collegiate learning assessment

WHITE PAPER

HOLISTIC TESTS IN A SUB-SCORE WORLD: THE DIAGNOSTIC LOGIC OF THE COLLEGIATE LEARNING ASSESSMENT

ROGER BENJAMIN, MARC CHUN
AND RICHARD SHAVELSON

[collegiate learning assessment]

HOLISTIC TESTS IN A SUB-SCORE WORLD: THE DIAGNOSTIC LOGIC OF THE COLLEGIATE LEARNING ASSESSMENT (CLA)

Roger Benjamin and Marc Chun, CAE
and Richard J. Shavelson, Stanford University

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 to 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 are 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 example of the Collegiate Learning Assessment, 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.

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 Information Age," which makes strengthened higher-order skills essential. In the Information Age 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 Information Age 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, new media such as 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, analytical 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, analytical 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.[1] 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

[1] Namely: The institution--not the student--is the primary unit of analysis (although additional sampling permits analysis of subgroups). Its value-added approach is central. The focus is on higher-order skills rather than content. Students write open-ended essays instead of reply to multiple-choice questions. There are no "right" answers; rather the student is evaluated on the quality and the extent to which they make a reasoned, reflective argument.

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 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 Collegiate Learning Assessment (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.[2] While some of these abilities that are used in addressing CLA tasks can be named, for example "quantitative reasoning" (at least hypothetically) the CLA takes the view that the whole is greater than the sum of the parts and does not attempt to pull the abilities apart in an artificial manner.

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 score 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

[2] See (a) our responses to the questions from the NASULGC Learning Outcomes Technical Work Group; which will be posted on the AASCU website, as well as (b) a recent paper "The Collegiate Learning Assessment: Facts and Fantasies," (Klein et al.) forthcoming in *Evaluation Review*; which may be accessed from www.cae.org/cla.

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 (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 student 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.

To this end, the CLA is committed to placing the capacity of building CLA-type tasks (and scoring rubrics) in the hands of colleges and their faculty; this will be accomplished through a new program called CLA in the Classroom. For this program, we are working on a guide for building tasks and scoring performance. We also plan to create a repository of such tasks, contributed by interested colleges, on our website so as to make as many tasks as possible widely available.

As a first step, the CLA will release the first "disclosed" performance task as a way of providing a concrete example of what we are trying to test and how we go about it. Indeed, some colleges might want to widely administer the task and score students' performance to get first hand understanding of how the CLA works and how faculty might expand the number of tasks for teaching and learning purposes.

With CLA-type tasks in hand, programs and individual faculty, perhaps in collaboration with a campus teaching-learning center, would be in a position to collect systematically their own diagnostic

information about the strengths and weaknesses of students and programs. Faculty members would be able to engage with students in classroom discussions or one-on-one (in some cases asking a student to "think aloud") during their performance. They would also be in a position to debrief students about the task afterwards, asking about their performance. And faculty could, jointly, examine and analyze students' written work, developing hunches about how to improve that performance. In these ways, students' specific strengths and weakness would become evident in the course of teaching.

Such information about student thinking and performance combined with other student work samples (e.g. classroom assignments, papers, problem sets, oral presentations) would enable programs and teams of faculty to situate students' analytic reasoning, critical thinking, problem solving and communicating given what else they know about their students. For example, knowing that students did well on math problem sets might eliminate the concern about students' skills in reading a table of numbers. Rather, it might suggest that the students were not well prepared to write cogently about quantitative evidence. This might lead to a conjecture and subsequent data collection that by focusing on more essays for the class, their performance might be improved.

Our goal, then, is not for the faculty to re-create the CLA at their campus. Rather, we believe that using CLA-type tasks would provide a means for campus programs, departments and faculty to have a better conversation with their students about where their performance could be improved relative to higher-order skills. Sub-scores of tasks are not required for campuses to have the means to understand what explains the scores and what might be done to improve the performance of their students. It involves programs and faculty directly in the assessment work while honoring the knowledge faculty have about their students.

In addition to releasing a disclosed CLA performance task, we intend to develop a set of materials that will support faculty in using the materials in their classrooms, field test the materials in a range of types of institutions, offer workshops for faculty in the CLA consortia of institutions to develop their own versions of the CLA tasks, and evaluate whether or the extent to which increased use of the CLA-type assessments, teaching techniques such as presenting many more problems and cases for students to analyze and write about results in creased CLA scores and thus, we assert, increased higher order skills.

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.

References

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.

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

Benjamin, R. (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, pp. 8-30.

Benjamin, R. and S. Klein (2007). Assessment Versus Accountability in Higher Education: Notes for Reconciliation. Forum Commission Paper. Paris: UNESCO, pp. 27.

Benjamin, R. and R. Hersh (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), pp. 7-10.

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

Chun, M. Looking Where the Light is Better: A Review of the Literature on Assessing Higher Education Quality. Peer Review 4, no. 2-3 (2002): 16-25.

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

Klein, S., G. Kuh, M. Chun, L. Hamilton, and R. Shavelson. (2005). An Approach to Measuring Cognitive outcomes Across Higher-Education Institutions. Journal of Research on Higher Education, vol. 46, no. 3, pp. 251-276.

Immerwahr, J. (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.

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

Shavelson, R.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, R. (2007). Assessing Student Learning Responsibly: From History to An Audacious Proposal. Change January/February, pp.26-33.

Simon, H. (1996) The Sciences of the Artificial. Boston, MA: MIT Press.