

Response to the Phase Two Report of the  
Working Group on Student Learning Outcomes and Assessment

Joint Task Force on Accountability  
University of Massachusetts Amherst  
October 14, 2011

The Phase Two report from the Working Group on Student Learning Outcomes and Assessment (WGSLOA) represents careful consideration of a broad range of issues stemming from the successive iterations of the “Visions” project and its strong focus on the evaluation of the Commonwealth’s public institutions of higher education. It follows logically from the Phase One report, and acknowledges many of the questions and issues raised following publication of that report. The Joint Task Force on Accountability at UMass Amherst (UMAJTFA) is grateful for the opportunity to comment on the Phase One report and for the effort exercised by WGSLOA in attempting to respond. As a result of this effort we believe the key issues are clearly defined.

In our opinion, the Phase Two report is especially effective in defining and articulating the implications of the “tension” to which it frequently refers: that between assessment for the purposes of program improvement and evaluation for the purposes of accountability. The stark juxtaposition of these two perspectives in the Phase Two report reveals the central dilemma facing the BHE in its quest: can a valid evaluation system be constructed that is faithful to the demands of both perspectives?

This question has national relevance, but for this purpose we want to answer it in the context of the Vision project and the Phase Two report. As we understand it, the goal of the Vision project is to develop a faculty-driven, segmentally focused, state-level, comparative evaluation system capable of satisfying the accountability desires of many constituencies, including prospective students and their families and a wide range of local, state and national policy-makers. The Working Group seeks to accomplish this by “piggy-backing” on student outcomes assessment conducted for the purposes of institutional improvement.

It is important to recognize at the outset that this is something that has so far never been accomplished. Despite decades of effort on the part of institutions, governing bodies, accrediting groups, and higher education organizations, a system of comprehensive, comparative evaluation, rooted in student outcomes assessment, has remained out of reach. It is also important to note that higher education is replete with rankings, ratings, and other evaluative efforts that, while persistent, do not accomplish what the Vision project proposes. In thinking about the current effort in Massachusetts it is useful to consider why other efforts have proved unsuccessful, and we will return to that larger context from time to time in the comments that follow.

For the purposes of clarity, we will address each of the dimensions outlined above, and then talk a little about interdependencies among them. Taking the topics in reverse order:

- I. Satisfying accountability desires means, in short, providing reliable answers to the questions that the target audience is asking. This would seem to be the threshold consideration. A system that fails to answer the right questions, or that provides unreliable answers, fails to achieve its basic purpose. Even worse would be a system that misleads by purporting to measure more than it really does. This is especially important given the “high stakes” orientation of the Vision project and its stated desire to guide decision-makers.

The Phase Two report lays out a conceptual sketch of how this might be approached: guided by the AAC&U’s LEAP goals, various kinds of indirect, direct, and “embedded” assessments would be combined into a “composite” score, which would then be used for the purposes of comparison among state partners.

The fundamental question is whether such an approach would be likely to produce reliable answers to the “right” questions.

1. With respect to the organizing goals, we believe the answer is a partial “yes.” The LEAP framework has been developed with strong collaboration across many constituencies, and its goals represent a very useful structure within which to talk about learning objectives and outcomes for a subset of undergraduate education related to general education and a “liberal arts” perspective. This emphasis is not surprising, given that LEAP was developed by an association of predominantly liberal arts institutions. UMass Amherst shares an interest in general education and the liberal arts, and we in fact have adopted the LEAP framework in our ongoing efforts to assess our general education program.

The partial “no” represents the flip side of LEAP’s focus. While we share a liberal arts mission with most other institutions, we also have learning objectives related to undergraduate research, professional education, and other dimensions of what we do that — while relevant to LEAP — we might approach in a somewhat different way were the universe of comparison just public research universities. The practical implication of this is that a LEAP-oriented system would tend to present an incomplete picture of UMass Amherst. We believe there will be a strong tendency for the Vision system — whatever it is — to become “the” way in which Massachusetts public institutions are evaluated, so the prospect of being viewed through a restricted lens is troubling. As the Working Group noted in quoting Alexander Astin: “assessment efforts should not be concerned about valuing what can be measured but, instead, about measuring that which is valued.” Otherwise we are in the position of the drunk searching for lost keys under a streetlamp because “the light is so much better there.”

