evaluators appropriately assign scores to student term papers within and between schools when the assignments are procedures were different, such as being or not being allowed to get help from other students or faculty? The same goes when the evaluation of a student's work includes self-evaluations that may or may not have been coached by others. Thus, what may seem to be "authentic" may just be a sham.

  The only feasible way to assign fair and valid scores to student work is to require that all students produce that work under the same conditions and instructions. In short, the assessment must be truly standardized and the work products in the portfolio must be created without the direct assistance of faculty, classmates, or others. Only in that way can we assess what the student is capable of doing.

  There are, of course, other problems with portfolios that preclude their being used effectively in large scale assessment programs. Evaluators must compare disparate work products and record their assessments of the quality of these products on a common scale. This is analogous to the "best in show" portion of a god show where the judge compares each "contestant" to the ideal characteristics for that contestant's breed, the weight to assign to each of those characteristics, and then somehow takes all of these evaluations and weights into considerations in deciding which one of the 7 contestants is "best." Its statistical voo doo with chutzpah.

[6] The points that follow regarding portfolios were contributed by my colleague Steve Klein.

In comparison to these alternatives, performance assessments have shown promise of meeting the three essential criteria for evaluating assessments that can be effective in higher education. This does not mean we have completed the journey of adapting the performance assessment paradigm to undergraduate education. Performance assessment should be extended to majors and perhaps to additional skills such as perspective taking, among many other possibilities.[7]

The Relevance of the Peer Review Analogy for Reporting Scores

Higher education produces at least three principal public goods; research, undergraduate education and service. For research, the higher education community long ago adopted the principle that research, research grants and potential publications should be governed by peer review (Bush, 1945). By peer review is meant anonymous critiques of research proposals, draft articles or book manuscripts submitted for publication consideration by journals and book publishers. Some scholarly associations also demand requests to give papers on panels at annual meetings to be peer reviewed as well. There are strict rules governing this process. The author's name is typically omitted from the draft article submitted for publication consideration. In turn, the reviewers' names remain anonymous. The rationale for this process is to maintain as much objectivity in the review process as possible. Both the submitting authors and reviewers need to know that their arguments, ideas, or critiques will be protected from public scrutiny. Otherwise, there is no incentive for a researcher to submit drafts for publication because they might be embarrassed by the content of the critique. Similarly, the reviewers do not want their identity to be made public. The peer review process, at its best, thus generates honest appraisal of research that benefits the authors even if their paper is not accepted for publication. Peer review makes research a process of continuous improvement. Further, research standing is documented by the process of peer review; no professor, department or institution can simply assert that they are the best researchers in the region, state or nation.

Comparative-based standardized assessments, if they are the right kind (judged to be effective strategies by the tradeoffs they employ based on the above noted criteria of evaluation of assessments), should be viewed through the lens of the peer review concept. Both peer review

[7] With support from the Carnegie Corporation CAE has recently established an Institute for Performance Assessment to be directed by Richard Shavelson. The Institute will review existing applications of the performance assessment testing paradigm and propose others. The research agenda of the Institute will be developed with assistance from an advisory board of leading measurement scientists.

and comparative-based assessment attempt to provide an answer to the question, "Is it good enough?" As in the case of peer review of research, comparative-based assessments should be designed to be objective characterizations of an institution, offering evidence that is judged to be reliable and valid, and in particular, to have strong face validity (thought authentic by the faculty). Such assessment that meets the measurement science requirement of minimum standards of reliability and validity offers a powerful reality check for institution-based formative assessment efforts by asking the question, "Is what the institution doing to improve its teaching and learning good enough?" The fundamental point, by analogy, is that the organization that does the comparative-based testing should not make the results public. The testing organization should report assessment results for the institutions it tests back to them only. Otherwise, why would an institution, department or program want to permit comparative-based testing which, we argue, plays a critical role in making formative assessment more systematic? Of course, the question of what to report and who gets to report it under what review conditions are central issues. The peer-review analogy suggests our position.[8]

The peer-review analogy also suggests other food for thought. Individual authors do create resumes that document their publications that, as a corpus, provide a strong indicator of the scope and level of attainment of their research career. Departments, programs, colleges and universities aggregate publication and related evidence of research and make such evidentiary bases public. Such evidence is given credibility because one can judge how much of it is based on peer review.

Let me be as clear as possible, then, about the rules that govern reporting of the initial institutional CLA scores. They are and will only be reported back to the institution that contracted with CAE to take the CLA and the students who take the CLA tests. These rules have been promulgated by CAE's Board of Trustees that is largely made up of distinguished current and past presidents of colleges and universities.[9] To do anything else would cripple the usefulness of comparative-based tests to provide useful information to individual colleges and universities and would result in distorted results being made public.

[8] Cf Benjamin and Klein (2007) for a more detailed discussion of the relationship between assessment and accountability.

[9] See CAE Board Statement on assessment presented to the U.S. Department of Education Spellings Commission (2006). See document at www.cae.org.

Implications for Assessment and Accountability in Higher Education

The kind of assessment evidence discussed here may be used in accountability systems agreed to by the institution. But the individual colleges and universities should play an important role in deciding whether and what to make public about its assessments, both comparative and local based.[10] National and state organizations have recently launched pilot experiments such as the Voluntary System of Accountability (VSA) (Association of Public and Land Grant Universities (APLU)[11] and the American Association of State Colleges and Universities(AASCU)) and the University of Texas system of accountability. The Council of Independent Colleges has launched a consortium of colleges that is making progress in figuring out how to use the CLA results to good effect to improve teaching and learning. Other national organizations such as the Association of American Colleges and Universities are focused on promoting a broad set of liberal arts goals and measuring them with portfolios. The Council for Higher Education Accreditation and most of the regional accrediting organizations have or are moving to requiring evidence of student learning improvement. And Robert Connor, president of the Teagle Foundation, with other leading higher education leaders, has created an Alliance to focus attention on the need to systematically improve teaching and learning, with assessment as an essential ingredient.

A Key Design Feature: Respect for Faculty Autonomy

The CLA performance tasks measure the higher order skills noted above. However, unlike many tailored assessments that "have the answers to assessment they attempt to recruit faculty to accept," the CLA program does not proscribe any particular teaching and learning approach for faculty to follow. We recognize the importance of faculty autonomy in their role as educators. In higher education authority we delegate to departments, programs, and, of course, faculty broad discretion to decide what to teach, how to teach, and how to assess. The CLA program overall, and CLA Education Services, in particular (see chapter 6), is built on the assumption that a successful assessment program in higher education needs to be designed to not only recognize the importance of faculty autonomy in teaching and learning but create strategies that encourage

[10] See Benjamin and Klein (2007).
[11] Formerly the National Association of State Universities and Land Grant Colleges (NASULGC).

faculty to develop problem and case-based assessments appropriate to their fields of expertise. In our view, there is a virtually unlimited lode of faculty driven ideas to give content to the shift underway in undergraduate education to student centered teaching; problem and case oriented curricula and text materials, and open-ended assessments which the CLA performance assessments are designed to provide models for.

From the point of view of the CLA it should be a time of experimentation that places the focus on improving teaching and learning. However, experimentation should be judged by how well the assessment strategy stacks up against the criteria for evaluation of assessments presented here. At a minimum, that means an either/or strategy; only formative or only summative tests should be seen as overly simplistic. As I recently wrote,

> We should apply to any proposed strategies for assessment and accountability the same logic by which Vanevar Bush developed the peer review research policy that respects the diversity of institutions and American higher education's institutional design, with its decentralized governance structure and respect for faculty autonomy. But we need to engage the kind of layered comparison strategy described here to improving the quality of teaching and learning in higher education.[12] If we did not have the ability to do the kind of comparative assessment described here, the status quo might be acceptable. But we now can do sophisticated, meaningful, appropriate comparisons – and, since we can, there is no legitimate argument against doing so (Benjamin, 2008, p. 55).

The argument that some critics make that any comparisons between institutions is inappropriate is not credible. In fact, the reverse is inappropriate. Without appropriate comparisons that serve to link individual campus efforts, current formative assessment strategies are doomed to remain isolated silo activities at colleges and universities that, in the end, cannot be built upon to create the more systematic approach to student learning improvement many in the academy seek to achieve.

This is a time for serious innovation in assessment and accountability, and we should encourage fresh efforts to do so while subjecting them to rigorous scrutiny which, above all, means employing the standards of reliability and validity developed by assessment scientists as well as the critiques of faculty who ultimately decide whether to assess and how to do it.

[12] By which I mean developing and using assessment instruments such as performance assessments that can be used for both summative and formative purposes.

ORGANIZATION OF THIS MONOGRAPH

This monograph is divided into six chapters[13], which together explicate the logic of the CLA for assisting faculty in using the CLA in what is designed to become a continuous system of improvement. The chapters draw on previously published papers, with slightly adapted and updated versions brought together here. Readers are encouraged to review the full papers for more detail.

Chapter 2, a paper by Benjamin, Chun and Jackson (2009), places the CLA in the context of the assessment and accountability movements in higher education. The chapter discusses how a focus on accountability overlooks the most important parts of the CLA, which is to introduce performance assessment to the academy. The chapter introduces the full scope of services the CLA program provides, most importantly to faculty in the classroom.

Chapter 3, which draws on a white paper by Benjamin, Chun and Shavelson (2007) lays out the chain of diagnostic logic from the initial step of assessment at the institutional level to the classroom and back to the institution. The chapter argues for a more systematic set of steps to improve teaching and learning.

Chapter 4, a paper by Benjamin (2009) details why the faculty role is central in assessment in higher education and must be the basic building block for any meaningful accountability system as well. If the CLA performance tasks are tests worth teaching to, what are the diagnostic properties of the tasks?

Chapter 5 describes the performance tasks in detail. This chapter, based on a document prepared by Hardison, Hong, Chun, Kugelmass and Nemeth (2009), lays out the architecture of the CLA tasks: their design and scoring criteria, and actual student responses at low to high ranges. (Note: under its charter, CAE does not produce national assessments of the progress of student learning in higher education. However, a recently published monograph provides a methodology for individual colleges and universities to set their own minimum cut-off points for CLA performance or hoped for maximum performances(Hardison and Vilamovska (2009).

Chapter 6 concludes this monograph with a description and discussion of how we move the assessment work to the classroom. This chapter by Chun introduces CLA Education, including the Performance Task Academies, the Performance Task Library and two new reports (the Student Diagnostic Report and the Institutional Diagnostic Report).

[13] This introduction is the first chapter.

Section I

The Logic of the CLA's Assessment Role

2. THE COLLEGIATE LEARNING ASSESSMENT'S PLACE IN THE NEW ASSESSMENT AND ACCOUNTABILITY SPACE[14]

Roger Benjamin, Marc Chun and Christopher Jackson

INTRODUCTION

Over the past five years, the higher education community has engaged in a national conversation that has focused on if and how to measure student learning. This is particularly noteworthy given that American higher education typically acts as fifty separate state systems and not as a unified whole; one must take notice when there is this level of collective discussion. Arguably, the assessment and accountability space has been permanently changed. The confluence of a number of factors helps to explain what has happened: (1) the entrance of new organizations actively supporting or undertaking assessment (e.g., the Teagle and Lumina Foundations and the Collegiate Learning Assessment); (2) the actions of the federal government (e.g., the Spellings Commission, and indirectly, No Child Left Behind); and (3) new initiatives put forth by national organizations (e.g., the Voluntary System of Accountability (VSA), sponsored by the Association of Public and Land-grant Universities (APLU) and the American Association of State Colleges and Universities (AASCU)).

However, given the simultaneity of these factors and the way that the assessment and accountability space has been altered over such a short span of time, it is not surprising for there to be confusion and a conflation of these conceptually connected yet nevertheless separate dimensions. It is the purpose of this chapter to help clarify the nature of the CLA, and to detail how it is separate from these other initiatives (and in particular the VSA).

What is the CLA?

The Collegiate Learning Assessment (CLA) was founded by the Council for Aid to Education in 2000 (then a subsidiary of the RAND Corporation) with one goal: to improve teaching and learning. The CLA experienced a rapid rise to the national scene. After two years of intensive development, the 2002-03 academic year pilot testing of the CLA measures and analytical approach proved so successful and the demand from institutions so immediate, that the

[14] Excerpted and adapted from Benjamin, Chun and Jackson (2009) The Collegiate Learning Assessment's Place in the New Assessment and Accountability Space.

program launched officially in the fall of 2004. Although the student testing component of the CLA is likely the most recognizable, the CLA actually includes four inter-related programs that address dimensions of (1) measurement science (CLA Testing); (2) consideration of institutional- and student-level factors that may correlate with measured performance (CLA Analysis); (3) curriculum and pedagogy (CLA Education); and (4) more general social science empirical work (CLA Research).

CLA Testing includes the use of our distinctive open-ended performance-based measures to assess students' skills in critical thinking, analytic reasoning, problem solving and written communication. Cohorts of entering and exiting students participate, and a matrix sampling approaches used to conduct cross-sectional or longitudinal analyses. Either approach provides information about the college's or university's overall value added (the student learning gains made at the institution after controlling for initial ability, and including comparisons to similarly situated institutions). Although many colleges and universities choose to use the minimum sample sizes that permit such institutional analysis, the CLA encourages institutions to do additional "in-depth" sampling to allow analyses of colleges within a larger university, comparisons of different groups of students, or consideration of other factors that may be of interest. (CLA Testing features include Cross-Sectional Sampling, Longitudinal Sampling, In-Depth Sampling and the CLA Institutional Report.)

CLA Analysis moves beyond this summative analysis, and encourages colleges and universities to use the student-level data to do formative analysis by investigating correlations of performance with institutional factors or locally collected assessments. In other words, institutions can drill down to understand the contributing factors that explain their institutional and sub-institutional scores. (CLA Analysis features include the CLA Student Level Data File and the CLA Local Survey; CLA Student Diagnostic Reports and CLA Institutional Diagnostic Reports are in development.)

CLA Education focuses on curriculum and pedagogy, and embraces the crucial role that faculty play in the process of assessment. As performance-based measures become increasingly recognized as the most authentic way of teaching the skills that institutions value in their students and that employers demand of college graduates, the CLA has worked to facilitate the connection of institution-wide results to classroom-level and faculty work on student learning.

The flagship program in this area is CLA in the Classroom[15]; the CLA in the Classroom Performance Task Academies provide faculty members with tools for creating and scoring their own content-embedded performance measures. No other assessment program has a comparable companion program. CLA takes seriously the informal mantra that we should create tests worth teaching to. If we believe students should be able to demonstrate higher order skills through performance tasks, we should not only assess such skills, but also help to create opportunities for students to develop and practice these skills. Other CLA educational efforts include informal gatherings at national meetings (Coffee [cla]tches), as well as free web conferences and newsletters to facilitate the sharing of best practices (CLA Spotlight and the CLA Pulse). Consortia[16] of institutions also provide means to share best practices. (CLA Education programs include CLA in the Classroom, CLA in the Classroom Performance Task Library, CLA Consortia, CLA Spotlight, CLA Pulse.) A new website also has been formed to host performance tasks developed by faculty. This website will be a clearing house for faculty to share performance tasks and also alert CAE measurement scientists to tasks that can be the basis for new CLA measures.

CLA Research includes empirical studies conducted by the CLA research staff to consider larger psychometric, sociological, political and economic analyses of the data, typically looking across institutions participating in the CLA. A current collaboration with the Social Science Research Council (SSRC) illustrates the type of independent research CLA encourages. White papers and other research reports can be found on the CLA website.[17]

Taken together, it is clear that the goal of the CLA is to promote the improvement of teaching and learning through this full range of programmatic offerings. Formative assessment—not accountability—is the raison d'etre of the CLA. The entire CLA consists of a number of programs that, taken together, form a continuous system of teaching and learning improvement.

[15] For more information, visit www.claintheclassroom.org.
[16] In 2002, the Council of Independent Colleges (CIC) with CAE jointly established the first CLA consortium of private liberal arts colleges. These 47 institutions share best practices used to administer the CLA as well as improve curriculum and pedagogy based on their CLA scores.
[17] Visit www.cae.org/cla.

