Classroom Assessment and University Accountability

MICHAEL S. MILLER
DePaul University
Chicago, Illinois

ABSTRACT. Classroom assessment measures the knowledge that universities have *added* to a student's stock of knowledge. Yet the community wants an individual who has attained or surpassed a particular *level* of knowledge. The central argument presented in this article is that the goal of accountability requires that the primary concern of faculty, and of assessment, be the measurement and certification of the level of knowledge achieved, rather than value added. The additional attention being paid to value added may actually reduce the ultimate level of learning.

Faculty, as participants in classroom learning, are sensitive to three important signals: those sent by students about learning; those sent by colleagues about curriculum and standards; and those sent back and forth between the university and the community about our graduates. The first of these signals, known now by the rubric "assessment of classroom learning," has come center stage at many universities. The effort to devise new methods to assess learning is driven in part by the belief that constructive responses obtained from these techniques will allow a university to evaluate more fully what has been accomplished in its teaching and curriculum. Once this is done, the university may better understand the content of the signal it sends to the market; it will, in that sense, become accountable to itself, to its students, and to the community.

Though there are myriad directions to take in the assessment of learning, universities must direct their limited resources to the assessment issues that will have the greatest marginal effect on the quality of the university's product. Attention currently is being directed at the first signal. Faculty are being directed to gather feedback frequently from students on how and how much material is being learned, so they may adjust their teaching during the semester to make student learning more effective.[1]

In this article, we suggest that universities focus instead on accountability to the community; that is, on the signals it sends to the market. This requires that we specify exactly what we, and eventually the community, think the signal contains. Grades, the most common market signals, are of utmost concern to employers, admissions officers of graduate programs, and students, but they are of minor importance in the current thrust of classroom assessment. This has resulted in a mismatch between the goals and practices of classroom assessment and the desires of the market. Assessment today is grounded in a desire to evaluate how well students are learning, to consider the knowledge we have added to an individual. Yet the market wants a signal of whether or not an individual has attained or surpassed a particular *level* of knowledge.

The central argument presented here is that the goal of accountability requires that the primary concern of faculty, and of assessment, should be the measurement and certification of the level of knowledge that students have achieved. Second, the additional attention that is being paid to value added may actually reduce the ultimate level of learning. As a result, the focus of assessment for the time being should be on the professor, academic standards, and grades and their meaning vis-à-vis the student's level of knowledge. Attention to value added can be valuable only after any uncertainty surrounding the level of knowledge has been settled. This conclusion has major implications for assessment research and application.

A Model of Classroom Production

The factors that could be considered in the construction of a model of classroom learning are substantial in number and complexity. Though such complexity could make modeling intractable, economists make such problems manageable through the methodology of positivism. The tools and language of the economic theory of production provide us with a structure we can apply to the important facets of learning in a university setting, especially because as-

sessment is concerned with how learning is produced in the classroom. This methodology forces us to render the process of classroom learning to its essential components, find the causal relationships between these components, and then use this model to glean useful inferences and develop effective policies. Let us begin with a specification of the output and inputs within a conventional university setting.

The Output

The output produced by education of a student S is knowledge attained, or KN_s. KN_s can be ascribed to a single course (in the short run), a group of courses, or a degree. To simplify the issue for the moment, our focus will be on a single course, and we shall assume that the student's intention is to take the university's certification of the course and performance to the market. The market in this case is comprised of employers, graduate programs, and other professors for whom this course serves as a curricular prerequisite.

The Inputs

The major determinants of KN_{sc} are four: a student S_i (for $i = 1,2,....,j$), who brings to the mix a certain level of natural ability, an endowment of knowledge retained from earlier education and experiences, and an expenditure of time over the duration of the course; the professor P_i (for $i = 1,2,...., k$), who chooses a text, reading list, assignments, course handouts, a philosophy of education, and graded materials; the curriculum C and standards applied to that curriculum; and other inputs R, which most notably include the level of resources and capital (e.g., library facilities, computer labs, classroom, teaching assistants and tutors).[2] The resultant short-run production function, which is course and student specific, is:

$$KN_{sc} = f(S_i, P_i, C, R).$$

The effect of an increase in each input on KN_{sc} is positive, but with diminishing marginal returns. To simplify the analysis, we may assume reasonably that for any given course the levels of R and C are constant, so any and all changes in KN_{sc} during the quarter or semester will come from the actions and interventions of the student and/or the professor only.