2. With respect to the proposed reliance on indirect, direct, and embedded assessments, if we again ask whether such an approach would be likely to produce reliable answers to the “right” questions, at this time we believe the answer is “no:”
  - a. While indirect assessments like NSSE are very valuable (and used by us for many purposes), in the view of most observers they do not provide the kind of insights that can support summative assessment, especially when aggregated and compared far from the program level. In fact, we understand that NSSE itself is currently expressing the need for caution when using these data for such purposes. By definition, indirect assessments are not reflective of actual student work, which we believe must serve as the basis for any meaningful evaluation of institutional performance. Indirect assessment can be a vital tool for formative evaluation, in part because it can serve as a guidepost to issues that deserve closer scrutiny. But indirect assessments can rarely support the weight of summative judgments with respect to student performance.
  - b. Direct assessments avoid this fundamental weakness, but being “direct” is not enough: the assessment must also be robust enough to carry the weight of summative judgment. The Phase Two report, like the Phase One report, rejects the notion of direct assessment in the form of a single, universal test for a range of reasons eloquently detailed in the reports. We fully agree, and we are very grateful that our concerns about testing are reflected in the conclusions of the Working Group. But the use of licensure exams, admissions exams, and the like, as proposed in the Phase Two report, suffers from many of the flaws of testing in general, with the additional weakness associated with the lack of universality. At UMass Amherst, a tiny fraction of undergraduates participate in licensure exams, and those are in highly specialized areas like teaching and nursing. It simply does not seem reasonable to believe that overall institutional performance can reliably be abstracted from results for a few test-takers in narrow professional fields. The GRE and similar exams, while incorporating a broader set of learning objectives, are still very limited in terms of student participation. At UMass Amherst only about 15% of graduating seniors take these exams, and they are obviously a self-selected population that cannot be said to fairly represent the institution as a whole. Moreover, the test-taking patterns at different institutions undoubtedly vary widely. For example, we imagine that UMass Amherst has a higher proportion of GRE-takers than UMass Boston; UMass Lowell has a higher proportion of nursing students than UMass Amherst, and hence the NCLEX exam would have greater weight in any “composite” measure utilizing direct assessments; and so on. It is difficult to

see how this approach would produce the kind of data necessary to support summative judgments.

- c. Embedded assessment is obviously the gold standard for both formative and summative assessments. The direct observation of student work in the context of clearly articulated learning outcomes has become accepted as the best — and at this point, the only — approach capable of providing deep and reliable insights into student performance. We therefore agree with the Working Group to the extent that it focuses on embedded assessment. The Working Group does not, however, address how embedded assessments at the program level could be translated, normalized, and aggregated in ways that would produce a reliable institutional assessment capable of supporting summative judgments. We are deeply interested in expanding the scope and effectiveness of embedded assessment on the campus, but we believe that the validity of using such assessments as the basis for institutional evaluation in a summative context has not been demonstrated nor even deeply explored. Adopting a statewide evaluation system with this approach at its core would therefore seem premature at best.

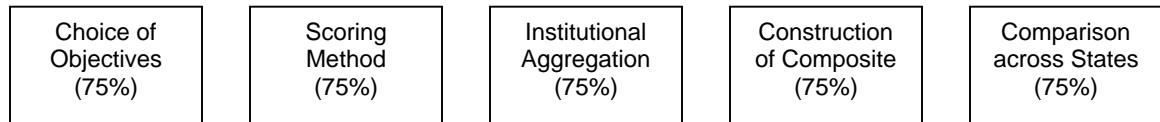
Finally, we want to comment on the Working Group's call to use indirect, direct, and embedded assessment in combination. Here the distinction between formative and summative evaluation is perhaps most obvious and most important. For the purposes of formative evaluation, where the objective is to glean insights from many sources, identify questions deserving of further study, and try a variety of approaches to promote program improvement, employing multiple perspectives makes complete sense. What may not be visible through one lens may come into focus through another. But for summative evaluation, where the objective is to form conclusions and drive decision-making, the use of multiple perspectives may or may not add power, especially if the tools are weak individually. To use a prosaic example, if one goes to the doctor for an annual physical and the blood pressure gauge is slightly off, the scale is a few pounds light, and the thermometer isn't working that day, the physician's overall assessment of health — and the plan for follow-up diagnosis and care — may not be severely compromised. The totality of evidence and observation by an experienced practitioner may support a sound assessment. But if one is visiting the surgeon as a follow-up to a liver transplant, then decisions regarding as to the success of the procedure and follow-up treatment demand precise measurement of temperature, blood chemistry, and so on. The summative judgment — "is this liver working right?" — cannot reliably be made on the basis of incomplete or ambiguous data. Moreover, multiple perspectives are only of value if they add useful information. Supplementing a faulty

observation with a poorly calibrated instrument does not produce a better diagnosis. And while we understand that the Phase Two report presents only a conceptual scheme for an evaluation system, our concern is that the concept itself is flawed for the reasons cited above.