THE PLACE OF THE CLA IN THE ACCOUNTABILITY SPACE

As suggested above, there has been much confusion about the CLA, particularly with perceived connections to policy efforts. Some critics have claimed incorrectly that the CLA is a product of the federal government, no doubt given the endorsement by the Spellings Commission. Although we appreciate that the US Department of Education sees value in the work we have undertaken, there has been no direct nor indirect connection between the CLA and any government agency. Also, some have mistakenly suggested that the CLA is a higher education version of No Child Left Behind. In fact, we have explicitly noted how every key feature of the CLA stands diametrically opposed to the corresponding features in NCLB.[18]

But perhaps the most immediate challenge has been the confusion with the VSA. The VSA[19] (and public reporting through the common web template the College Portrait) is a program that, as stated on their website, was designed "to improve public understanding of how public colleges and universities operate." The website notes, "The College Portrait provides consistent, comparable and transparent information on the characteristics of institutions and students, cost of attendance, student engagement with the learning process, and core educational outcomes. The information is intended for students, families, policy-makers, campus faculty and staff, the general public, and other higher education stakeholders." The College Portrait requires institutions to share information in three areas: (1) consumer information, (2) student experiences and perceptions, and (3) student learning outcomes. For the student learning outcomes section, VSA recommends that institutions use one of three assessment programs: ETS' Measure of Academic Proficiency and Progress (MAPP), ACT's Collegiate Assessment of Academic Proficiency (CAAP), or the CLA.

We support the goal of the VSA to assist higher education institutions in providing greater transparency and recognize that higher education constituent groups and accrediting agencies are increasingly requiring institutions to provide evidence of student learning outcomes. However, CLA's participation in the VSA should not be taken to mean that the CLA is primarily an accountability tool. There are three key distinctions between the CLA and VSA that are most salient: program purpose, stage of development, and focus on faculty. Each will be discussed below.

[18] See Klein (2008) How the CLA Differs from No Child Left Behind.
[19] For more information about the VSA, visit www.voluntarysystem.org.

Difference #1: Program purpose

The first difference is the overall purpose. The VSA by design focuses on assessment for accountability[20]; the CLA, by contrast, has always been driven by a commitment to assessment for improvement[21]. These are not wholly incompatible, but they are nevertheless distinct.

It is important to note here that the VSA utilizes just one of the components of our program: CLA Testing. Furthermore, VSA requires only institution-level results; the in-depth sampling and sub-group analyses encouraged by CLA are not part of the VSA's recommended protocol. Moreover, participating in CLA in the Classroom, the analytical research or the other policy work of the CLA goes above and beyond the VSA. This is not to say that the VSA should incorporate those elements into the College Portrait. The mission of their program is quite specific. However, when the work of CLA and VSA get mistakenly conflated, it may be easy to overlook the fact that the CLA involves far more than just institution-based assessment.

While institution-based value-added scores may be used to meet accountability requirements, it is important to recognize that separating out the institution-based CLA score can undermine the vision of the CLA as a whole. These scores are designed to be only a signal to faculty and administrators about where their institution stands compared to similarly situated institutions. Indeed, if institutions do not commit to understanding what leads to their institution-level scores (as suggested by the other features of CLA Testing, as well as CLA Education and CLA Analysis), they will not reap the full benefits of what the CLA offers.

Difference #2: Stage of Development

A second difference between the CLA and the VSA is their respective stages of development. The CLA has already been established as a reliable and valid measure and analytical approach, and is in its fifth year of widespread use throughout the United States and abroad.[22] The student learning outcomes component of the VSA's College Portrait allows institutions to have their own four-year pilot phase. It is the institutions that are doing piloting

[20] The VSA's rationale is to assist institutional efforts to get out ahead of potential accountability mandates. The primary stated objectives of the VSA are to help institutions: (1) demonstrate accountability and stewardship to the public, (2) measure educational outcomes to identify effective educational practices and (3) assemble information that is accessible, understandable and comparable.

[21] See Benjamin and Klein (2006).

[22] For example see and access the numerous technical publications on methodological issues on the CAE website.

testing as they figure out how to use the tools to assess their student learning. It is not the case that the CLA is being tested. VSA's strategy is laudable, in that it allows and encourages institutions to test out the different measures as they prepare to report out publicly such sensitive data for the first time. So although the VSA allows institutions to engage in their own pilot phase, it is important to recognize that the CLA itself is well beyond the pilot phase.[23]

Difference #3: Role of Faculty

A third major difference between the CLA and VSA is the role of faculty. Again, the CLA was created to improve teaching and learning. Notably, CLA in the Classroom was designed to address the specific needs of faculty, and to ensure that they are central to the assessment process. CLA in the Classroom provides professional development to aid faculty in improving teaching and learning in their own classrooms, and has already met with great success in helping to improve teaching practice and encouraging student skill development. The VSA's efforts to make information available to multiple constituencies through public reporting is laudable but there is no direct role for faculty in their program. We believe that faculty must be central to the process in order to ensure the greatest sustainable success in systemic and systematic improvement of student learning.

CONCLUSION

The focus on institutional level scores, as encouraged by the VSA, can unwittingly obscure the complexity of the teaching and learning enterprise. We trust that the creators of the VSA understand that institution-based value-added scores are only one indicator of success in student learning just as graduation rates only partially indicate the success level of a college's undergraduate education program. In the case of the CLA, we offer a comprehensive set of programmatic offerings to improve practice and understanding that will in turn promote better teaching and learning.

[23] MAPP and CAAP are engaging in their own experimentation as they attempt to duplicate some form of the CLA-designed value-added protocol.

We, all those involved in higher education, must acknowledge that we are only at the start of the long road ahead if we are to succeed in getting our colleagues to focus on the systematic steps needed for a continuous improvement in teaching and learning.

3. HOLISTIC TESTS IN A SUB-SCORE WORLD: THE DIAGNOSTIC LOGIC OF THE COLLEGIATE LEARNING ASSESSMENT[24]

Roger Benjamin, Richard Shavelson, and Marc Chun

Some complex tasks easily lend themselves to be separated into constituent parts. The production of automobiles was improved with the division of labor and the introduction of the assembly line, where there was one autonomous unit working on building engines, another working on tires, and another installing windows. Arguably, separating out these tasks enabled managers to better control the production process, and understand where inefficiencies occurred. Mechanisms to measure the performance of each separate part enabled the whole assembly line to produce cars faster and cheaper.

It is easy to see how tempting it would be to apply the same logic to the assessment of learning in higher education. Given the explosion of knowledge, a similar division of labor and approach to measurement are reasonable-sounding strategies – and in many cases this does in fact make sense. One prominent example is specialization by academic discipline, where experts teach what they know best; no one faculty member could or should be expected to know all there is to know about all subject areas. And within a particular classroom, tests are administered to determine the sub-areas where students have mastered material. A well-designed chemistry test, for example, divided into topic areas (say) could help a faculty member determine if students had mastered the rather different and distinct tasks of understanding how to read the periodic table vs. how to titrate chemicals vs. how chemical bonds work. Students' abilities to integrate these separate skills into an understanding of the nature of the physical universe is also valued, and could also be assessed accordingly.

Similar attempts have been made to divide "higher-order" skills (i.e., those that faculty as a whole are responsible for developing in students) into component parts such as critical thinking, analytic reasoning, problem solving and written communication. We argue that although such attempts have been made and sub-scores reported, such efforts might be misguided. Using the CLA as an example, we suggest that there is another way of looking at the diagnostic process from a holistic perspective and we seek to show how, in the end, such a process might have more salutary effects on teaching and learning that the traditional componential approach.

[24] Adapted and excerpted from Benjamin, Chun and Shavelson (2007).

IMPORTANCE OF HIGHER-ORDER LEARNING

We believe there are three key rationales for focusing on the development of higher-order skills. First, recent theories of learning stress the importance of improving students' ability to structure their own learning experiences that help them use what they have learned in new settings. Simon (1996) argues that the meaning of "knowing" has changed from being able to recall information to being able to find and use it. Bransford et al. (2000, p. 6) note that the "… sheer magnitude of human knowledge renders its coverage by education an impossibility; rather, the goal is conceived as helping students develop the intellectual tools and learning strategies needed to acquire the knowledge to think productively." Under these conditions the proponents of the new learning theories argue that active learning and assessment of learning become critical because students must learn to recognize when they understand a subject and when they need more information (NRC, 2001).

Second, most college mission statements reference the need to improve higher-order skills. The need to focus on these skills is supported not only by recent national movements – such as the Greater Expectations program of the American Association of Colleges & Universities (AAC&U, 2002) – but also by the general public as well as parents (Immerwahr, 2000). What makes the reform agenda urgent is the growing realization that we are in the midst of a new phase of social and economic development, often dubbed the knowledge economy which makes strengthened higher-order skills essential. In the knowledge economy there is a shift away from the Industrial Age's focus on developing an adequate supply of material goods and services (such as health, education, policing, and social welfare) to a focus on monitoring and improving the number and quality of those goods and services.

Third, advances in information technology have made information the primary instrument for citizens to access wants throughout the economy and society. This fundamental shift in the economy, society, and polity has occurred in those countries advanced enough economically to warrant designation as postindustrial or knowledge-based societies (Benjamin, 1980, 2003; Hage and Powers, 1993). In this new environment, individual and collective choices become much more numerous, complex, and often are in conflict, requiring citizens to be able to sort them out. Information about choices is also much greater and more widely available than before, as well as more immediate due to the Internet, cable television, blogs, cell phones, and personal computers.

Under these conditions concentrating on content in education remains important, but is no longer enough. Critical thinking, analytic reasoning, problem solving and written communication skills are also needed. Students need to be able to judge the quality of information sources associated with recommendations and arguments. They must determine if arguments are concise, logical, built on plausible assumptions and linked to credible evidence. In short, students need to be better able to sift through documents, materials, graphs, figures and oral arguments to arrive at reasoned, reflective positions.

Measuring Higher Order Skills, and the Diagnostic Logic of the CLA

The CLA measures institutional (or programmatic) contributions to the development of these higher-order skills holistically. The CLA does not claim to measure all of undergraduate education, nor does it claim to capture all aspects of critical thinking, analytic reasoning, problem solving and written communication skills. Rather, our claim is that the holistic tasks themselves have a high degree of overlap with the stated mission of most colleges; they also have important face validity because faculty agree that graduating students should be able to perform these tasks at an acceptable level. Therefore, we claim that any definition of higher-order skills will include, among its characteristics, the attributes the CLA tasks measure. In turn, we are able to argue that increasing CLA scores increases higher order skills.

The CLA has a set of features that taken together uniquely characterize its measurement and analytical approach. The CLA performance tasks present realistic problems and assess students' ability to use the provided information in order to create a justified solution guided by a series of questions. To address the problems successfully, students need to think critically and evaluate the information they are provided. If that information includes quantitative information (e.g., arrest rates, consumer demand over time) students need to reason with quantitative information in the context of this concrete problem. If the information includes works of art, students need to reason spatially or musically in the context of this specific situation. Moreover, students often need to make a decision about a course of action that balances multiple (and possibly conflicting) goals, values or perspectives. And they must communicate that decision in writing clearly and cogently with a rationale linked to the information provided and the reasoning they applied.

The CLA, then, assumes that multiple abilities will be brought to bear on a concrete problem and that students will vary in the ways in which they use their abilities to respond to the problem or task at hand (e.g., Shavelson et al., 2002).

THE CLA: A DIFFERENT APPROACH

The Temptation of Creating Sub-Scores

While there might be agreement about the viability of such measures, there are different perspectives about how to report out scores. Attempts have been made with other measures to define critical thinking as a discrete set of sub-skills that can be broken out separately, and then arranged along a series of dimensions. What are the constituent parts of critical thinking that can be identified? How can we break down problem solving to smaller, manageable pieces? Often, the assumption is grounded in the notion that understanding such sub-parts will easily allow for an educational response.

However, a rather different approach is at the heart of the CLA; here, these skills are honored as being interrelated and therefore assessed and scored holistically. The CLA reflects the recognition that these higher order skills are inherently intertwined in a complex manner, both in the tasks and the responses to them. While some of these abilities that are used in addressing CLA tasks can be named, for example the make-an-argument measures writing, critique-an-argument benchmarks analysis skills, and the performance tasks track critical thinking, the CLA takes the view that the whole is greater than the sum of the parts and does not attempt to pull the abilities needed to answer performance tasks themselves apart in an artificial manner (more in Chapter 5).

The CLA in its holistic approach to assessing these higher-order skills differs from typical standardized tests of college learning. These other tests employ multiple-choice and short-answer test questions that can be and are (up to the limits of design and reliability) divided into constituent parts. That way sub-scores can be created and both item and sub-scale analyses can be used to describe and explain relationships between and among those sub-scores. These sub-scores are then added up to arrive at some total score that is interpreted as reflecting achievement or learning. Such tests contrast significantly with the CLA, then, in their underlying

philosophical approach (Shavelson, 2007). But this leaves us with a challenge. If we agree that the scores should not be separated, how should a campus respond? How will they know what to do to improve if their scores are below expected, and how will they know what they are doing well if their scores are above expected?

We note that whereas the higher education community has been "trained" on sub-scores, we suggest that the CLA demonstrates how holistic measures and scores more naturally integrate into teaching and learning improvement.

The Initial Focus: Starting with the Institution

CLA score reports are prepared in ways that focus on the institution, and provide information about overall value added. After an institution receives its results it must consider and understand the relative contribution of a number of factors to the institution's performance on the CLA. Perhaps paramount among these factors is the campus' academic program – academic majors, general education, and other learning opportunities. The question is whether the academic program provides opportunities for students to learn to tackle the kind of problems the CLA represents, and to learn from their instructors to improve their performance.

More importantly, using the CLA means that the campus commits to going beyond narrowly defined disciplines or subdisciplines and a smorgasbord of largely unrelated courses to meet the holistic learning goals the colleges express. That is, colleges commit to an integrated vision of learning, teaching and assessment that focuses on improvement of higher order skills. This is important because higher-order skills exhibit public-good-like characteristics by which is meant they possess the quality of jointness of supply, i.e., produced by many contributors. In this case no one department, course or major produces them and all graduates should have them. Professors may and do argue that since they do not teach these skills in their departments, they should not be evaluated for their achievement by students. Nor should assessment of student learning be focused on their acquisition. This position has carried the day until very recently. Now, employers, commentators, higher education reform groups and observers of higher education increasingly argue that it is these public good-like skills that are precisely what undergraduate education should improve: that narrow content or specialization should not be the major focus of undergraduate education. Undergraduate education should teach students how to think and not just train them to be proficient in a specific academic field. From this perspective,

the institution, not the department, becomes the initial focus of assessment because no one department produces or improves these skills.

With this recognition, initial institution-level performance reports may lead campuses to modify the CLA portion of its learning assessment program as well as other factors in the program. These modifications include, but are not limited to, expanding the use of the CLA (perhaps through in-depth sampling), setting the CLA in the context of other assessments, and analysis of other quantitative and qualitative data about student performance.

Discussions of how to improve that performance and how subunits might perform then become pertinent questions. Conjectures about how to improve performance might be tested out in, say, variations in a general education program. Conjectures about sub-unit performance might be tested out in smaller units and levels at the institution. For example, a university might ask: How do individual colleges vary in their value added CLA performances? What distinguishes, say, the engineering college that performs better than expected than the business or arts and science colleges? Similarly, smaller liberal arts colleges might assess the relative performance of selected departments and programs on the CLA. Institutions might consider what are the effects of critical factors such as transfer versus "native" student performance, gender differences, and ethnic/racial differences? What is the relative contribution of factors such as class size, presence or absence of core curriculum, student-centered versus lecture format-based instruction? What does an audit of the types of assessments used in the classroom show?