The Measurement of Output

A primary duty of professors is to find a valid and reliable way of conveying to the student and others the level of KN_s achieved. Unfortunately, there is no objective measure of knowledge. As a result, we must find a proxy for knowledge that recognizes the immaterial nature of knowledge, yet is meaningful to the person receiving the information. By convention we have chosen grades as that proxy.[3] The various grades reflect some predetermined level of knowledge as stated in a university *Bulletin*. For example, an A is often described as superior performance, a C as an average performance, an F as a failure, and so on. The professor first establishes that a student has achieved a level of knowledge, then translates this level into a grade consistent with the department's (or university's) standards. The grade G will be measured numerically in grade points, subject to $0 \leq G \leq 4$, which assumes an A = 4, B = 3, and so on. So, the output for a given course ($c = 1,2,...g$), for a given student for whom the grade awarded was G is G_{sc}. The proxy for the cumulative attainment of knowledge over courses 1 through g would be contained in the grade point average.[4]

If universities are to be accountable, they must be precise in specifying what KN_{sc} and G measure. There are two measurement approaches to consider. The first is the philosophy behind current classroom assessment practices. Angelo and Cross (1993, p. 5) argued that "[c]lassroom assessment is a formative rather than a summative approach to assessment. Its purpose is to improve the quality of student learning, not to provide evidence for evaluating and grading students; consequently, many of the concerns that constrain testing do not apply." Thus, assessment is applied to the process of learning and precedes the awarding of a grade, and its intent is to maximize the difference between what was known coming into the class and what is known at the end of the class (i.e., value added). To do so we must assess learning on a consistent basis throughout the semester. The sooner we can make adjustments in the actions of the student and professor, the more likely we are to create a larger KN_{sc}.

This approach, however, creates a paradox for accountability: It results in a measure of output that is meaningful to an individual student, but of little meaning or relevance to the market. Maximizing the size of value added, a laudable goal, does not necessarily imply that a given student has achieved even an average mastery of the material. A high value for KN_{sc}, if it represents value added, tells us only that a student learned a lot between the first day and last day of class; it cannot be relied upon to tell us in any absolute sense how much is known. Such a concept of output is of limited use if the ultimate goal is market signal and accountability to the community.

The grade signal requested by the market requires that we focus our attention on the other approach to measurement, one that states KN_{sc} and G in more absolute terms. The market wants to know if a student has achieved a predetermined level of knowledge consistent with the various grade designations, not the highest value added given the quality of the inputs. If we assure that G measures a level of knowledge, and is awarded based on some underlying standard in the choice and application of that grade, then the meaning is clear to all—to the student, the professor, and the market. An important inference from this is that assessment of learning should begin by specifying, or clarifying, this underlying standard. Only then can we have confidence that the signals we send to the market have meaning; only then are we accountable.

These two approaches do not, on the surface, appear to be contradictory. If assessment is formative, then adding some assessment to a regimen that focuses on the level of knowledge might be an aid in maximizing the size of KN_{sc}. Rather than being complementary, however, the two approaches present us with a second paradox: Attempts to expand the formative aspect of learning may reduce rather than increase the ab-

solute level of knowledge obtained by the end of the course. This paradox arises because of several of the characteristics of classroom production. First, there is a time constraint. The number of hours available to the professor to convey a body of information is finite. We are thus faced with a cost/benefit choice for every classroom activity. The more class time is spent assessing or re-teaching the points that remain muddy in the students' minds, the less time we have to present new material. Of course, all of us who teach spend time each class asking questions, and asking for questions, to check for understanding. But assessment techniques require that we use additional methods, and additional class time, to confirm how much students are learning as the semester progresses. Assessment per se is valuable; more assessment, however, may be no more valuable than less assessment. We may conclude that additional assessment activities will maximize the output of knowledge if, and only if, the marginal benefit of the assessment exceeds the marginal cost (the loss of class time for new material).

The second characteristic of classroom production that creates this paradox is the variation in the abilities and preparation of the students in the class. A class of fully prepared students with exceptional abilities and motivation would require little to no assessment. One comprised of ill-prepared students of questionable ability with little motivation would require a profound amount of assessment just to figure out how to proceed. In the former case, assessment would reduce output; in the latter, it would increase output. Because most classes are comprised of students of varying backgrounds, we may be led to conclude that assessment as described by Angelo and Cross (1993), done in the classroom, taking class time, would be indispensable. Though plausible, there is another, more obvious, method that could be used to maximize the output of knowledge of the class: assessment of abilities and preparation *before* allowing students to enter the classroom. Such assessment, and standards, would minimize the variability of student ability in the class.