3. With respect to the proposed conversion of various forms of assessment into a “composite student learning indicator,” we must ask again whether such an approach would be likely to produce reliable answers to the “right” questions. In this case we believe the answer is clearly “no.” As we understand it, the idea is to create a weighted metric composed of varying proportions of direct, indirect, and embedded assessment results, with this “composite” serving as the basis for judgment and comparison of the institution. This seems unworkable for several reasons:
  - a. As noted above, the composite’s components do not seem individually up to the task, so combining them in any proportion would not seem to produce the needed outcome.
  - b. We are not aware of a proven method for creating such a composite. On what basis would the proportions be determined? What would a composite that reflected such diverse perspectives actually convey? What phenomena would one actually be observing via the composite?
  - c. In general, we do not believe the case has been made that overall institutional performance can be expressed in a single metric.
  - d. Even if it were somehow possible to craft a valid single measure, we do not see how that would respond to the needs of different audiences. “Performance” always means performance with respect to something. In higher education there are many such “somethings”: efficiency, affordability, love of learning, opportunity, career success, personal attention, and so on. These values are often in competition. What single metric could capture the trade-offs and interdependencies among such a broad set of interests?
  
4. With respect to the use of the proposed composite indicator for comparative purposes with other state partners, we also believe the answer is “no.” First, each state’s composite would suffer from the weaknesses described above. Second, there would undoubtedly be some variation from state to state in how assessments are designed and implemented, program mix, demographics, and so on. The composite is by definition an attempt to normalize the assessment data, but in so doing it would inevitably obscure institutional difference that might be important to understand.

Ultimately, the conceptual model proposed in the Phase Two report suffers from a familiar evaluation dilemma: the further one gets from the actual phenomenon

being observed, the weaker the reliability of the observation. In all multi-step evaluations one must attend carefully to the error introduced at each step, because the probabilities of error are multiplicative. This is easy to see in the illustration below:



So in this example, there is error possible in the choice of objectives to assess (e.g., they are incomplete or poorly framed); there is additional error in the actual scoring of student work (e.g., inter-rater variability); additional error in the process of turning 70 program scores into one institutional score (e.g., sampling, weighting); additional error in constructing a single composite score (e.g., component error, weighting); and additional error at the point of comparison (e.g., different upstream decisions at all steps). If each step in the process is 75% “accurate” (i.e. introduces 25% error), then by the time the final summative judgment is made (e.g., “Massachusetts is performing better than Missouri”), the probability that the apparent answer truly reflects performance is roughly 24%. Even if accuracy at each step could be brought up to 90%, at the end of the chain the probability of a true answer would be scarcely better (59%) than flipping a coin. Even if error could be reduced to 5% at each step, our confidence in the final judgment would only be about 77%.

We must therefore ask what an acceptable level of confidence would be. To us, this returns us to the purpose to which the evaluation is put. For example, if the question being asked is, “which institutions ought to be asked to take a closer look at their assessment practices,” then a fair amount of error may be acceptable. But if the question being asked is, “what state has shown the greatest increase in student performance over the past five years,” then a very modest amount of error would make it impossible to give a confident response. [We note that this is the same challenge confronted by the APLU in developing the Voluntary System of Accountability (VSA). The VSA methodology introduced many opportunities for error, and the resulting summative judgment is limited to “performs above expectation,” “performs at expectation,” and “performs below expectation” — and there is considerable doubt as to whether even those rough judgments will prove defensible.]

We believe this is important because it is our understanding that the Vision project seeks to make specific and important judgments about higher education in the state. Students, parents, and policy makers want to make good decisions, and for that they need reliably true information. If we implement a system in which underperforming institutions are declared to be doing good work, and vice versa, then we would seem to be defeating the whole purpose of the effort.

- II. Comparative means that assessment results will be used to compare states or segments in terms of measured performance, presumably with an eye toward identifying those with greater or lesser performance. The Vision documents make frequent reference to comparison and concepts such as “a leading state.”

As noted above, one obstacle to valid comparison is ensuring that each participating state conducts its evaluation according to the same definitions and methods. The Working Group proposes forming a consortium of states (through the LEAP framework), which could help promote consistency in approach and hence reduce some forms of error.

Even if it were possible to align assessment practices across states, some important differences would remain. For example, pre-K through 12 educational systems vary significantly from state to state. A student emerging from a highly effective high school, for example, might be expected to be better equipped for college and hence more likely to show performance gains when compared with a student from a challenged high school. State-to-state differences in the distribution of more-effective and less-effective high schools would therefore be likely to leach into higher education performance assessments.

Similarly, states vary significantly in terms of which students attend public institutions. For example, in California by policy nearly all in-state UC students graduated in the top 5% of their high school classes. In other states a broad distribution of students attend public institutions.

Making meaningful comparisons, then, would require identifying and controlling for important variables, which would be a daunting challenge. Without doing so, however, one would be unable to know whether a measured performance difference reflected the effectiveness of the higher educational institution, some background variable, or a combination of both. Without such knowledge the judgments that could be made about the higher educational institution would be quite limited.