Shifting the Focus: Moving from the Institution to the Program into the Classroom

It is of course insufficient to focus solely on the institution as whole. The second part of the process is to determine ways to move attention to the work of faculty and activity in departments or programs and ultimately in the classroom. We believe that for programs to improve, they need information about (1) where they are going (goals), (2) where they are at present, (3) how to close the gap, and (4) a mechanism for monitoring, feeding back information, and providing incentives for getting there. While our focus here will be on the diagnostic use of the CLA, we want to emphasize that unless the program has in place mechanisms for putting assessment information into action by testing conjectures as to how to improve teaching and

learning and keeping progress at the forefront of programmatic missions, all the assessment in the world won't improve things. As the saying goes, weighing a pig doesn't make it fatter.

The challenge is that scores that are generated from the holistic performance tasks themselves do not readily lend themselves to disaggregated sub-scores that can be readily analyzed. Given the interrelated nature of education, knowledge and skills, it is hard to disentangle these elements based on a student's response. For example, if in a CLA performance task there were a key table of numbers a student did not refer to in her/his response, we do not know if it is because the student (a) did not know how to read the table; (b) did not have the analytic skills to realize the importance of that table given the task; (c) was able to recognize the importance, but did not have the writing skills to communicate this; (d) had all of these skills, but was just uninterested or not motivated to perform; and so on. As a measure of performance, these reasons matter less in terms of giving a score, but as a diagnostic tool for faculty and a campus, these reasons matter more. If a faculty member had a sense of which (or which combination) of those factors were associated with performance, she/he would have a better idea about what to do about it.

Thus, whereas it is important for assessment to initially focus on the institution as a whole, when it comes to teaching and learning, the locus for change occurs at the department or program level on into individual classrooms. Given the holistic nature of the CLA measures it may not seem entirely obvious how to translate the results in a way to effect change in departments or programs that in turn affect classroom teaching and coordination among courses.

The conceptual shift needed here is to understand that to do program-level work that impacts the classroom, the diagnostic power of the CLA comes not from the score results, but rather from the CLA measures themselves. Thus, it is important to rehearse the logic of why the important next step is to take the CLA measures directly into academic programs and departments as a central way in which the CLA is used. The intent is to ultimately impact classroom teaching and learning in a coherent, coordinated way such that progress toward departmental and program goals can be monitored and various conjectures for improvement tried out and tested.

One useful thing that can be done with CLA-type tasks is to put them in the hands of faculty members so they can be used both as a focus for program or departmental planning, monitoring, and feedback, and as a classroom tool for teaching and learning. If the tasks used in

the CLA are the kinds of tasks colleges say they want their students to succeed at, and we have evidence that they are, it seems reasonable to incorporate them into program and classroom teaching and learning activities. By making a wide variety of such tasks available, we believe this will increase the capacity of students to solve problems, think critically, and communicate their ideas not only on the classroom-embedded tasks themselves, but on similar types of tasks that students encounter in life. For more on these efforts, see Chapter 6.

FINAL THOUGHTS

The CLA offers an alternative to assessment approaches that rely heavily on multiple-choice tests and separate sub-scores. The CLA builds on a rich legacy reflected in the development of assessment tools and ways to think about assessing higher education (Shavelson, 2007), but more importantly how we think about the work of higher education and the role of faculty in using assessment data. We argue here that holistic tasks can be a key component of the way we measure the work of our colleges and universities, and that we should resist the temptation to divide these scores into sub-scores when such information lacks significant meaning. We also acknowledge the role of the faculty as a whole in developing these skills. To return to the example that opened this essay, we have learned much from efforts to re-think the manufacturing process, and to recognize the value when all team members come together in addition to working in their separate units, and that a commitment to that collective goal – be that building cars or educating students – is enhanced. We argue that this can be done by empowering faculty with the tools they need to interpret, re-create, and use assessment data at the program and ultimately at the classroom level, while still linking this together at the institutional level.

4. THE IMPORTANCE OF THE FACULTY IN THE AGE OF ASSESSMENT[25]

Roger Benjamin

There is increased interest in assessment and accountability of learning in higher education. This is signified both by the Spellings Commission Report (Commission on the Future of Higher Education) and the initiatives of several of the national associations of higher education to respond to the call for greater assessment and accountability (such as the VSA). Many colleges and universities are signing up for programs that call for them to make public information about their student learning outcomes. The increased attention is also demonstrated by the number of articles on the subject in the press and journals related to higher education. However, within these discussions, the importance of the faculty is either overlooked or given short shrift. This demonstrates a fundamental problem about how we are currently approaching reforms in both assessment and learning. Of course assessment instruments must be reliable and valid. However, if the assessment instruments do not directly assist the faculty to improve the teaching and learning results of their students, they will not be accepted and used by the faculty. As a consequence, little progress in the assessment and accountability movements in higher education is likely to occur. It is, therefore, essential that any effort in learning assessment directly contributes to the work of the faculty as educators.

In the discussion that follows I will provide evidence for this assertion by laying out the assumptions and strategy of the CLA and, in particular, its focus on the faculty as the ultimate "customer." The analysis moves through four steps, all of which directly speak to the importance of faculty efforts in teaching and learning. I first discuss why performance assessment, as a new approach, is important to higher education. I then justify the strategy of comparing institutions and, in particular, value-added comparisons of student learning growth between institutions. Next, I present the CLA template for continuous learning improvement which is designed to get performance assessments into faculty hands and, therefore, the classroom. Finally, I present the argument about why the contribution of the assessment instrument to teaching and learning should be the central focus in the higher education assessment and accountability discussion.

[25] Adapted and excerpted from Benjamin (2008)

THE RATIONALE FOR PERFORMANCE ASSESSMENT

The current assessment regime, dominated by multiple-choice tests, is no longer sufficient in the new knowledge economy. For a century, multiple-choice tests have been the principal assessment method in education. This made sense in America's industrial era of development, by mirroring the focus on the mastery of content demonstrated by students' ability to recall facts. Today, we live in an economy dominated by information and services rather than physical goods. In this knowledge economy it is more important to be able to access, structure, and use information than merely recall facts. This places a premium on the ability of students to reason, assess the relevance of information, and make arguments; in short, think critically. This effort to focus on critical thinking skills is being implemented in classrooms across the country, where faculty are arming their students to navigate a constantly changing world. The manner in which we assess students must reflect these interests. Multiple-choice tests may present examples of correlations and causation and ask students to identify whether each is correctly or incorrectly applied. However, responding to such choices passively is very different from asking students in performance assessments to actively critique a case study that presents an argument about data in which correlation or causation are misused. And, it is important to underline the requirement in the knowledge economy for citizens to actively shape the information at their disposal rather than simply respond passively to choices put before them.

Assessment must catch up with an emerging reform agenda in higher education resulting from our new understanding of student learning. At the most basic level this involves understanding that the meaning of knowledge itself is undergoing a significant shift. New theories from cognitive science stress the importance of improving students' ability to structure learning experiences that help them use what they have learned in new settings. Simon (1996, p. 43) argues that the meaning of "knowing" has changed from being able to recall information to being able to find and use it. Under these conditions, the proponents of the new learning theories argue that active learning becomes critical because students must learn to recognize when they understand a subject and when they need more information (Pellegrino et al., 2001).The implications for higher education are profound. Consider the assumptions which structure and guild higher education in the industrial era. The lecture format was the norm with students seen as passive receptacles receiving the content provided by lecturers. The role of higher education

was to transmit knowledge. Faculty and administrators were comfortable with these assumptions, because even though it was understood that knowledge was progressing in multiple fields, most shared the view that there was a stable, enduring stock of knowledge that graduating seniors should know. Under these circumstances content was emphasized and multiple-choice tests were preferred as the assessment tool.

Over the past two decades it has become clear that a new vision of undergraduate education is developing in response to the changing definition of knowledge. It is comprised of three parts:

- A shift from the lecture format to a student-centered approach that emphasizes analytic-based writing. Faculty are much more interested in having their students participate actively in their learning and students appear equally interested in doing so. Although evidence is still in the formative stage, it appears that colleges that emphasize analytic-based writing produce students who do well on assessments that benchmark higher order skills.

- A change in emphasis from the current focus in curriculum and texts on content to case- and problem-based materials that ask students to apply what they know to new situations. This is reflected in curriculum reform and is also resulting in textbook publishers substituting this new approach for solely content filled volumes. The graduate business school emphasis on the case approach to learning may be an early example of this strategy.

- Change in assessment from multiple-choice and short answer formats to open ended essays that are better aligned with the first two parts of the reform. The CLA is one such constructed essay approach designed to meet the needs of the first two parts of the reform. There are likely to be other approaches developed in the future.

The changes in how we approach teaching and learning have resulted in the need to examine the way in which we assess undergraduate education. However, it is essential that these new means of assessing learning, such as the CLA, be based on and unified with the work of educators in the classroom. None of the three changes operate independently; changes in style, content and assessment must be a collaborative effort between faculty, administrators and students.

Criteria for Evaluation of Assessment Instruments

Before moving to the next three phases of this discussion, it is important for the reader to understand the criteria for evaluating assessment instruments. It is useful to review the six

criteria listed below together, because they are all critical to the creation of successful assessment instruments.

- *Reliability and validity*: If a measure is not reliable and its interpretation not meaningful, the other criteria are irrelevant.

- *Impact on curriculum and instruction*: If the assessment instrument is not found to be authentic and useful for improving teaching and learning in the classroom, it will not be effective. Faculty will be correct to not accept it.

- *Cost*: The assessment instruments must be affordable to both the institution and individual students.[26]

- *Time*: Time for assessing students is precious. Faculty do not wish to use up excessive amounts of time testing, since this negatively affects the amount of time left for instruction. It is also important that the assessment instrument not be too cumbersome to the institution by taking an excessive amount of time to derive results.

- *Standardization:* To be able to compare results of an assessment, either with prior classes of the same course, between courses and majors within an institution, or between institutions, it is important to make the testing time, conditions and assessment instrument uniform. This is what is meant by standardization in a technical sense. Without meeting this criterion, assessment tests remain useful for specific courses only.

- *Fairness:* Time and resources must be the same for all students taking the test. Equally, racial/ethnic groups and majors must not be advantaged or disadvantaged by the assessment instrument.

Different forms of assessment may not be able to fulfill all six of these criteria on the institutional level. Multiple-choice tests may score highly on time, cost and standardization but fail to create meaningful information for instructors to use in their curriculum and instruction. There have also been numerous questions raised to their fairness among different racial and ethnic groups, which is an important issue for all assessment tools.

The Case for Comparison Is Strong

As noted above, comparison between institutions, if possible, is a good goal since it is likely the great diversity in American higher education will yield interesting lessons for teaching and learning best practices. Eventually, comparativists would argue, this search strategy may

[26] This and the criteria that follow are taken from Klein (2002), p. 27.

lead to better understanding of the systematic steps to take to improve teaching and learning (Benjamin, 2008).

In the case of the CLA, we argue that the approach to emphasize is value-added comparison of broad abilities called higher order skills (critical thinking, analytic reasoning, problem solving and written communication) at the institutional level. The institution is the appropriate initial unit of analysis to start with because the growth of higher order skills is a cumulative process for students; i.e., the sum of all the educational experiences they have, not just of a single course or even solely the coursework in their major. An important, first order educational question to ask is, "How much did our teaching lead students to learn?" To answer this question, one needs to know the extent to which student improvement is comparable to that of equally able students across time within the institution or between similarly situated institutions.

To summarize, the main focus of comparisons should be made not between absolute levels of achievement but between the amount of value that colleges and universities add (Klein et al., 2007 and 2008). However, this leads to two different and somewhat contradictory counter-arguments by those who object to the use of value-added comparisons at elite institutions. The first is that elite institutions should only be compared to peer institutions that face the same challenges with similar students. The second argument is that elite institutions might want to compare the absolute level of student attainment they achieve to that of their peers since this is the scale they excel on. Both of these arguments can be accommodated in a voluntary system of assessment where institutions can choose what to focus on: value-added or absolute levels of performance.

An additional benefit to comparability is the possibility of a shared vocabulary and dialogue between faculty both inside of institutions and with colleagues at other institutions across the United States. Beyond assessing institutions, the CLA is an instrument that can provide a means for creating a meaningful dialogue on learning. By providing a reference point for discussion, the CLA enables faculty and institutions to collaborate and truly define their learning objectives.

Moving Beyond the Initial Assessment of the Institution to the Faculty

The initial institutional comparison provides faculty and administrators a useful signal about where their institution stands in comparison to others. Inputs, such as admission qualifications of entering students, per student endowment, internal processes such as instructor/student ratios, and outputs such as retention, graduation rates, and CLA results of graduating students should also be studied to develop an efficient description of the factors that correlate (or not) with CLA results. However, it is critical that this information then be used to benefit teaching and learning. In order to do so, there are two important issues that should be examined as well.

First, the question of which colleges (if the institution is a university), departments or programs (if the institution is a college) are particularly strong or weak contributors to the institution-level CLA should be answered. Second, an audit of existing assessments used should be conducted to determine how extensive constructed essay tests, case and problem assessments are in common usage in the college, in comparison to multiple-choice tests.

More generally, faculty and administrators at a college or university should consider joining or founding a consortium of similarly situated institutions for the purpose of sharing best practices used to improve teaching and learning. The consortium approach is proving quite attractive to institutions and systems of institutions around the country.

Finally, the most important step is to get published CLA tasks into the hands of the faculty and assist them to:

- Use them in their classroom
- Develop their own performance tasks
- Choose case studies and problems for text and curriculum material instead of only content
- Test students in the classroom with CLA-like tasks that they can use to diagnose the strengths and weaknesses of students.[27]

These steps comprise a continuous system of improvement. The institution's global score provides an important signal that triggers an internal focus on what correlates – positively and negatively – to the institution-level score. Is it instructor/student classroom ratios, factors relating to the quality of student life, teaching and learning style, or a combination of these and

[27] See chapter 6 for additional information on this topic.

other variables catalogued in IPEDS? After developing an efficient description of these factors in relationship to CLA results, the follow up steps relate to understanding what led to those results and deciding what improvement goals make sense for institutions. The faculty will then be able to implement the best practices chosen to achieve the new student learning goals. The cycle can then begin at the institutional level over again, since assessment should be a continuous activity.

The Role of the Faculty

Discussion about assessment and accountability tends to focus on policy issues or the reliability or validity of assessment instruments. These are, of course, important issues. But discussion of the relevance of the assessment instrument to teaching and learning is either absent or approached as an after thought. To reiterate, the assessment instrument must be reliable and valid, but the threshold question is the instrument's relevance to faculty in the classroom. The relevance of the assessment instrument to the faculty in the classroom should take precedence over its technical dimensions and larger policy debates over whether or how assessment or accountability should occur.

The faculty needs to be the focus of assessment because individual instructors are at the center of teaching, learning, and curriculum matters. This includes whether and how to assess the quality of teaching and whether and how to use assessment instruments as a diagnostic tool to assist their efforts to improve teaching and learning. The implication of this point is that faculty buy-in is critical to the future of assessment and accountability in the academy. Until it is clear that testing organizations have developed assessment instruments that are accepted by faculty as valuable aids to their instruction, it is unlikely that we will move the policy debates on assessment and accountability forward in higher education. Thus, our focus should be on encouraging the faculty to use assessment instruments like the CLA that are in line with their teaching and learning goals..

If the faculty buy in to using assessment instruments as central ingredients in their work, good things on the other fronts, such as appropriate accountability and use of assessment evidence for internal governance and diagnostic purposes, will occur.[28] However, this must begin by the faculty recognizing the inherent value of assessment to their own work as teachers. This

[28] For example, faculty may be able to reclaim governance over the undergraduate curriculum (Benjamin, 2007).

will only occur if, in fact, the tools themselves are proven to be effective; it is with this goal in mind that we created the CLA.