Variation in ability is unavoidable, but what is avoidable is the placing of students in a class for which they do not have the requisite tools for success. It would be more efficient for the university to make assessments of preparation before allowing students to enter a schedule of classes rather than having four or five professors in a full-time student's course load make four or five separate assessments. Reducing the variability of student preparation can be done most efficiently in two ways: through a selection process at the point of admission (or once admitted, through remedial education); and through a grading system that assures that the standards are maintained and that unprepared students are not passed on into subsequent courses. Should we not be able to assume in a course in microeconomics that a student who earned a passing grade in differential calculus can differentiate simple expressions? Yes, if grades measure a level of knowledge, not necessarily value added.

Allowing substantial variability in student preparation to continue, where the pace of the course will be set by the least prepared students to the detriment of the remainder of the students, assures a reduction in this output. Assessment of university admissions, course prerequisites, and grading standards demands our attention more than assessment of learning in the classroom. Once those matters are put to rest, then classroom assessment merits attention.

Assessing the Output of the Classroom

Because we cannot observe the amount a student has learned directly, we must devise a method through which students reveal to the professor how much has been learned. How do we do this within the time constraint of a semester? Do we ask students to assess through self-reflection the quantity of knowledge produced? No. In spite of our hope that students can tell us how much they have learned, especially when placed within the anonymous, nonthreatening environment of classroom assessment techniques, students do not possess a sufficient depth of knowledge or experience to make such a judgment.[5] To think otherwise would confer upon them a level of wisdom inconsistent with their status as students. Additionally, the market wants a measure not invalidated by the profound conflict of interest a system of self-reporting would create. To depend on an unreliable estimator would not be in the interest of students, the market, or the university. Only performance, as evaluated by someone competent to do so, matters. Obviously, the professor is the only one in a position to specify KN_{sc}.

The nature of the university-student relationship precludes any focus on the student's opinion or situation in the estimation of output. Additionally, because grades are to be a measure of the level of knowledge upon completion of the course, professors must ignore extraneous issues, such as the possibility that one student may be able to master the material with little expenditure of time while another barely passes in spite of trying hard. The amount of knowledge a particular student possesses on entering the course is irrelevant as well. Otherwise, we would need a separate grading scale, and signal, for each individual student. Grades must be narrowly defined, or they become meaningless. How hard a student works, or how great are the hurdles he or she conquered to get even an average grade, may reflect well on character, but muddy the signal sent by grades if included in their computation. Character, hard work, and the like can be transmitted to the market via other means.

If the professor is to be the arbiter of the size of KN_{sc}, we must have a sound understanding of the basis on which the KN_{sc} is revealed. Contrary to the assertion that graded assignments are not assessment, the production model leads us to infer the opposite. Graded materials such as examinations, presentations, labs, and reports are the *only* basis through which students can reveal the level of KN_{sc}. especially in a system that requires we certify KN_{sc} only days after the course has ended. And graded materials are clearly assessment. The standard dictionary definitions of the verb "to assess" are "to set or determine the amount," "to evaluate or appraise." Graded materials do exactly that in such a way as to require the student to reveal this level of knowledge, which is then

deemed superior, average, or unsatisfactory in the granting of a grade. Ideally, we would follow a student throughout his or her education and career, and decide according to observed behavior the amount of knowledge that he or she gleaned from coursework. But the real world requires we do it now.

If graded assignments are the tools that an individual professor uses to assess the level of knowledge, then accountability can come in the form of peer evaluations of the quality of those assignments. Such an evaluation would fulfill two roles: We would observe the basis on which the grade was granted, and we could determine whether the designation of that grade is consistent with the university's standards. Thus, we are assessing the level of KN_{sc} by evaluating the quality of the inputs that were part of the production process.

The proposal, then, is to focus on the right-hand side of the production function to determine valuable information about the left-hand side. The initial reaction could be that this does not further accountability. Though we may be assessing, we are assessing inputs, not output. This argument, however, is specious. An implication of the production-function approach is that we can infer the size of the output (KN_{sc}) from information about the quantity and quality of the inputs.[6] Consider the following generalized production function, and the causal, predictable relationships we use to draw conclusions about the size of output.