- III. A state-level system is one in which a primary purpose of assessment is to characterize performance at the level of an entire, multi-institution and multi-segment state. This represents one of the most unusual features of the Vision project. We are unaware of successful efforts elsewhere to establish a single metric capable of reliably measuring performance across all institutions in a state. In this case, it is proposed that state-level assessment also be robust enough to support state-to-state comparison. To be successful, we believe several challenges will have to be met:

1. As noted earlier, state higher education systems reflect a wide diversity of institutional types and missions, ranging from two-year technical and community colleges to graduate research universities. A first challenge is therefore to find ways of measuring performance that capture that diversity. So far, the Working Group has focused on a subset of

institutional activity related to general education-related goals. While these are important, they do not provide an adequate basis for defining the total higher educational performance of a state.

2. While most states use a traditional “three-tier” approach in organizing higher education, the nature, scale, and distribution of institutions vary widely. A single “state” measure would therefore have to account for these differences in some way. While weighting or other kinds of adjustments are possible, each introduces error. It has not yet been demonstrated that a single “state” measure can reliably represent actual student performance across different institutions, so we believe “proof of concept” in this regard should be an early priority of the Working Group.

The Working Group proposes that state partnerships in the context of the LEAP framework could be a useful first step, and we agree. It will be important to reach consensus among state partners as to the goals and uses of the assessments, so that a common understanding can guide the research design. It does not seem possible to develop a system in which state-to-state comparison is possible without such an understanding.

- IV. A segmentally focused system is one in which the object of analysis is not individual institutions but “segments,” organizational constructs intended to group institutions of like mission. Success in this sense would require two things: 1) demonstrating that findings within sectors reasonably approximate findings for institutions within sectors (i.e. sector results do not misrepresent institutional results); and 2) sector results mean the same thing when used in a comparative context from state to state.

Again, as with state-level analysis, there are many challenges to creating a meaningful “segment” metric. For example, some states, like California, have relatively homogeneous university, state college, and community college sectors. Many other states, including Massachusetts, have considerable variation within sectors. We are not aware of current approaches that satisfactorily account for the wide variation in segments across the country.

- V. A faculty-driven assessment effort is commonly understood to have two meanings: 1) faculty are responsible for determining how the attainment of educational goals is measured; and 2) the faculty are integrally engaged in the assessment activity. The first is important because effective assessment cannot be divorced from the articulation of learning objectives and curricular design, which are the province of the faculty. The second is important because the faculty must have confidence in and actively embrace assessment tools in order ensure their credibility and durability.

The Working Group has stressed the importance of faculty leadership in the assessment effort from the beginning, and UMAJTFA applauds this insistence on faculty ownership. As noted above, without such ownership the chances of building a reliable and durable assessment system seem remote.

To date, however, in both internal campus discussions and in conversations with faculty at other institutions, we have not yet observed the level of support necessary for the success of the approach outlined by the Working Group. We believe this stems from several causes:

1. First, the approach itself presents many unanswered questions. The Working Group asserts that program-level assessment findings can be normalized and summarized at the segment and state levels, combined in some way with indirect measures (such as NSSE) and GRE and other test and licensure results, and then converted through some form of weighting into a score that would permit valid comparisons across states. But as noted throughout this document, serious, unresolved obstacles are evident at each level and across the model. The Working Group does not provide a conceptual or evidentiary basis that would explain how the model it proposes could produce reliable and accurate results. Until this gap is filled, it is difficult to see how the model could attract meaningful faculty support.
2. Second, also as noted earlier, the Working Group acknowledges but does not resolve fundamental differences between formative and summative assessment. In particular, there seems to be a failure to appreciate the difference between the kinds of evidence that can inform formative judgments and the kinds that are capable of carrying the much heavier weight of summative judgments. Most faculty are trained researchers, very much attuned to the standard of proof to which their own findings and arguments are subjected every day. If their professional judgment and experience suggest to them that an analytic approach is flawed, they are not likely to choose to be associated with it. By proposing that important judgments about the state's higher education enterprise should be based on evidence not appropriate to the task, the Working Group is undermining the potential for faculty participation.
3. Finally, we sense a troubling ambiguity with respect to the purpose of this assessment. On the one hand, the focus seems to be on student performance. But on the other, the objective seems to be positioning the state. Both may be valid purposes, but faculty are acutely aware that the national trend toward competing claims in the form of rankings, ratings, and other comparisons has resulted in a great deal of bad analysis. In this context, concern over methodological weaknesses is heightened. For faculty to embrace an evaluation plan, we believe it will be essential to demonstrate its rigor and validity, and to allay fears that it will become the servant to one agenda or another. In our view, that process begins with an acknowledgment of

the challenges described throughout this document, and with a commitment to make judgments only in proportion to the strength of the evidence.