CONCLUSION

If the human capital school demonstrates the importance of education, the implications of the knowledge economy and recent theories of learning place the focus on improving the higher order skills of the next generation of students. These developments, in turn, create an urgent need to develop and implement a testing paradigm that measures and stimulates these skills. That new paradigm is the field of performance assessment. Performance assessment-based tasks are realistic problems, scenarios or simulations that both faculty and students find useful to teach to since there are no clear "right" answers; rather, the question is the quality of reasoning the student exhibits. Thus, performance assessment is better aligned with the new vision of teaching and learning emerging throughout the higher education sector. However, it is essential that this paradigm be intimately connected to the work of the faculty in the classroom. At CAE we are making this connection through the CLA in the Classroom program that has already reached hundreds of faculty. The goal of the faculty academy part of this program is to train faculty to design and use performance assessment. We are now posting the best examples of faculty constructed tasks on the CAE website. Many institutions participating in CLA testing are providing the necessary support to their faculty to implement this new approach to assessment.

Since performance assessment is rarely taught in schools of education or psychology departments, the field of performance assessment is underdeveloped. We need to encourage foundations and universities to devote more resources to the development of assessment scientists dedicated to performance assessment. In this way we can develop a greater critical mass of assessment scientists who can lead the way beyond the current dominant multiple-choice testing paradigm, which was central to our teaching and learning needs in the industrial era.

Educational institutions are being challenged to do a better job educating tomorrow's workforce. A college education has never been more necessary for productive participation in society. Employers now seek individuals able to think and communicate to meet the requirements of the new "knowledge economy." This means that the skills taught in higher education are changing. There is less emphasis now on acquiring content knowledge and more

on finding information, using reasoning to solve problems, and using communication skills to convey this information.

Performance assessment tools are therefore necessary not only to evaluate whether students are learning the skills expected of them in today's workforce, but also to spur educational advances in teaching such skills. And that is the central role of the faculty.

Section II

CLA Implementation

5. THE ARCHITECTURE OF THE CLA[29]

Chaitra Hardison, Esther Hong, Marc Chun, Heather Kugelmass and Alex Nemeth

INTRODUCTION

The CLA is comprised of three types of prompts within two types of tasks: the Performance Task and the Analytic Writing Task. Most students take one task or the other. The Analytic Writing Task includes a pair of prompts called Make-an-Argument and Critique-an-Argument.

The CLA uses direct measures of skills in which students perform cognitively demanding tasks from which quality of response is scored. All CLA measures are administered online and contain open-ended prompts that require constructed responses. There are no multiple-choice questions. The CLA tasks require that students integrate critical thinking, analytic reasoning, problem solving, and written communication skills. The holistic integration of these skills on the CLA tasks mirrors the requirements of serious thinking and writing tasks faced in life outside of the classroom.

This chapter provides the reader with an excerpted example of a retired Performance Task and an example of an Analytic Writing Task. The Crime Reduction Performance Task was delivered as part of the CLA from fall 2005 through spring 2007, after which it was retired. The Make-an-Argument and Critique-an-Argument prompts presented here to represent the Analytic Writing Task were not delivered as part of the CLA, but they were developed by our measurement science team and underwent initial field-testing. They remain in the same spirit, format, and construction as our "live" Make-an-Argument and Critique-an-Argument prompts.

Representative student answers at the high, moderate, and low ranges are also presented for a performance task, Make-an-Argument and Critique-an-Argument as well as the characteristics of these responses that qualify them for a particular level. These answers may be of interest to colleagues at participating CLA institutions because they can see what answers at different levels look like and the extent which the scores of their students fall into these categories.

[29] Adapted and excerpted from a report prepared by Hardison, Hong, Chun, Kugelmass and Nemeth (note: primary work completed by the first two authors; the last three names are alphabetical).

Please note that these examples were not chosen to represent the range in CLA prompt topics. Rather, they reflect how prompts with different scenarios can assess similar concepts (e.g., the concept of causation versus correlation appears in both the Crime Reduction Performance Task and the Weddings Critique-an-Argument prompt) as well as how prompts with different main concepts can be presented through similar scenarios (e.g., both the Crime Reduction Performance Task and the Government Funding Make-an-Argument prompt present crime as a policy issue).

TASK DESCRIPTION

Performance Task

Each Performance Task requires students to use an integrated set of critical thinking, analytic reasoning, problem solving, and written communication skills to answer several open-ended questions about a hypothetical but realistic situation. In addition to directions and questions, each Performance Task also has its own document library that includes a range of information sources, such as letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes or transcripts. Students are instructed to use these materials in preparing their answers to the Performance Task's questions within the allotted 90 minutes.

The first portion of each Performance Task contains general instructions and introductory material. The student is then presented with a split screen. On the right side of the screen is a list of the materials in the Document Library. The student selects a particular document to view by using a pull-down menu. On the left side of the screen are a question and a response box. There is no limit on how much a student can type. When a student completes a question, he or she then selects the next question in the queue.

No two Performance Tasks assess the exact same combination of skills. Some ask students to identify and then compare and contrast the strengths and limitations of alternative hypotheses, points of view, courses of action, etc. To perform these and other tasks, students may have to weigh different types of evidence, evaluate the credibility of various documents, spot possible bias, and identify questionable or critical assumptions. Performance Tasks also may ask students to suggest or select a course of action to resolve conflicting or competing strategies and then provide a rationale for that decision, including why it is likely to be better than one or more other

approaches. For example, students may be asked to anticipate potential difficulties or hazards that are associated with different ways of dealing with a problem, including the likely short- and long-term consequences and implications of these strategies. Students may then be asked to suggest and defend one or more of these approaches. Alternatively, students may be asked to review a collection of materials or a set of options, analyze and organize them on multiple dimensions, and then defend that organization.

Performance Tasks often 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.

All of the Performance Tasks require students to present their ideas clearly, including justifying their points of view. For example, they might note the specific ideas or sections in the document library that support their position and describe the flaws or shortcomings in the arguments' underlying alternative approaches.

Analytic Writing Task

Students write answers to two types of essay prompts, namely: a "Make-an-Argument" question that asks them to support or reject a position on some issue; and a "Critique-an-Argument" question that asks them to evaluate the validity of an argument made by someone else. Both of these tasks measure a student's skill in articulating complex ideas, examining claims and evidence, supporting ideas with relevant reasons and examples, sustaining a coherent discussion, and using standard written English.

A "Make-an-Argument" prompt typically presents an opinion on some issue and asks students to write, in 45 minutes, a persuasive, analytic essay to support a position on the issue. Key elements include: establishing a thesis or a position on an issue; maintaining the thesis throughout the essay; supporting the thesis with relevant and persuasive examples (e.g., from personal experience, history, art, literature, pop culture, or current events); anticipating and countering opposing arguments to the position, fully developing ideas, examples, and arguments; crafting an overall response that generates interest, provokes thought, and persuades the reader;

organizing the structure of the essay (e.g., paragraphing, ordering of ideas and sentences within paragraphs); employing transitions and varied sentence structure to maintain the flow of the argument; and utilizing sophisticated grammar and vocabulary.

A "Critique-an-Argument" prompt asks students, in 30 minutes, to critique an argument by discussing how well reasoned they find it to be (rather than simply agreeing or disagreeing with the position presented). Key elements of the essay include: identifying a variety of logical flaws or fallacies in a specific argument; explaining how or why the logical flaws affect the conclusions in that argument; and presenting a critique in a written response that is a grammatically correct, organized, well-developed, logically sound, and neutral in tone.

TASK DEVELOPMENT

Task development occurs through an iterative process. A team of researchers and writers generate ideas for Make-an-Argument and Critique-an-Argument prompts, and Performance Task storylines, and then contribute to the development and revision of the prompts and Performance Task documents.

For Analytic Writing Tasks, multiple prompts are generated, revised and pre-piloted, and those prompts that elicit good critical thinking and writing responses during pre-piloting are further revised and submitted to more extensive piloting.

During the development of Performance Tasks, care is taken to ensure that sufficient information is provided to permit multiple reasonable solutions to the issues present in the Performance Task. Documents are crafted such that information is presented in multiple formats (e.g., tables, figures, news articles, editorials, letters, etc.).

While developing a Performance Task, a list of the intended content from each document is established and revised. This list is used to ensure that each piece of information is clearly reflected in the document and/or across documents, and to ensure that no additional pieces of information are embedded in the document that were not intended. This list serves as a draft starting point for the analytic scoring items used in the Performance Task scoring rubrics. During revision, information is either added to documents or removed from documents to ensure that students could arrive at approximately three or four different conclusions based on a variety of evidence to back up each conclusion. Typically, some conclusions are designed to be supported better than others. Questions for the performance task are also drafted and revised during the

development of the documents. The questions are designed such that the initial questions prompt the student to read and attend to multiple sources of information in the documents, and later questions require the student to evaluate the documents and then use their analysis to draw conclusions and justify those conclusions using information from the documents.

After several rounds of revision, the most promising of the Performance Tasks and the Make-an-Argument and Critique-an-Argument prompts are selected for pre-piloting. Student responses from the pilot test are examined to identify what pieces of information are unintentionally ambiguous, what pieces of information in the documents should be removed, etc. After revision and additional pre-piloting, the best functioning tasks (i.e., those that elicit the intended types and ranges of student responses) are selected for full piloting.

During piloting, students complete both an operational task and one of the new tasks. At this point, draft scoring rubrics are revised and tested in grading the pilot responses, and final revisions are made to the tasks to ensure that the task is eliciting the types of responses intended.

SCORING CRITERIA

This section summarizes the types of questions addressed by CLA scoring of all task types. Because each CLA task and their scoring rubrics differ, not every item listed is applicable to every task. The tasks cover different aspects of critical thinking, analytic reasoning, problem solving, and writing and in doing so can, in combination, better assess the entire domain of performance.

Assessing Critical Thinking, Analytic Reasoning and Problem Solving Skills

Applied in combination, critical thinking, analytic reasoning and problem solving skills are required to perform well on CLA tasks. We define these skills as how well students can evaluate and analyze source information, and subsequently to draw conclusions and present an argument based upon that analysis. In scoring, we specifically consider the following items to be important aspects of these skills.

Evaluation of evidence

How well does the student assess the quality and relevance of evidence, including:

- Determining what information is or is not pertinent to the task at hand;
- Distinguishing between rational claims and emotional ones, fact from opinion;
- Recognizing the ways in which the evidence might be limited or compromised;
- Spotting deception and holes in the arguments of others; and
- Considering all sources of evidence?

Analysis and synthesis of evidence

How well does the student analyze and synthesize data and information, including:

- Presenting his/her own analysis of the data or information (rather than "as is");
- Committing or failing to recognize logical flaws (e.g., distinguishing correlation from causation);
- Breaking down the evidence into its component parts;
- Drawing connections between discrete sources of data and information; and
- Attending to contradictory, inadequate or ambiguous information?

Drawing conclusions

How well does the student form a conclusion from their analysis, including:

- Constructing cogent arguments rooted in data/information rather than speculation/opinion;
- Selecting the strongest set of supporting data;
- Prioritizing components of the argument;
- Avoiding overstated or understated conclusions; and
- Identifying holes in the evidence and subsequently suggesting additional information that might resolve the issue?

Acknowledging alternative explanations/viewpoints

How well does the student acknowledge additional perspectives and consider other options, including:

- Recognizing that the problem is complex with no clear answer;
- Proposing other options and weighing them in the decision;
- Considering all stakeholders or affected parties in suggesting a course of action; and
- Qualifying responses and acknowledging the need for additional information in making an absolute determination?

Assessing Writing Skills

Analytic writing skills invariably depend on clarity of thought. Therefore, analytic writing and critical thinking, analytic reasoning, and problem solving are related skills sets. The CLA measures critical thinking performance by asking students to explain in writing their rationale for various conclusions. In doing so, their performance is dependent on both writing and critical

thinking as integrated rather than separate skills. We evaluate writing performance using holistic scores that consider several aspects of writing depending on the task. The following are illustrations of the types of questions we address in scoring writing on the various tasks.

Presentation

How clear and concise is the argument? Does the student…

- Clearly articulate the argument and the context for that argument;
- Correctly and precisely use evidence to defend the argument; and
- Comprehensibly and coherently present evidence?

Development

How effective is the structure? Does the student…

- Logically and cohesively organize the argument;
- Avoid extraneous elements in the argument's development; and
- Present evidence in an order that contributes to a persuasive and coherent argument?

Persuasiveness

How well does the student defend the argument? Does the student…

- Effectively present evidence in support of the argument;
- Draw thoroughly and extensively from the available range of evidence;
- Analyze the evidence in addition to simply presenting it; and
- Consider counterarguments and address weaknesses in his/her own argument?

Mechanics

What is the quality of the student's writing?

- Is vocabulary and punctuation used correctly;
- Is the student's understanding of grammar strong;
- Is the sentence structure basic, or more complex and creative;
- Does the student use proper transitions; and
- Are the paragraphs structured logically and effectively?

Interest

How well does the student maintain the reader's interest?

- Does the student use creative and engaging examples or descriptions;
- Does the structure, syntax and organization add to the interest of their writing;
- Does the student use colorful but relevant metaphors, similes, etc.;
- Does the writing engage the reader; and
- Does the writing leave the reader thinking?

For specific information about the scoring, see Hardison et al. (2009).

SCORING PROCESS

Score Sheet

There are two types of items that appear on a CLA score sheet: analytic and holistic. Analytic scoring items are particular to each prompt and holistic items refer to general dimensions, such as evaluation of evidence, drawing conclusions, acknowledging alternative explanations and viewpoints, and overall writing. We compute raw scores for each task by adding up all points on all items (i.e., calculating a unit-weighted sum).

Performance Task scoring is tailored to each specific prompt and includes a combination of both holistic and analytic scoring items. Though there are many types of analytic items on the Performance Task score sheets, the most common represent a list of the possible pieces of information a student could or should raise in their response. These cover the information presented in the Performance Task documents as well as information that can be deduced from comparing information across documents. The analytic items are generally given a score of 0 if the student did not use the information in their response, or 1 if they did. The number of analytic items varies by prompt.

Performance Task holistic items are scored on four or seven-point scales (i.e., 1-4 or 1-7). There are multiple holistic items per Performance Task that require graders to provide an evaluation of different aspects of critical thinking and reasoning in the student responses. These holistic items include areas such as the student's use of the most relevant information in the Performance Task, their recognition of strengths and weaknesses of various pieces of information, overall critical thinking, and overall writing.

Critique-an-Argument score sheets also include a combination of analytic and holistic scores. Critique-an-Argument analytic items are a list of possible critiques of the argument presented in the prompt. In addition, a few holistic items are used to rate the overall quality, critical thinking and writing over the entire response.

Make-an-Argument score sheets contain only holistic items scored on four or seven-point scales (i.e., 1-4 or 1-7). The holistic items include ratings for various aspects of writing (e.g.,

organization, mechanics, etc.) and critical thinking (e.g., reasoning and logic, sophistication and depth of treatment of the issues raised in the prompt) as well as two overall assessments of writing and critical thinking.

For all task types, blank responses or responses that are entirely unrelated to the task (e.g., writing about what they had for breakfast) are assigned a 0 and are flagged for removal from the school-level results.

Scoring Procedure

During the 2007-2008 CLA administration, all scoring was conducted by trained scorers. Starting in fall 2008, a combination of machine and human scoring is being used.

All scorer candidates undergo rigorous training in order to become certified CLA scorers. Training includes an orientation to the prompt and score sheet, instruction on how to evaluate the scoring items, repeated practice grading a wide range of student responses, and extensive feedback and discussion after scoring each response. After participating in training, scorers complete a reliability check where they score the same set of student responses. Scorers with low agreement or reliability (determined by comparisons of raw score means, standard deviations and correlations among the scorers) are either further coached or removed from scoring.

PERFORMANCE TASK: CRIME REDUCTION

In this section, we present you with excerpts from a retired CLA Performance Task called "Crime Reduction" as follows:

- Introduction
- A "Document Library" consisting of the Instructions and seven primary source documents
- Question #1 (the first out of a total of three Crime Reduction questions)

We also present representative student answers to question one at the high, moderate, and low levels as well as provide the characteristics of these responses that qualify them for a particular level. This section therefore should be of practical value to colleagues at participating CLA institutions and observers who wish to get a sense of what the range of answers from low to high on the CLA looks like. (See also Hardison and Vilamovska forthcoming)

Below is a screen shot taken of the beginning of the Crime Reduction task; this is how the students would have seen it on-screen:

We then go in-depth with the first of the three Crime Reduction questions, explaining the scoring items associated with the first question and providing you with three actual student responses to the question accompanied by a brief explanation of what characterizes one response as a "high" response, one as a "moderate" response, and one as a "low" response.