Assume we have any representative student, professor, and course. Thus, $KN = f(S, P, C, R)$. In the short run, one or more of the inputs must be assumed constant, so let us assume that R, C, and S are constant. As for P, assume that the only aspect we vary is the quantity and quality of the graded and nongraded materials brought to the class, denoted p_m. We may reasonably posit that an increase in the quantity/quality of p_m will result in a higher level of KN. Take, for example, examinations. We may assume that a poorly written examination with unclear questions covering material unrelated to the course contains less information about the knowledge the student has learned than one comprised of fair, well written questions related to the course topic. This is true regardless of how the student scores in either examination. This quality-output relationship is true of all the inputs, but p_m has a particular relevance to accountability because of its role in determining the size of KN. As a result, we may conclude that an unbiased evaluation of the quality of p_m is an evaluation of the reliability and validity of the grades granted. Unwillingness to make this inference calls the entire evaluation process in education into question. We do, and must, make subjective evaluations about the quality of materials, work, and performance.

Long-Run Considerations

Clarity regarding grades and standards as our first assignment in accountability is important in the long run as well as the short run because of the incentives that underlie the actions of the various participants in the production of KN. There are three agents within the production function, with three different objective functions, which can have profound effects on standards: (a) students, whose short-run objective is to maximize the current value of their certification; (b) administrators, whose objective over the short and long run is to retain standards within constraints of budgets, enrollments, and giving; and (c), faculty, who face the dual objectives of retention of standards by their grading and curricular decisions, while pursuing personal job satisfaction and security. The potential exists for all three agents to undermine standards in pursuit of short-run objectives. That said, the responsibility for, and power over, standards should fall primarily on the shoulders of the faculty. Let us consider why students may not be the best source of information on learning.

The participants with the shortest time commitment to the university are students. Except for that small number of individuals who attend a given university through several undergraduate and graduate degrees, students usually come and go in 5 years or less. Once admitted to a university of any particular reputation, students have the most to gain and least to lose from a relaxation of standards. The objective of a majority of students is to earn certification of an education that they can sell to the highest bidder. What they want is that first job, or admission to a particular graduate school. The market will, other things equal, offer more to a graduate with a higher GPA. So, the maximization of the GPA is a reasonable objective. An individual's performance on the job or in the graduate program will be the dominant gauge of the person after the market has purchased the degree. Thus, the initial market value of high grades and a degree are immense, but that value falls substantially in relation to other market signals as time passes.

With these incentives in place, students would be rational to strive for the highest GPA even if the route is grade inflation (i.e., a reduction in standards).[7] A drop in standards via grade inflation can only improve a student's chances of employment or academic admission. The long-run effects of such inflation—the reduction of the reputation of the university—would not affect them. They will already have cashed in on the value of the degree and grades immediately after graduation. By the time the market detects that the grades earned do not mean what was thought, the original individual will be far enough removed as to suffer no consequences. We may thus conclude that student input is not reliable in the determination and retention of standards and accountability. Students are not competent to provide useful information for the assessment of KN, and their incentives would lead them to act in such a way as to have the university lower rather than maintain standards.

The incentives for administrators are more varied, but pose no less of a problem. Within the production function, the administration has influence over KN via the purchase and allocation of resources and capital (R), often the quality and quantity of the students admitted (especially in private institutions), and the compensation of the faculty. The time horizon of the administration's stake in the university is certainly longer than that of students, but success will likely be evaluated in relation to short-run accomplishments in enrollment management, budgetary control, and university donations. Evaluating admin-

istrators on such criteria is reasonable. A problem arises, however, if the administration's decisions adversely affect the faculty's application of grading and academic standards.

The incentive structure is organized in such a way as to have administrators perceive students as customers rather than students. The only way for budgets to balance is to generate sufficient credit hours. The two ways to generate credit hours—adding new students every year, and keeping the students already enrolled—potentially conflict with the enforcement of standards. A faculty that insists on high standards will be working at cross purposes with the administration. Would a student want to attend a university where C's dominate, when another university is available that has the same reputation (currently) and B's dominate? Of course not, unless the external market is keen enough to realize that a C student from a university with standards can possess more knowledge than a B student from a university without standards. In general, a move from a professor-student relationship to a company-customer relationship will lower standards.