Introduction

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

While your personal values and experiences are important, please answer all the questions solely on the basis of the information above and in the Document Library.

Write your answers in the box below each question. You can write as much as you wish; you are not limited by the size of the box on the screen.

Document Library

Here, we provide brief descriptions (i.e., not the full text) of each of the documents that students needed to examine in order to answer all three of the Crime Reduction questions.

Document 1: Instructions

Pat Stone is running for reelection as mayor of Jefferson, a city in the state of Columbia. Mayor Stone's opponent in this contest is Dr. Jamie Eager. Dr. Eager is a member of the Jefferson City Council. You are a consultant to Mayor Stone.

Dr. Eager made the following three arguments during a recent TV interview: First, Mayor Stone's proposal for reducing crime by increasing the number of police officers is a bad idea. Dr. Eager said "it will only lead to more crime." Dr. Eager supported this argument with a chart that shows that counties with a relatively large number of police officers per resident tend to have more crime than those with fewer officers per resident.

Second, Dr. Eager said "we should take the money that would have gone to hiring more police officers and spend it on the XYZ drug treatment program." Dr. Eager supported this argument by referring to a news release by the Washington Institute for Social Research that describes the effectiveness of the XYZ drug treatment program. Dr. Eager also said there were other scientific studies that showed the XYZ program was effective. (continued on next page)

Third, Dr. Eager said that because of the strong correlation between drug use and crime in Jefferson, reducing the number of addicts would lower the city's crime rate. To support this argument, Dr. Eager showed a chart that compared the percentage of drug addicts in a Jefferson zip code area to the number of crimes committed in that area. Dr. Eager based this chart on crime and community data tables that were provided by the Jefferson Police Department.

Mayor Stone has asked you to prepare a memo that analyzes the strengths and limitations of each of Dr. Eager's three main points, including any holes in those arguments. **Your memo also should contain your conclusions about each of Dr. Eager's three points, explain the reasons for your conclusions, and justify those conclusions by referring to the specific documents, data, and statements on which your conclusions are based.**

Document 2: Investigator's Memo

This is a memorandum written by a private investigator hired by Mayor Pat Stone to look into any possible connections between Dr. Eager and the XYZ drug treatment program.

Document 3: Newspaper Story

This is an article in the local paper, Jefferson Daily Press, entitled, "Smart-Shop Robbery Suspect Caught: Drug-Related Crime on the Rise in Jefferson." The article describes a robbery that occurred at a Smart-Shop store where the suspect was arrested within hours of it being reported by the owner. According to the article, the suspect appeared to be "high on drugs he had purchased with some of the money taken from the store."

Document 4: Police Tables

Two tables are presented from the Jefferson Police Department. They provide data for the city's five zip code areas. Table 1 presents crime statistics: percentage of adults who are drug users; number of robberies and burglaries; number of residents; and number of robberies and burglaries per 1,000 residents. One sees that as the percentage of drug users increases, the number of robberies and burglaries increases; thus it appears that Dr. Eager may be correct. However, if you look at the percentage of drug users against the number of robberies and burglaries per 1,000 residents, you see that there is no relationship.

Table 2 presents demographic characteristics: percentage of offenders living in Jefferson who are drug users; and percentage of residents who are college graduates.

Document 5: Report on XYZ

This is a research brief from the Washington Institute for Social Research titled, "XYZ drug treatment works in Clarendon." It highlights the effectiveness of the XYZ drug treatment in the small city of Clarendon.

Document 6: Crime Statistics

This figure comes from the State of Columbia's Department of Public Safety. It looks at crime statistics by county for the year 2000. There are 53 counties in Columbia. The figure plots the relationship between the number of police officers per 1,000 residents in a county (y-axis) against the number of robberies and burglaries per 1,000 residents (x-axis). Overall, there is a positive relationship.

Document 7: Dr. Eager's Chart

This is the chart that Dr. Eager used during the TV interview to show the relationship between the number of crimes committed and drug use in Jefferson. The chart is based on data that were provided to Dr. Eager by the Jefferson City Policy Department. Specifically, the chart was created from the data in Table 1 of Document 4.

Document 8: Research Abstracts

This document contains three research abstracts gathered from an online search where the search terms are: drug prevention, success, XYZ Drug Treatment Program. After reading the three

research abstracts, students might point out specific strengths and weaknesses (i.e., in research design) in each of the three studies.

This section provides an in-depth look at Question 1 of Crime Reduction. Here, we provide actual student responses to Question 1 from students who took the Crime Reduction Performance Task online as part of the CLA. These student responses represent different levels of performance (high, moderate, and low) as well as the characteristics of these responses that qualify them for a particular level. We did not modify the student responses for content or length, nor did we make edits for spelling or grammar.

Question 1

Mayor Stone has asked you to evaluate each of Dr. Eager's three main points. The Document Library on the right side of the screen contains materials that you should use in preparing your analysis of Dr. Eager's points. Please take a few minutes now to skim through these documents.

Document 6 contains the chart Dr. Eager used to support the claim that Mayor Stone's proposal for reducing crime "will only lead to more crime." Do you agree or disagree with this statement? Use the box below to explain why you reached this conclusion. In other words, why do you believe Dr. Eager's statement regarding this matter does or does not make sense? Be specific as to the strengths and limitations of Dr. Eager's position on this matter and the information in the documents (and any other factors you considered) that led you to this conclusion.

Central Aim of the Question

The question is trying to ascertain whether the student agrees or disagrees with Dr. Eager's statement that hiring police will only lead to more crime. To be correct, the student should disagree with Dr. Eager on this point. Why? The main concept here is correlation versus causation. Can the student distinguish between the two concepts? The contention that communities with more police have more crime is specious. It implies that police cause crime. It is more plausible that communities with more crime have hired more police to deal with the problem. You cannot draw anything conclusive from Dr. Eager's chart (Document 6); you can not know anything with certainty simply based on the chart. A student might argue that the points on the plot are too scattered to infer any linear relationship – this is incorrect.

Analytic Scoring Items for Question 1

Below are the four analytic scoring items for Question #1 from the Crime Reduction score sheet. These items are accompanied by explanations, also below.

1. Agrees with Eager or asserts that more police cause more crime
   __ YES __ NO
2. Suggests that more crime **might** necessitate more police
   __ YES __ NO
3. Says correlation does not mean cause (or causality could go either way)
   __ YES __ NO
4. Says a third variable could cause both crime and police to be correlated
   __ YES __ NO

Item 1
The scorer checks this item if the student agrees with Dr. Eager on this specific point (the relationship between crime and police). If the student agrees with Dr. Eager on this point, this is **incorrect**. This should raise a red flag as it indicates that perhaps the student does not correctly understand correlation and causation.

Item 2
The scorer checks this item if the student does not agree with Dr. Eager because more crime might necessitate more police. "Might" is a key word here; the student should express uncertainty rather than a certainty in the explanation.

- Some strong responses: "a more likely explanation might be" or "this could be the cause"
- Some weak responses (these are ones stated with certainty): "obviously this is what happened" or "clearly"

Item 3
The student must capture the intent, even if the exact words "correlation does not mean causation" are not used. It is important to emphasize intent because students may not always use the correct technical terminology of the concept that they are trying to express (such as "correlation"), but they can express this concept adequately.

- Example of intent expressed: "Two things might go together, but this one doesn't lead to the other"

Item 4
This item recognizes the instance where the student proposes a third variable not covered by the documents. It allows him/her to entertain different, feasible possibilities. The third variable suggestion must make sense. At the most basic level, the student might just reference the possibility of a third variable. At an even higher level, the student might provide an explanation for including this third variable.[30] It is important to distinguish between a third variable that might explain the cause for both crime and police to be correlated versus an explanation that describes why more police might cause more crime or more crime might cause more police.

- An example of a third variable: Wealthy communities can afford to hire more police and also attract more crime

[30] It is noteworthy if your student even recognizes a third variable. It happens infrequently.

High Quality Response and Key Characteristics

I do not agree with Dr. Eager's claim that Mayor Stone's proposal for reducing crime "will only lead to more crime." His only support for the claim hinges on the document 6 chart that shows a weak correlation between the number of police officers per 1000 residents and the number of robberies and burglaries per 1000 residents. However, Dr. Eager is mistaking correlation for causation and failing to understand the alternate explanations for such a correlation. More than likely higher volumes of robberies and burglaries per 1000 residents are occurring in concentrated urban areas or poorer neighborhoods with crime problems. As a result more officers will naturally be allocated to these areas rather than to other areas with low crime rates. However, that does not mean that the increase in police officers in these areas is causing the extra crime. By only observing correlation and not examining the underlying circumstances, Dr. Eager is being shortsighted in his analysis. If anything the problem is that even though more police officers have been allocated to high crime areas, these problem areas still simply do not have enough police personnel to adequately deal with the problems. As such Mayor Stone's proposal possesses merit that Dr. Eager's claims fail to observe.

Characteristics of this high quality response:

- Evaluates the evidence
- Provides analysis and synthesis of the evidence (e.g., understands correlation versus causation and suggests an alter native reason for the relationship between crime and police officers)
- Draws appropriate conclusions (e.g., there is not necessarily a causal relationship between the variables displayed on the chart)
- Writes with clear organization and the response is easy to follow
- Shows strong command of writing mechanics

Moderate Quality Response and Characteristics

While it seems strange to say that an larger police presence will in fact lead to higher crime. In the case of Dr. Eager's argument, there in fact may be a valid point in that a higher police presence may address short term issues such as arresting the criminal who commits a robbery or burglary but may not take care of the long-term problem as to why that person commits that crime in the first place. In the case of document 6 which is the crime rates and police officers chart. There does appear to be a correlation between the number of police officers and the number of crimes committed. However, this chart can be misleading as it doesn't take into account other factors that wuold be important to consider in an issue such as this one. For example, the chart doesn't taken into account where these crimes are being committed and what the police presence is in those areas. It could be argued that the higher police presence is in response to a rise in crime in a particular area. We do not have any idea how long the crimes have been going on nor see the effect of having more police officers in one area does to that area's crime rate. The graph also doesn't take into account that higher population areas would have higher a higher number of police officers and a higher crime rate. This graph combines all the counties and creates this one standard in which areas with a small number of police

officers, which probably would have lower crime rates along with lower populations are made to appear that fewer police officers leads to fewer crimes. This graphs takes these numbers out of context and makes an extremely flawed argument that if taken into practice would lead to extremly detrimental results. That's why Dr. Eager's statement about more police leading to more crime is flawed and it presents an opportunity for the mayor to counter the Doctor's argument.

Characteristics of this moderate quality response:

- Evaluates the evidence
- Provides analysis and synthesis of the evidence (e.g., understands correlation versus causation and suggests other possible factors leading to the relationship, such as population or the possibility that the higher police presence is in response to a rise in crime in a particular area)
- Shows moderate command of writing mechanics (e.g., fragments)

Low Quality Response and Characteristics

I understand Dr. Eager's statement about crime. It is a valid statement that makes sence. Jefferson does appear to have a high percentage of crime rates caused by drug addicts. A successful drug treatment program would lower the crime rate, however, I believe that crime will always be out there. No matter what a city, state or country does, crime will always exist. Drugs and crime are always a bad combination. In this case, the charts report the greater the population using drugs, crime was on the rise. There are many great programs out there that will treat drug abuse; hence, a cut in crime rates. When they are appropriatly funded they are statistically proven to work. The university research abstracts conclude that 27% of people dropped out of the XYZ Drug Treatment plan, whereas 30% dropped out from the I Can plan. There were fewer arrests for those that completed the XYZ plan.

Characteristics of this low quality response:

- Accepts the document as it is (without critique) and does not interpret the information correctly (e.g., agrees with Dr. Eager, thus confusing correlation with causation)
- Interjects response with personal opinion, often without supporting evidence
- Interjects response with other information, though it is unclear why this information is presented (It should be noted that in the subsequent responses to the two other questions in this task, the student continues to use not always the most relevant or significant document to support statements)

Other characteristics of low quality responses:

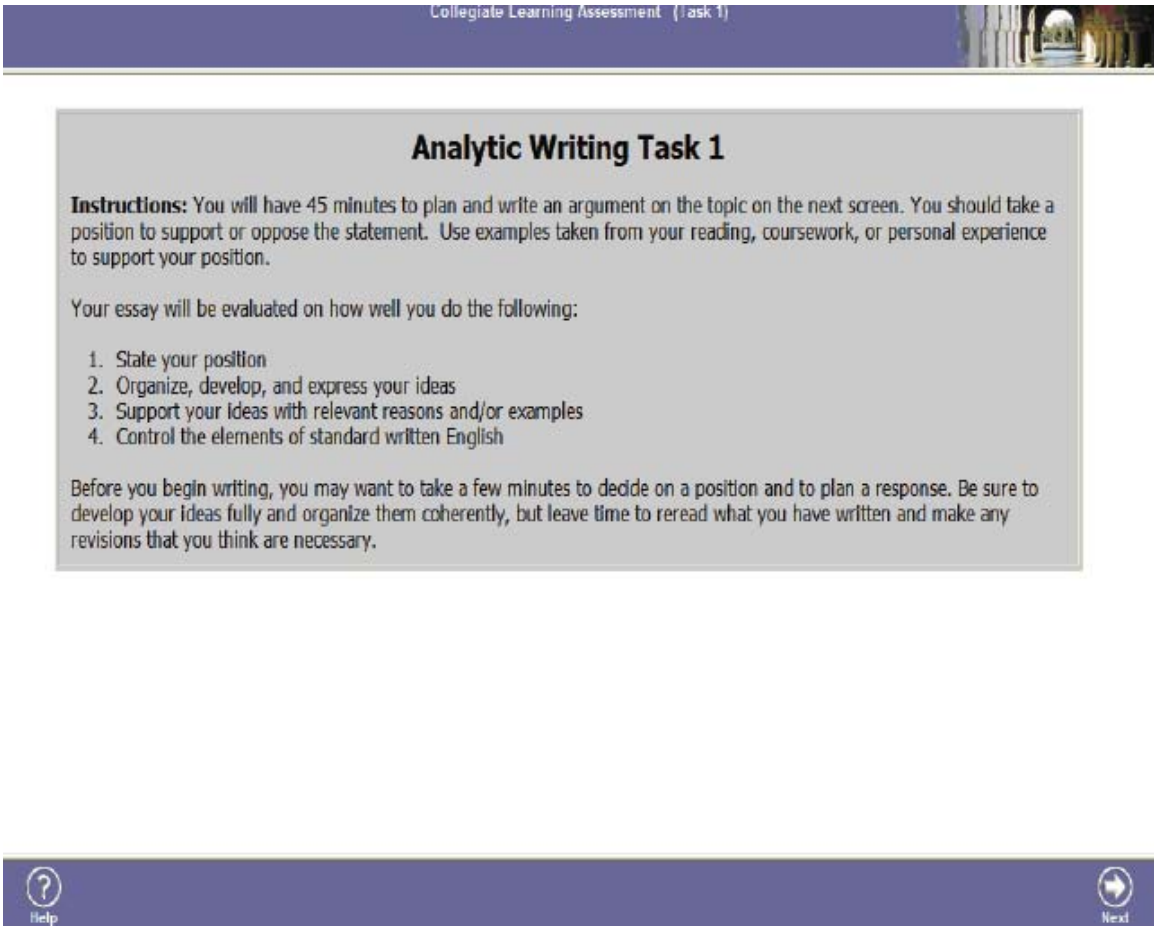
- Blank, extremely brief (i.e., only a few words or one sentence), unintelligible, or completely off topic answers.

MAKE-AN-ARGUMENT: GOVERNMENT FUNDING

In this section, we present a Make-an-Argument prompt called "Government Funding," sample responses at different levels of performance (high, moderate, and low), and characteristics of responses at each of those levels.

Introduction

Students are provided with the following instructions when taking Make-an-Argument:



Prompt

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

Scoring Criteria

In addition to consideration of the scoring criteria outlined earlier in this document, each Make-an-Argument response is assessed specifically on general logic, argumentation, and analytic writing skills:

- Clarifying a position and supporting it with evidence

- Considering alternative viewpoints or counter points to their argument
- Developing logical, persuasive arguments
- Depth and complexity of thinking about the issues raised in the prompt
- Level and sophistication of vocabulary, sentence structure, and grammar
- Organization and flow of information presented
- Generation of reader interest, provocation of thought
- Use of examples

Students can argue either side of the argument. Students can also argue that both have merit or neither has merit. No penalty is given for the perspective they take; however, they are expected to take a clear position on the issues in the prompt and support it.

<u>High Quality Response and Key Characteristics</u>

Government imposes order upon its citizens to pursue generally agreed-upon goals in society. An important function of American government, for example, is to protect the "life, liberty and the pursuit of happiness" of its citizens, a premise upon which the U.S. was founded more than two centuries ago. Guaranteeing this "inalienable right" through government action is easier said than done. In general, government does so by collecting taxes, enacting laws, and enforcing laws consistent with goals. Violating these laws, by definition, are crimes and the people who commit crimes are criminals. But the meaning of laws and the causes of crime are complicated. In all, there is no simple formula for investing taxpayer dollars and the statement oversimplifies the challenge of dealing with crime. While investing public dollars in crime prevention may have certain advantages, it is not necessarily "better spent" than "dealing with criminals after the fact."

Laws are reflections of moral beliefs of society, that is, what we collectively believe to be right or wrong. These beliefs often change over time, and even by communities within broader society. Furthermore not all laws, or crimes, receive the same levels of enforcement. For example, while we might universally agree that certain violent acts (e.g., murder, rape, armed robbery) are indeed crimes that ought to be prevented at high dollar cost, we might not agree that others (e.g., underage drinking, jaywalking) deserve the same attention. And certain laws which may have been important at the time or in the jurisdiction where they were written, they may no longer be relevant, although they may remain on the books. Given different interpretations, severity and changing nature of crime, it might be quite difficult (and costly) to create a program that effectively prevents crime in all its variety. Doing so would run the risk of addressing those crimes that either do not pose significant threat to "life, liberty and the pursuit of happiness" or, in the future, are no longer crimes at all. By contrast, dealing with criminals after the fact has the advantage of focusing resources on those who have indeed violated existing laws in society, in particular those laws society has chosen to enforce. This approach also allows society to reconsider laws for relevance in present-day society (i.e., through the courts) as violations occur, so that criminal behavior may be redefined as concepts of morality may change.

Furthermore, preventing crime requires that we understand why crimes occur, so that we may know how to intervene. But crime is complex, stemming from many, many conditions pertaining to society and its members. These factors may divide along lines of the classic debate in biology over "nature vs. nurture" as determinants of behavior. Interpreting

crime in this way, we might ask: Are criminals the result of the influence of their environment? Or are criminals born to commit crimes? If criminals are products of their environment, then crime prevention programs should address root causes of crime in society. But what are these root causes, and can they be disentangled from a combination of other factors? Are all people susceptible to the same causes, or does a crime prevention program need to accommodate all individual differences so that none will become criminals? Investing in a comprehensive crime prevention program that addresses all causes and all individuals would appear to be a costly proposition. It is difficult to imagine a program that could effectively do so, at any cost. Furthermore, addressing a root cause of crime would likely trigger a series of other causes that would need to be addressed. If, for example, robbery is related to high incidence of poverty and drug abuse, then crime prevention requires effective programs to address problems of poverty and substance abuse. But these, too, are complex problems related to issues of education, discrimination, mental health, and so forth. Where would the crime prevention program (and government investment) stop? By contrast, according to the "nature" argument, criminals are social deviants from birth. Addressing crime becomes a simple matter of identifying these individuals and removing them from society according to the crimes they commit, without any need to address social or environmental concerns. So long as the number of criminals is few, the cost of separating these individuals from society (e.g., by sending them to prison) will also be relatively small, and government funding might be "better spent" on this approach.

But my understanding is that the "nature vs. nurture" argument rages on, leading me to believe that neither determines an individual's behavior by itself. Sending individuals to prison, because they were born criminals, assumes that these people cannot become productive members of society. It denies these individuals their own "inalienable right," a reason many have come to the America in the first place. Whether or not this is the case, keeping these individuals imprisoned assumes further that laws, and therefore the definition of crime, never changes. Unjust imprisonment in the name of dealing with criminals can never be government funding "better spent" in the United States.

Neither investment in crime prevention nor investment in dealing with criminals by themselves can easily address the problem of crime in our society. Instead, some combination, along with investments in other societal improvements will be required to address problems of crime. More generally, how government funding should be spent to address the complex challenge of protecting citizen's rights to "life, liberty and the pursuit of happiness" is best determined by the continued interaction of lawmakers, law enforcement officials, the courts, and the citizenry, just as it has for more than 200 years.

<u>Characteristics of this high quality response:</u>

- Clearly elucidated thesis
- Well-organized
- Sophisticated use of vocabulary and mechanics
- Sophisticated, in-depth treatment of the issues
    - Acknowledges and discusses issues on both sides of the prompt
    - Raises uncommon points (e.g., the changing conception of crime)

- Clarifies the different meanings and purposes of key terms (e.g., government, crime, prevention)
- Supports points with helpful examples
- Applies concepts from their education (e.g., nature vs. nurture, laws are reflections of societal moral beliefs)
- Considers the consequences of their suggestions
- Logically developed; each idea builds upon the last

Moderate Quality Response and Key Characteristics

Government funding would be better spent on dealing with criminals after the fact as opposed to investing in programs intending to prevent crime. I say this because there will always be those who can outsmart the government. Criminals will always find new opportunities and means to commit criminal acts, even though the government will win occasional battles in the war on crime,.

Technology plays a central role in this ongoing battle between government and criminals. New weapons and tools in particular, increase the capabilities of those who commit crimes. Often these weapons and tools are more readily available to criminals than to the crime fighters! For example, criminals armed with so-called "assault rifles" enjoy a distinct advantage over cops who are not allowed to carry them. In some ways, our system of government hinders our ability to defend our society against clever and well-equipped criminals. While our system does change over time—police are now allowed to carry more powerful weapons in some locales— change occurs slowly.

Government can never get far enough ahead of criminals to anticipate criminal behavior and prevent crime because criminals will also have better weapons than crimefighters, Thus investing in crime prevention cannot be the best use of government funding. Instead, government funding should be spent on dealing with criminals after the fact.

Characteristics of this moderate quality response:

- Clear but limited thesis that focuses on a narrow aspect of crime (technology and weapons of criminals)
- Sentence structure is unvarying (subject, verb, object)
- Some arguments are unclear, or not clearly related to thesis. For example:
  - In paragraph 3, how do criminals having better weapons make it hard for police to anticipate them or their crimes?
  - Argument about criminals and technology is not clearly related to how funding should be spent. It is not clear whether or not the writer is suggesting that if police are equipped with better weapons, they will be better able to defend society from criminals
- Does not attempt to counter potential objections to the argument (e.g., the greater resources of the police force relative to a single individual criminal)

Low Quality Response and Key Characteristics

Crime is a huge problem around the globe, and mostly in America. Crime effects our everyday lives more than we even know and is the black hole into which billions and billions of dollars are sunken into each year on security products for people, and legal and justice cost.

About security products, we buy expensive alarm systems for our homes, bars for our windows, locks for our doors. We hire security guards to patrol our neighborhoods. If that was not enough, we store our valuables in banks and rent safety deposit boxes! And when we put "decorative" bars on our windows and fences around our yards, nobody will want to buy a home that needs so much security!

We carry mace in our purses, whistles on our key rings, we plan our schedules and routes to work to avoid certain neighborhoods. We get escorts to our cars in parking lots after dark. All of this costs money. And the only ones benefiting from all of this are the manufacturers of the products, the security guards and the lawyers.

Our great country deals with this dilemma in almost backwards fashion. The government could easily use these billions and billions of dollars spent on people stealing bread for there family to eat, to just provide bread for the families so they won't have to steal. Also, the funds could be used to develop programs in which people are trained to get jobs. Instead of force people into crime and forget about them, the U.S. government should uplift its own people into something greater, so that allot of these issues and crimes would cease to exist.

Characteristics of this low quality response exemplified in this sample:

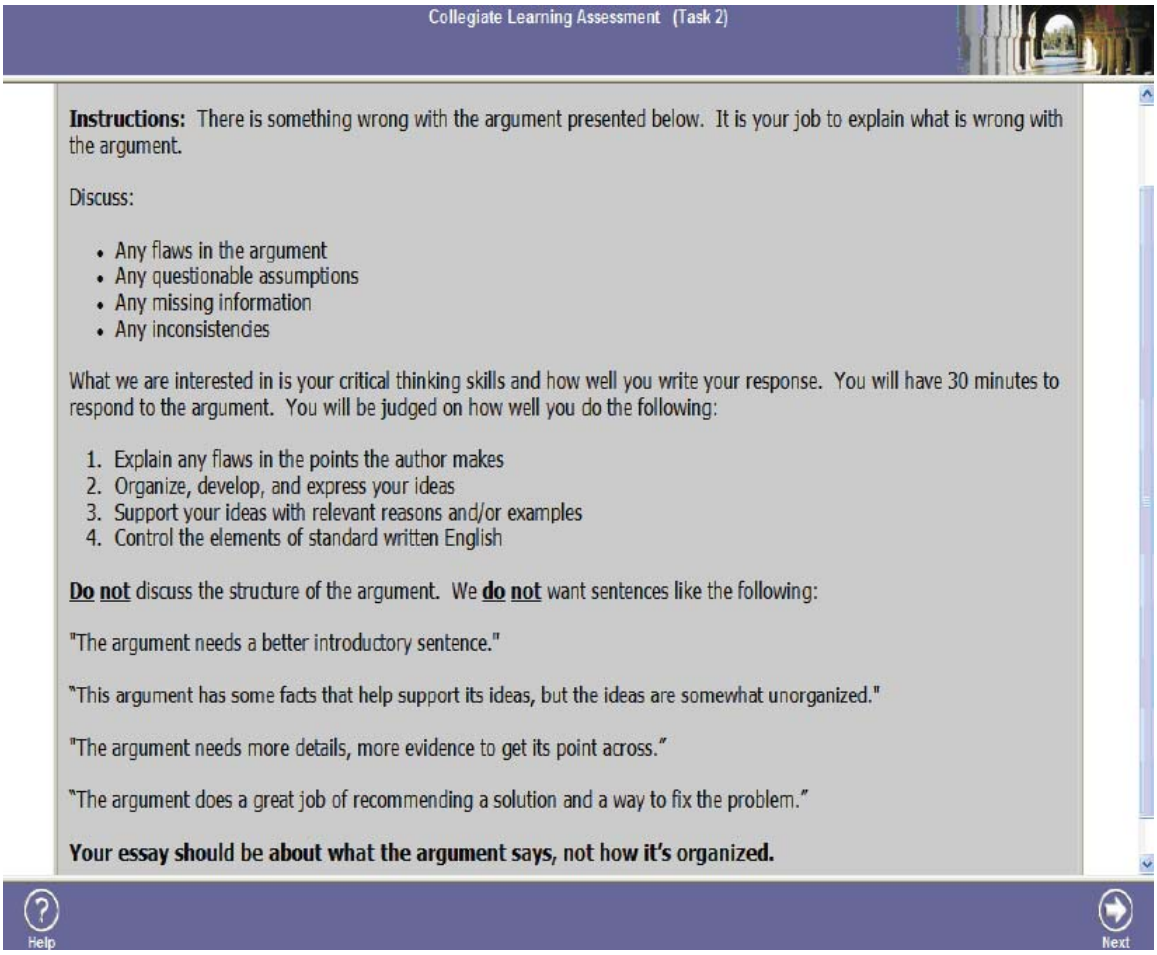
- Thesis is undeveloped
- Writing is adequate, but contains awkward constructions and mistakes in vocabulary and tense
- Does not address the main issues in the prompt
    - Argument is largely about our fear of crime
    - Never takes or supports a position about prevention vs. dealing with criminals
- Uses some good examples (e.g., bars on our windows), but largely to support our fear of crime
- Critical thinking is poor
    - Unclear why manufacturers, guards, and lawyers are the only ones to benefit from security devices
    - Does not try to counter the position that security devices can be effective in crime reduction
    - Opening contention is hyperbolic (billions and billions of dollars spent on security devices)

CRITIQUE-AN-ARGUMENT: WEDDINGS

In this section, we present you with a Critique-an-Argument prompt called "Weddings," sample responses at different levels of performance (high, moderate, and low), and characteristics of responses at each of those levels.

Introduction

Students are provided with the following instructions when taking Critique-an-Argument.



Prompt

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

good advice for young couples to have short engagements and choose a month other than June for a wedding.

Scoring Criteria

Each Critique-an-Argument response is assessed on the holistic scoring criteria (e.g., critical thinking, writing) as well as recognition and explanation of specific logical flaws in the argument. The logical flaws are prompt-specific; however, they cover a variety of common critical thinking concepts. For this prompt, some examples of logical flaws include:

- Number and proportion are not the same thing
  - The population and hence the number of weddings are growing, so the increase in the number of divorces may simply reflect an increase in population, and nothing more
  - A more appropriate measure is the proportion of marriages that end in divorce now compared to the past, or the proportion of June weddings ending in divorce compared to the proportions of weddings in other months that end in divorce
- Correlation is not causation
  - Getting married in June may not cause people to get divorced
  - June weddings may not cause long engagements
  - Long engagements may not cause divorce, even if June weddings do cause divorce

High Quality Response and Characteristics

There are several problems with this author's argument for avoiding divorce by shortening engagements and avoiding June weddings. One problem is that just because the number of divorces is going up, divorces are not necessarily a bigger problem now than they were last year or the year before. Every year there are more people in the United States (and on the planet) so that means that each year there are more marriages and probably more divorces. If the number of divorces goes up and the number of people on the planet also goes up by the same amount, then it means that the percentage of divorces would be the same. The writer doesn't tell us whether the percentage of divorces has gone up, down or stayed the same.

The author assumes that because so many divorces are from June weddings, it means that June weddings cause the divorces, or make the divorces more likely. Because we don't know whether the percentage of divorced couples has gone up, down or stayed the same, we don't know if divorces are more, less, or equally likely to happen these days. If more weddings happen in June (because as the writer points out, June weddings are so popular) we might also expect more divorces from weddings in June. If, for example, 80 percent of weddings happen in June, then we might expect 80 percent of divorces to happen to people who were married in June too. If the author is correct that 1 in 3 marriages end in divorce, then it may be the case that 1 in 3 June weddings end in divorce, 1 in 3 February weddings end in divorce, 1 in 3 July weddings end in divorce and so on.

Another problem is that the writer assumes that couples end up being engaged for a long time just so that they can get married in the summer months (like June). But couples might be engaged for long periods of time for a lot of other reasons too. For example, couples might stay engaged for a long time so that they can get to know each other better, and not rush into something too quickly. Or maybe they have lengthy engagements because weddings take a long time to plan. Both my parents and grandparents had long engagements and were married in winter, so clearly not all people are having a long engagements just so they can wait to get married in the summer months. Furthermore, my parents and grandparents both married young and are still married, probably because of the greater understanding for one another that they developed during their engagement. If this is true, then the writer's argument that couples should have short engagements to prevent divorces may not be justified.

The last problem that I see in the paragraph is that the author argues that avoiding June weddings will prevent divorce. But simply changing a wedding to May or July or any other month should not have any affect on whether or not a couple gets divorced. Divorce is caused by many complex issues in a relationship including communication, love, caring, respect, supportiveness, compromise, compatibility, and above all hard work at maintaining the relationship. If a couple wants to try to prevent getting divorced, they should work on these things, not simply avoiding a June wedding as the author suggests. My brother is divorced. Yes, he was married in June. But in my opinion the date of their wedding was the least of their problems.