Faculty are not without their conflicting incentives as well, which is all the more important considering that the individual faculty member assigns the final grade. Given the expanded role played by student evaluations of teaching in the promotion-tenure process, the chances are greater that grade inflation and a reduction in standards will occur. There is sufficient evidence in some disciplines that professors can buy higher evaluations with higher grades (consider, for example, Becker, p. 1,369). If a goal of the professor is tenure, then the incentive exists to please the students in their quest for high grades. The resulting improvement in the professor's evaluations (assumed by most to be valid and reliable indicators of teaching effectiveness) will further the professor's case for tenure. Yet, if standards and accountability are our goals, then considerable attention must be paid to the materials mentioned above (p_m) rather than to evaluations. At minimum, the grade distribution of the class evaluated should be considered along with the student comments. If the faculty members as a whole focus on the importance of standards in grades, materials, and curriculum in their deliberations on tenure, promotion, and curriculum committee work, then, and only then, will standards be maintained or reestablished.

Implications for Assessment Within the University

These observations lead us to conclude that the agenda for assessment should focus initially on the specificity of the meaning of a grade, and the quality of the materials the professor uses to evaluate a student's level of knowledge. These two tasks must be considered simultaneously; we must assess the quality/quantity by juxtaposing the performance of the student and the quality of the professor's inputs. In short, assessment must begin with evaluation of the university's faculty, curriculum, and standards, not with in-class evaluation of or by the student. The classroom assessment movement has expanded because we have failed; we have failed in our duty as educators to maintain our standards and enforce those standards in the admission of students, and in their progression from course to course.

With this stated, how do we improve our accountability to the student and the market? Three levels of assessment will take us a long way toward accountability:

1. Establish a set of standards on how much a student should know by the time he or she graduates from the university.
2. Evaluate the professor and his or her materials on their relevance to meeting this standard.
3. Evaluate how well the professor has applied these standards and materials to the measurement of *KN* and its translation into a grade.

NOTES

1. Among the driving forces behind accountability is the AACSB/The International Association for Management Education. Its standards for accreditation require members to evaluate instructional effectiveness on a continuing basis, leading many schools to create standing committees of faculty on classroom and program assessment. See AACSB standards C.2.2 and IN.2 (1993).

2. Normally, resources and capital are separate inputs. Though all can recognize that changes in the quantity and quality of resourcesand/or capital will affect the level of output, the distinction between the two inputs is not important in the context of this article.

3. This designation of grades as a measure of output is a reflection of what actually occurs, rather than what many educators would like to see. Despite all their drawbacks, a vast majority of programs use grades, and they are depended upon by employers and graduate schools. For example, several years ago when DePaul University was debating adding pluses and minuses to its letter-grading system, the initial reaction of the student government was immediate and negative. Then a representative of the students contacted several employers and was told emphatically that employers would much rather look at transcripts with plus/minus grades: to them, such signals contain information they can use in assessing candidates. The students then threw their full support behind the proposal. As an alternative, programs based on experiential learning often focus on attainment of competencies, rather than grades. The discussion of whether that is a superior measure of output is beyond the scope of this article.

4. Using the symbols of this article, the GPA is computed as:

$$\text{GPA} = \frac{\sum_{i=1}^{n} G_{Si}}{n}$$

5. This is not to say the same is true in the long run. After a sufficiently large number of experiences and/or the attainment of more in-depth knowledge, students can judge after the fact. But of what use would that be in the short run? Very little. It can be of immense usefulness to the university in the long run, however, as a means of validating or invalidating the university's initial stance on the quantity of knowledge.

6. Drawing such inferences on the quality of the output from the quality of the inputs is common, and reasonable. If you have two identical metal structures, but one is made of aluminum and the other of steel, we may infer with confidence that the latter is stronger.

7. This incentive is more pronounced if the student's tuition is being paid as a fringe benefit by an employer. Under many corporate fringe benefit plans, the amount of the tuition that is reimbursed is proportional to the grade received. For example, an A merits 100% reimbursement, a B 75%, and so on. Often a C is not reimbursed at all, and the financial incentive driving the student is to find a university where C's are seldom awarded. A university wanting to attract such a student body is not immune from these financial pressures.

REFERENCES

American Assembly of Collegiate Schools of Business (AACSB). (1993). *Achieving quality and continuous improvement through self-evaluation and peer review. Standards for accreditation in business administration and accounting* (Revised April 20, 1993). St. Louis: Author.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques. A handbook for college teachers* (2nd ed.). San Francisco: Jossey-Bass.

Becker, W. E. (1997). Teaching economics to undergraduates. *Journal of Economic Literature, 35*(September), 1347–1373.