Characteristics of this high quality response:

- Information is well organized. The reader knows exactly which part of the prompt is being critiqued at every point in the response
- Uses complex sentence structure and varied vocabulary
- Respectful and appropriate tone
- Uses examples (e.g., reasoned hypothetical examples and common knowledge) to support and illustrate valid points
- Identifies numerous flaws (complex and subtle)
- Explanation/justification: The response not only mentions numerous flaws throughout the argument, but also explains the flaws clearly, completely, and convincingly for the reader
- Demonstrates solid understanding of several important critical thinking concepts. For example:
    - The difference between interpreting proportions versus just raw numbers in statistics and how doing so can lead to different conclusions
    - Correlation is not causation

Moderate Quality Response and Characteristics

At first glance the paragraph that couples should avoid June marriages sounds well grounded in factual evidence. However, there is no information provided for the total number of weddings in each month. Minus the statement that June weddings are "popular" how can you tell if those June weddings are more common than May weddings? Or August weddings? The article implies that more weddings in June end up in divorce. Well, if there are twice as many June weddings, which seems to be supported

by that June is the most desirable month, then one can reasonably assume that there will be twice as many June weddings that end in divorce as well. We cannot conclude, from the data or arguments that being married in June ends up in divorce any more than being married in other months.

The argument for the shortening of engagements is also flawed. Short engagements likely mean less time to think about the decision of marriage. How can this be a good thing when ultimately the argument is for avoiding divorce? The paragraph seems to say that at people must avoid June weddings and that somehow length of engagement matters too. What if the couple gets engaged in April? Should they hasten their plans and get married in May to avoid the dreaded June? The paragraph suggests that doing so is better than waiting until July, or longer. What is the right amount of time to be engaged in order to avoid divorce? What is the best month to get married? Given differences among people, and therefore couples, and a lot of other factors, I think it depends on many things. But we can't conclude from the information or argument given that the answer is brief engagement leading to a wedding in a month other than June.

Characteristics of this moderate quality response:

- Writing is clear and somewhat organized
- Makes some substantive points
    - Divorce rates between years and months cannot be compared without knowing the total number of weddings per month
    - Notes the logical flaw with having brief engagement periods, and highlights with an extreme example
- Barely touches on other flaws
- Mentions the complexity of marriage and how what is right for one couple may not be right for another couple. Does not develop this point at all
- Points are partially, but not fully developed. The use of rhetorical questions and hypothetical examples is somewhat effective at illustrating their point; however, the rhetorical questioning is overused. The response would benefit from use of more varied examples to support points, and greater development of points

Low Quality Response and Characteristics

MY BROTHER GOT MARRIED LAST JUNE. I WAS THE BEST MAN, BUT I DON'T KNOW WHETHER THEY SHOULD HAVE A JUNE WEDDING AGAIN OR NOT. WE HAD A GREAT PARTY AFTERWERD, SO IT WAS STILL A LOT OF FUN DANCING, BUT I AGREE THAT JUNE WEDDINGS AREN'T A GOOD IDEA. OTHER MONTHS THAT ARE COOLER WOULD BE BETTER FOR DANCING. I THINK THAT MY BROTHER AND HIS WIFE HAVE A GOOD MARRIAGE, BUT THEY HAVE ONLY BEEN GOING OUT FOR A YEAR.

Characteristics of this low quality response:

- Lack of content: No critical evaluation of the logical argument presented. Appears to not fully understand how to critically evaluate an argument
- Writing is simple: short sentences, basic vocabulary

Other characteristics of low quality responses:

- Misses the purpose of the critique (e.g., "this paragraph needs a comma after the third word in the first sentence")
- Attitude/tone of writing: emotional, flippant and insulting
- Some statements are inaccurate
- Poor command of written English
    - Lengthy run-on sentences
    - Statements are poorly structured

Table 1.1
Summary of Task Differences

	Task Differences		
	Make-an-argument Tasks	Break-an-argument Tasks	Performance Tasks
Task complexity and real-world fidelity	Moderate	Moderate	High
Variety of acceptable responses	High	Low	Low
Logic, reasoning and critical thinking	Moderate	High	High
Evaluation of specific facts and information	Low	High	High
Generation of factual information from personal knowledge	High	Low	Low
Synthesis of disparate pieces of information	Low	Low	High
Basic clarity of written communication	High	High	High
Persuasive writing and advanced vocabulary, grammar, sentence structure, organization and interest	High	Moderate	Low

6. INTRODUCING CLA EDUCATION: BRIDGING THE GAP BETWEEN ASSESSMENT AND TEACHING/LEARNING

Marc Chun[31]

INTRODUCTION

This monograph has presented the logic of the Collegiate Learning Assessment, then showing how this logic relates to the core of education (namely teaching and learning). The previous five chapters have set up how the CLA can be establish a continuous cycle of improvement. The role of CLA Assessment Services is one half of this process, and the chapters to date have documented in a conceptual manner how the CLA has reconfigured the balance between assessment and accountability, how the CLA can provide a diagnostic bridge between the classroom and the institution, the crucial role of faculty, and deep exploration into the nature of the CLA measures. We now turn to the second half of the process, which is the practical, specific means by which CLA Education makes this happen.

Background

In 1990 the U.S. Department of Education stated as a goal that "the proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially." Despite such an important charge, a generation later this effort is still ongoing. The children born in 1990 might be preparing to go off to college in 2008, and yet then-presidential candidate Barack Obama stated that "We'll teach our students not only math and science, but teamwork and critical thinking and communication skills, because that's how we'll make sure they're prepared for today's workplace."

Consistent with Obama's claim that critical thinking and problem skills are needed in the workplace Halpern (1993) found that "virtually every business or industry position that involves responsibility and action in the face of uncertainty would benefit if the people filling that position obtained a high level of the ability to think critically."

The importance of critical thinking skills has also been recognized by national educational associations as well. The National Assessment of Educational Progress (commonly referred to as NAEP) "results suggest that although basic skills have their place in pedagogy, critical

[31] Marc Chun is the director of CLA Education.

thinking skills are essential" (Ballantine and Spade, 2007). And a report from the American Association of Universities ("Standards for Success"[32]) indicated that habits of mind (critical thinking, analytic thinking and problem solving) were essential for college success.

Although calls to action from the federal government, employers and national associations remind us of the importance of these skills, it is of course college and university faculty who should have the most important voice in this conversation. And in fact according to then-president of Harvard University Derek Bok (2006) in *Our Underachieving Colleges*, national studies have found that more than 90 percent of faculty members in the United States consider critical thinking the most important goal of an undergraduate education.

Given this unanimity of commitment (especially driven by faculty) to ensuring students have critical thinking skills one might reasonably expect that college graduates would develop these core abilities. But do they?

Turning first to college students, a 2006 American Institutes for Research (AIR) study found that 75% of two-year college students and 50% of four-year college students did not perform at proficient levels of literacy (here, meaning being unable to compare credit card offers with different interest rates or summarize the arguments of newspaper editorials).

More troublesome concerns arise when we look to college graduates. A 2005 National Center for Education Statistics study found that only 31% of college graduates could read a complex book and extrapolate from it. Parallel patterns are reported in the aforementioned AIR study that found that 20% of college graduates had only basic quantitative skills (operationalized as being unable to calculate the total cost of ordering office supplies, compare ticket prices, or calculate the total price of a salad and sandwich on a menu).

So, given that the federal government, employers and, most importantly, faculty are behind the charge to equip students with critical thinking and other higher order skills, how is it the case that students are not developing them by the time they graduate?

What We Know from the Literature

One possibility is that we don't have a standardized, coherent theory of practice about the best way to develop these higher order skills. In Pascarella and Terenzini's (2005) comprehensive, meta-analysis of the literature on higher education student development, they

[32] Visit www.s4s.org for more information.

summarize the evidence mounted to date about how to explain within-college effects on the acquisition of cognitive skills (with specific discussions of developing critical thinking skills). With respect to critical thinking, what we know from the literature boils down to just four points.

First, major field of study had little consistent connection to critical thinking, although they found that different fields of study led to the development of different reasoning skills. The main conclusion here could be that critical thinking skills are not the province of any one academic domain, and taken another way, it is a fair expectation that all students should be able to develop such skills regardless of field of study.[33] (Pascarella and Terenizini do note that the research suggests that exposure to natural science courses positive influence growth in critical thinking.) And overall, curricular experiences that require cross-disciplinary integration enhanced critical thinking.[34]

Second, there are particular student behaviors (outside of the classroom) that are associated with the development of critical thinking skills: namely, the use of computers, hours studied, number of non-assigned books read, writing experiences, library use, and specific course-learning behaviors have positive effects on critical thinking. Also, student-faculty non-classroom interactions were also found to have a positive effect. Research using the National Survey of Student Engagement (NSSE) has affirms the theoretical claim that student engagement is positively associated with student development.

Third, interactions with other students (particularly those who come from different backgrounds or have different points of view) have a positive effect. Interactions with peers that extend and reinforce the academic experience and opportunities to interact with those who are different are positively associated with critical thinking skill growth; involvement in diversity experiences and service learning do as well.

Taken together, these first three points indicate that critical thinking skills do develop across majors, when students are engaged, and when students come face-to-face with difference. But can it be taught?

There is good news and bad news. As we might hope and expect, the fourth point found in Pascarella and Terenizini's review of the literature is that critical thinking can indeed be taught. Unfortunately, however the observed effects have been quite modest. Pascarella and Terenzini

[33] That is, it is not only students who earn degrees in or take courses in philosophy who should develop critical thinking skills.
[34] As will be discussed below, performance tasks often embrace such cross-disciplinary approaches.

note that this may stem in large part due to differences in definitions of what constitutes "teaching critical thinking." As noted, with no shared conception of what pedagogical and curricular practices constitute such instruction, it is difficult to measure this. Further, since most of this research used multiple-choice or student self-reports to measure critical thinking, the research was hampered in it's ability to measure the key dependent variable.

So what does the research literature tell us? The preponderance of evidence about how to improve critical thinking skills falls into the first three points -- and much of that is largely beyond the scope of what happens in the classroom, and thus, beyond the scope of what faculty can do. They key is the fourth point; evidence does suggest that it's possible to teach these skills, but we must be more clear about what that means and how to do that. CLA Education (to be discussed below) seeks to be a practical response.

AUTHENTIC ASSESSMENT AND PERFORMANCE TASKS

Both CLA Assessment Services and CLA Education share a common starting point: the key to our work in assessment as well as the development of curriculum and pedagogy in this area is the *demonstration* of critical thinking and other higher order skills. As noted previously by Chun (2002, 2006), we often are left "looking where the light is better," finding simple and convenient proxies for these skills that faculty agree are most important, and that given the "iron triangle" of assessment that seems to require that we choose between faster, better and cheaper forms of assessment, all too often better is sacrificed for fast and cheap. The result often has been assessment that serves as a compliance exercise rather than a means to improve teaching and learning.

Authentic assessment and performance tasks provide a means of bridging the chasm between institutional assessment and classroom practice; between the summative and the formative.

The authentic assessment movement informs the work of the CLA. Grant Wiggins (1990), considered to be the guru of authentic assessment, noted:

> Assessment is authentic when we directly examine student performance on worthy intellectual tasks. Traditional assessment, by contrast, relies on indirect or proxy ‘items’ -- efficient, simplistic substitutes from which we

> think valid inferences can be made about the student's performance at those valued challenges.

Wiggins argues that authentic assessment presents students with an array of tasks that reflect best instructional activities: writing, doing research, engaging in oral analysis, collaborating with others.[35] Further, it requires students to effectively perform using acquired knowledge, often by completing tasks that are intentionally ambiguous and not clearly structured. This presents challenges and roles that help students rehearse for the situations they will face (and doing so, gain confidence); more so, Wiggins notes that authentic assessment achieves validity when the task simulates larger world "tests" of ability. Authentic assessment challenges students to create complete and justifiable performance, answers or products, in response to a meaningful prompt. And finally, authentic assessment gains validity and reliability through appropriate scoring criteria for the product.

All CLA programs embrace these principles of authentic assessment, and this big idea of looking for the demonstration of key skills takes the form of *performance tasks* (which can be seen a specific form of authentic assessment). Michael Hibbard[36] (2000) described central features of performance task assignments or projects that are:

- Engaging (they grab students' attention)
- Activating (more student work leads to being drawn into the task)
- Authentic process (steps taken mirror those of similar performance for similar audience)
- Authentic product (product or performance similar to that of "larger world")
- Essential (connected to important—and not trivial—standards)
- Integrative (knowledge, thinking skills, problem-solving skills, and writing are utilized together)
- Embedded (used with—and not added-on-to—instruction)
- Appropriate structure (sufficient explanation such that students understand the task)
- Feasible (tasks are possible to complete given time and other resources)
- Equitable (Fair to students based on background knowledge)

[35] It is surprising, perhaps, that academics have their students complete such engaging classroom work, and yet accept standardized, multiple-choice tests to assess student learning.
[36] Hibbard writes about middle school science, but arguably the same principles are applicable to other domains.

All of these features are included in the CLA Performance Tasks.[37] The authentic nature, focus on key higher order skills, and means of delivery are all hallmarks of the CLA. Additional features that Hibbard notes (but aren't a part of the performance tasks that are used as part of CLA Assessment Services) include:

- Feedback and revision loop (a chance to make revisions)
- Group work and individual work (a balance between the two approaches)
- Promotes deeper understanding (expects students to go beyond surface-level knowledge)

These last three points are central for using performance tasks as a classroom activity to promote learning, which although is beyond the scope of CLA Assessment Services, these features are built into performance tasks created by faculty as part of CLA Education (to be discussed below).

What may seem obvious upon reviewing this list is they way that, as noted above, performance tasks can bridge the gap between assessment and classroom practice. What CLA has done is capitalize on this point. The same tools that can be used as an effective means to improve learning can also be used to assess learning. It is by connecting these two features that systematic improvement of teaching and learning can and will occur. There is a simple poetry to this stance: if we want students to develop critical thinking skills, we should have them practice higher critical thinking skills. If we have students learn and develop critical thinking skills in an authentic, practical manner using performance tasks, we can assess the growth in their skill development using performance tasks. Here, assessment and teaching/learning become one-and-the-same. CLA has taken on as an informal mantra the idea that you can teach to the test once you find the right test. Another way to frame this is that if we share a common set of outcomes, we can teach to those outcomes and we can assess those outcomes -- and accomplish both of these with the same approach.

CLA EDUCATION

[37] See the explication of the goals and underlying structure of the CLA performance tasks in Chapter 5 of this monograph.

CLA Education, a major new division of the CLA, focuses on curriculum and pedagogy; CLA Education explicitly recognizes the central role of faculty in the process of developing higher order thinking skills like critical thinking. CLA Education -- which launched some programs in 2008 and will have a full launch in the fall of 2009 -- and includes the CLA in the Classroom programs that provides faculty with curricular and pedagogical tools to help students practice and develop higher order skills. CLA in the Classroom includes the Performance Task Academies (which provide faculty members with tools for creating and scoring their own content-embedded performance measures) and the Performance Task Library (which serves as a repository for these tasks). Each will be discussed below. In the 2009-10 academic year, CLA Education is launching the Student Diagnostic Report and the Institutional Diagnostic Report, also discussed below. Other CLA Education efforts include informal gatherings at national meetings (the CLA coffee [cla]tches), as well as free web conferences and newsletters to facilitate the sharing of best practices (the CLA Spotlight and the CLA Pulse).

Performance Task Academy

The first Performance Task Academy was offered in March 2008. The basic two-day Academy includes a mixture of mini-lectures, small group activities, large group discussion, and independent work. The Academy is designed to be an introduction for faculty who have little to no experience creating performance tasks, and a chance for developing tasks for faculty who already have some practice.

The first day of the Academy focuses on authentic assessment practices, performance tasks, rubrics, and providing diagnostic feedback. The second day begins with a series of activities to help faculty learn how to put these concepts into practice, and then they spend the balance of the day creating an actual performance task that they might use in their own classrooms. They receive feedback from each other, as well as from the facilitators.

The number of participants is capped to keep activities highly interactive. Institutions are encouraged to send teams of at least two faculty members to the Performance Task Academy. To date, approximately 500 faculty members have participated in one of the 20 Academies offered in the inaugural year; these Academies have been offered across the country, as well as in Korea and Japan.

Feedback from faculty about the Performance Task Academy has been overwhelmingly positive, and perhaps revealed what faculty found to be lacking in other faculty development activities. One associate professor of psychology noted, "Generally if I come back from a conference/workshop with one small nugget of good info, I am satisfied. Here, however, I was amazed that the entire two days were fabulous. I never felt bored or wished things would move along more quickly." A professor of English noted, "This stuff gives us something concrete to work with, to analyze and evaluate. It's much better than just waving our hands and saying, "well, you know, it's about 'critical thinking'.""

Other participants' responses indicate that they appreciated the active learning aspects built into the workshops. One faculty member wrote, "The facilitators were the best I've ever had at a workshop. I was particularly impressed by the connection between teaching/learning and assessment and that there was a focus on authentic assessment and changing the culture of teaching and learning at our institution." Another commented, "The workshop was developed to ensure maximum learning. The two-day workshop consisted almost entirely of active learning. It is refreshing when presenters practice what they preach."[38]

Participants also valued the opportunity to work with other faculty. "Highly effective model that balances individual work time with group work," wrote one participant, continuing, "I've made important contacts that I will maintain as I continue to develop my performance tasks."

New workshops will be launched in the 2009-10 academic year. This will include the Advanced Task Development Academy, which will be a special four-week program, in which initial in-person training will be provided on creating performance tasks, then faculty will work for four weeks (with web-based workshops to support task development), and then another in-person workshop will be held to share and get feedback on the finished performance tasks. Additionally, a Course Development Academy will instruct faculty on how to create an entire course centered on the use of performance tasks. CLA measurement scientists are eager to examine whether critical thinking skills can be intentionally improved over the period of one course that is dedicated to that improvement.

[38] Moving forward, the plan will be to have an independent, outside evaluator observe and document the experiences at an Academy.

Finally, there has been interest among faculty about being "certified" in developing the skills to create such performance tasks; some faculty have indicated that they would be interested in including this in their tenure files. Accordingly, we are exploring how to provide such certification and to create a community of faculty members committed to using the performance task pedagogy and curricular approach to promote development of these higher order skills.

Performance Task Library

Faculty from across the country who have participated in one of the Performance Task Academies have created course projects and assignments that utilize authentic assessment approaches from the workshop. These performance tasks include a range of topics that stem from different disciplinary orientations, and often cross disciplinary boundaries (in ways that reflect how "real world" issues don't fit neatly into one or another academic domain.[39]

Some have been designed for high school classes, some for undergraduate courses, and some for graduate student programs. Some tasks are appropriate for general education courses, while others include specific disciplinary content.

In one performance task designed for an introductory biology course, the professor created a scenario where a friend was diagnosed with Stage III metastatic breast cancer, and the student is asked to help recommend one of two drug treatments as proposed by two different doctors. The student receives documents from mostly real sources, including information from the National Cancer Institute, the Mayo Clinic, pharmaceutical companies, newspaper articles, and adapted data from research studies, among others. The biology content is woven into the performance task: the friend in the performance task asks the student why the doctors were not concerned by a benign breast tumor, and asks the student to explain the difference between a benign and malignant tumor. This question requires the students to cell division in a situated manner -- which in another situation might have been assessed through a decontextualized, multiple-choice test.

In another performance task, students would assume a role on the local city council, which is charged with voting on and approving the upcoming lease agreement for the municipal vehicle fleet. Given increases in fuel costs, the city council wants to adopt a more economical and

[39] Again, these performance tasks use the same approaches and elements of those used in the CLA, and utlize the principles of authentic assessment and performance tasks.

environmental option over the traditional gasoline vehicles that make up the current fleet. The top two supported options are currently either hybrid or biodiesel based vehicles. Such a performance task could be used in courses in environmental studies, biology, political science or economics.

Other topic areas have included performance tasks crated by faculty on air pollution, art exhibition catalogues, debates between intelligent design and evolution, drunk driving reduction, employment discrimination, university policy, autism, farm subsidies -- clearly a tremendous range.

Student Diagnostic Report and Institutional Diagnostic Report

In order to facilitate further the improvement of student learning, CLA Education is launching in the 2009-10 academic year the Student Diagnostic Report (SDR) and the Institutional Diagnostic Report (IDR). These programs allow an individual faculty member to assess the students in her or his own class; using a CLA Performance Task no longer in rotation as part of CLA Assessment Services. Student responses are scored by a trained reader, but rather than controlling for initial ability and providing analyses of overall student growth (as is part of CLA Assessment Services), responses are scored relative the basic demonstration of the key higher order skills in an absolute sense. Using a scoring rubric that is being explored to in the future anchor diagnostic scoring[40] of the CLA, such that faculty would receive sub-scores of student performance on (a) analytic reasoning and evaluation, (b) problem solving, (c) persuasiveness, and (d) mechanics. The SDR reports back individual student results (with careful caveats to note that the results should not be used for any form of high-stakes decision making), but instead (and perhaps more importantly given the focus on teaching and learning) as one piece of evidence that could be used as part of a diagnostic conversation with students about their skills; this would be part of a larger, ongoing process focusing on student development. The IDR aggregates these results the whole institution.

The program permits faculty to have on-demand assessment (that can be conducted off-cycle from the formal CLA), since it need not be scored with and reported back relative to the rest of the institutions participating in the CLA .

[40] The closest parallel would be the more commonly understood trait scores, although the CLA is taking a slightly different tact here.

Final field testing of the SDR and IDR is being completed in the spring of 2009.

SUMMARY AND CONCLUSION

As Benjamin, Chun and Jackson (chapter 2) note, the CLA is strategically placed in the context of assessment and accountability, and it is multifaceted approach to engage in assessment, education, research and policy ensures that these components will inform one another. Benjamin, Chun and Shavelson (chapter 3) outlined the role of diagnostic logic employed by the CLA, which provides a justification for the work of CLA Education, and Benjamin (chapter 4) focuses on faculty and how the CLA is a test worth teaching too (with measures outlined by Hardison, Hong, Chun, Kugelmass and Nemeth (chapter 5).

Despite consensus about the need for students to develop higher order skills like critical thinking, research has shown that students are not developing them. Although we can argue that faculty are the key player in processes of teaching and learning, the research has provided limited lessons learned about the role of faculty. We have hints that faculty can teach these skills, but little evidence given the divergence of ways in which "teaching critical thinking" has historically been operationalized.

CLA Education takes a straightforward approach: that students will develop these skills if they have opportunities to practice them. Drawing on the literature on authentic assessment and performance tasks, CLA has found a key link between assessment and classroom practice. When substantive classroom work focuses on development of these skills, but assessment tools are used whereby multiple choice tests attempt to capture proxies or self-reports of learning, it is not surprising to find a disconnect between the two (and for faculty to be frustrated by the lack of validity in the measures used to assess their work). CLA connects the two by using performance tasks that serve both purposes. This coherent approach has served to galvanize faculty, because they see the connection between assessment and pedagogy. A dean of Academic Affairs noted:

> One of our faculty members was skeptical of the utility of the CLA given her own professional needs and concerns. However, after she participated in the workshop on designing a unique, course-based performance task to assess student learning, she was a convert. Immediately, she saw how a performance task—especially one that she could tailor to reflect the concerns of her science course—could energize student learning and serve as a platform on which students could synthesize and integrate knowledge from various different disciplines and courses.

The continuous system of improvement described in this monograph makes possible the direct connection between assessment and classroom teaching and learning. Having a clear mandate to develop critical thinking skills sets the stage for this work to occur, and recognizing the role of faculty is crucial. But after reconfiguring the landscape of assessment and accountability, and by recognizing that authentic assessment and performance tasks (vs. multiple choice tests, self reports and other inadequate proxies) may finally provide the tools not only to directly connect assessment and classroom practice, but more importantly to do it in a way that faculty endorse, it is only now that the real work can begin. It will not be easy work, knowing that our current system has graduated generations of students unable to extrapolate from a complex book or calculate totals costs of orders. And there may be new challenges teaching faculty how to do this, given that our research to date has not revealed a clear theory of practice that informs how to develop critical thinking skills. But now that the stage is set, we can abandon all that distracted us and confused the process. We are ready to move forward. We can now return to learning.

REFERENCES

Arum, Richard, Josipa Roksa, with M. Velez, Learning To Reason and Communicate in College: Initial Report of Findings From the CLA Longitudinal Study. New York: Social Science Research Council, November 2008.

Association of American Colleges and Universities (AAC&U) (2002). Greater Expectations: A New Vision For Learning As A Nation Goes To College. Washington D.C.: AAC&U.

Ballantine, Jeanne H. and Joan Z. Spade (Eds.) (2007). Schools and Society: A Sociological Approach to Education, Third Edition. Wadsworth Publishing.

Benjamin, Roger (2008). "The Case for Comparative Institutional Assessment of Higher-Order Thinking Skills" Change, Vol 40, No. 6, November/December: 51-55.

Benjamin, Roger (2008). The Importance of the Faculty In The Age Of Assessment. New York, NY: CAE.

Benjamin, Roger and Stephen Klein (2007). "Assessment Versus Accountability In Higher Education: Notes On Reconciliation." United Nations Educational, Scientific and Cultural Organization (UNESCO) Commissioned Paper Series: 1-26.

Benjamin, Roger, Marc Chun, and Richard Shavelson (2007). Holistic Tests In A Sub-Score World: The Diagnostic Logic of the Collegiate Learning Assessment, New York, NY: CAE.

Benjamin, Roger. (1980). The Limits of Politics Collective Goods and Political Chance in Postindustrial Societies. Chicago, IL: University of Chicago Press.

Benjamin, Roger. (2003). The Environment of American Higher Education: A Constellation of changes. The Annals of the American Academy of Political and Social Science, vol. 585, January: 8-30.

Benjamin, Roger. (2007) "Recreating the Faculty Role in Governance in Research Universities," in Joseph C. Burke (ed.) Fixing the Fragmented University. Bolton, MA: Anker Publishing, 2007: 70-98.

Bok, Derek (2006). Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More. Princeton University Press.

Bransford, J., A. Brown, and R. Cocking (eds.) (2000). How People Learn. Washington D.C.: The National Academy Press.

Bush, Vanevar (1945). Science: The Endless Frontier. Washington, D.C.: United States Government Printing Office.

Chun, Marc (2002). Looking Where the Light is Better: A Review of the Literature on Assessing Higher Education Quality. Peer Review, 4 (Winter/Spring): 16-25.

Chun, Marc (2006). Faster, Better, Cheaper: The Iron Triangle of Higher Education Assessment. New York, NY: CAE

Council for Aid to Education (2006). Board Statement on Assessment. www.cae.org

Graf, Gary and Carol Birkenstein (2008). A Progressive Case for Educational Standardization: How Not to Respond to the Spellings Report. Academe On line, May/June. http://www.aaup.org/AAUP/pubsres/academe/2008/MJ/Feat/graf.htm

Hage, Jerald. and Charles Powers (1993). Post-Industrial Lives, Roles and Relationships in the 21st Century. Beverly Hills, CA: SAGE Publications.

Halpern, Diane (1993). Assessing the effectiveness of critical thinking instruction. Journal of General Education, 42(4): 238–254.

Hardison, Chaitra and Anna-Marie Vilamovska (2009). The Collegiate Learning Assessment: Setting Standards for Performance at a College or University. Santa Monica, CA: RAND.

Hersh Richard and Roger Benjamin (2002). Assessing Selected Liberal Arts Outcomes: A New Approach. Peer Review, Winter/Spring, Vol. 4, No 2/3: 11-15.

Hersh, Richard and Roger Benjamin (2002). Measuring the Difference College Makes: the RAND/CAE Value Added Assessment Initiative. in Value Added Assessment of Liberal Education. Peer Review, vol. 4, no. 2/3 (winter/spring): 7-10.

Hibbard, Michael (2000). Performance-Based Learning & Assessment in Middle School Science. Eye on Education.

Immerwahr, John. (2000). Great Expectations: How The Public and Parents--Whites, African-Americans, and Hispanics View Higher Education. San Jose: CA: National Center For Higher Education Public Policy, August.

Klein, Stephen (2002). Direct Assessment of Cumulative Student Learning in Peer Review, Vol. 4, No. 2/3: 26-28.

Klein, Stephen, David Freedman, Richard Shavelson, Roger Bolus (2008). Assessing School Effectiveness. Evaluation Review, 32, 6: 510-525.

Klein, Stephen, George Kuh, Marc Chun, Laura. Hamilton, and Richard Shavelson. (2005). An Approach to Measuring Cognitive outcomes Across Higher-Education Institutions. Journal of Research on Higher Education, vol. 46, no. 3: 251-276.

Klein, Stephen, Marc Chun, Laura Hamilton, George Kuh, and Richard Shavelson (2005). An Approach to Measuring Cognitive Outcomes Across Higher Education. Research in Higher Education 46.1: 251-276.

Klein, Stephen, Roger Benjamin, Roger Bolus, and Richard Shavelson (2007. "The Collegiate Learning Assessment: Facts and Fantasies." Evaluation Review 31.5: 415-439.

Koretz, Dan, Brian Stecher, and Stephen Klein, Dan McCaffrey, and Edward Deibert (1993). Can Portfolios Assess Student Performance and influence Instruction? The 1991/92 Vermont Experience. RAND Institute on Education and Training and UCLA: CSE Technical Report 371, National Center for Research on Evaluation, Standards, and Student testing. Los Angeles, CA: CRESST/UCLA.

National Research Council (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington D.C.: The National Academy Press: 44-51. 59-104.

Pascarella, Ernest and PatrickTerenzini (2005). How College Affects Students: A Third Decade of Research. San Francisco, CA: Jossey-Bass.

Pellegrino, James. W., Norm Chudowsky, and Robert Glaser (eds). (2001)Knowing What Students Know: The Science and Design of Educational Assessment. Washington D.C.: The National Academy Press.

Ragin, Charles (1989). The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. Berkeley, CA: University of California Press.

Shavelson, Richard (2007). "Assessing Student Learning Responsibly: From History to an Audacious Proposal." Change 39.1: 26-33.

Shavelson, Richard (2007). A Brief History of Student Learning: How We Got Where We Are and a Proposal for Where to Go Next. Washington, DC: Association of American Colleges and Universities.

Shavelson, Richard (forthcoming). "The Collegiate Learning Assessment," in The Quest to Assess Learning and Hold Higher Education Accountable. Stanford, CA: Stanford University Press.

Shavelson, Richard J., R. W. Roeser, H. Kupermintz, S. Lau, C. Ayala, A. Haydel, S. Schultz, G. Quihuis and L. Gallagher (2002). Richard E. Snow's Remaking of the Concept of Aptitude and Multidimensional Test Validity: Introduction to the Special Issue. Educational Assessment, 8(2): 77-100.

Shavelson, Richard, Amy Kurpuis and Matt Bundick with Richard Hersh, Daniel Silverman, Corey Keyes, and Lynn Swaner (unpublished paper, 2009). On Assessing Learning Broadly and Responsibly.

Simon, Herbert (1996). The Sciences of the Artificial. Boston, MA: M.I.T. Press.

The College Portrait: the VSA Template. www.voluntarysystem.org

U. S. Department of Education (2006). The Secretary of Education's Commission On The Future of Higher Education. Washington D.C.: Government Printing Office.

Wagner, Tony (2008). The Global Achievement Gap. Boston, MA: Basic Books.

Wiggins, Grant (1990). The Case for Authentic Assessment. Practical Assessment, Research & Evaluation, 2(2).