EVALUATING THE QUALITY AND QUANTITY OF GRADUATE STUDENT RESEARCH

Suzanne Ortega
Dean of the Graduate School and Vice Provost for Advanced Studies
University of Missouri – Columbia

Two fundamental assumptions from sociology shape my thinking about the evaluation of graduate student research productivity. First, the meaning of any social process, including research and evaluation, is shaped by the context within which it occurs. Second, the whole is always larger (or at least different) than the sum of its parts.

The most important analytic elements of a social context include: the intended uses of the process outcome and the various audiences to receive the "message" about the outcome. As a sociologist, I see two forces involved in shaping the context for graduate student evaluation. Accreditation requires that institutions be able to use objective indicators of student learning outcomes to map their planning efforts. Also, senior administrators and university communities want to utilize resources to improve the quality of graduate education (a proxy for which is often an increase in the number of graduate programs ranked in the top 10 or in the top quartile of the next NRC study of doctoral programs). In terms of audience, there are two that overlap: academic and non-academic, the latter including industry and governmental leaders, and the public at large. The academic audience is internal to an institution and is composed primarily of university administrators, faculty, and current graduate students. Each of these audiences will have a slightly different use for evaluation and assessment data on graduate research productivity. As a result, each will have a differential stake in the efficiency and/or comprehensiveness of the assessment process and each will be interested in a somewhat different set of outcome measures.

Let's look at how our thinking about assessment correlates with the premise that the whole is always something larger (or at least different) than the sum of its parts. Universities are more than the sum of the departments that comprise them. Graduate programs are more than a simple sum of individual student learning outcomes. This also leads us to believe that the quality of graduate research is more than a simple aggregation of the number of graduate student papers presented or published.

Joan Lorden's model for measuring research productivity is an excellent framework for the remainder of my remarks. I will focus on: goals, audiences, values, and practicality.

Aside from satisfying the mandates of accrediting agencies such as the NCA, we might first ask what our **goal** is in graduate program assessment (and by implication graduate research). I would submit that there are two major goals. The first goal is to provide the information necessary to create self-directed plans for improvement in graduate student learning outcomes and in the overall quality of graduate programs. The second major goal is to provide the data necessary to guide resource allocation decisions. Resources can, of course, range from the very tangible to the much less tangible. This can include hiring new graduate faculty, providing enhanced graduate stipends and benefits, and assisting with professional travel and development opportunities, as well as the more intangible aspects of acquiring prestige as a result of achieving some highly desired and difficult to attain outcome.

There are two **audiences** for assessment data and each is differentially interested in one of the two goals above. Internal audiences include all stakeholders within the university. Certainly students and staff are impacted by the perceived prestige of an institutions' graduate research profile and by the relative proportion of resources that flow to it. However, I would like to focus on the two internal audiences that seem to have the most impact on the way research is assessed and used. Administrators—Chancellors, Provosts, Research Administrators, and Graduate Deans, for example—primarily look at research assessment as a tool for making strategic decisions about the use of existing resources and as a platform from which to argue for more external resources, whether that be prestige or funding. Of course, administrators are also interested in improved educational outcomes, but on a day-to-day basis, I believe most are willing to trust another internal audience—the faculty—to make sure that improvements in graduate research training are taking place and that those improvements are reflected in the assessment data they produce.

Faculty, of course, have a major stake in evaluating the quality and the quantity of graduate research. Yet, I would have to say that I have used up more of my reserve of good-will capital with faculty on the assessment issue than on anything else. Even though I keep telling faculty that they should be interested in assessment as a strategy for improving the things they care about, i.e., the preparation of the next generation of scholars and researchers, they believe I actually want a quick and efficient way of allocating—or more frightening still, reallocating—resources. In fact, I suspect most faculty end up going along with our standard graduate research assessment procedures only because they are worried that if they don't comply, they might lose funds. They are skeptical at best that any new resources or opportunities will be forthcoming as a result of an honest evaluation of either graduate research quantity or quality. The ambivalence of faculty may be attributed to the competing and in some ways contradictory use of data to: (1) make resource allocation decisions (an approach that many faculty fear and resent) and (2) make informed, self-directed decisions about program improvement strategies.

Externally, we can divide audiences into two subtypes, other academics and non-academics. When we speak in the language of graduate program rankings and prestige, I would submit that our primary, but not our exclusive, audience is composed of other members of the academic community. This is particularly true with reference to the National Research Council (NRC), where the primary indicators of graduate faculty quality—number of publications in refereed journals, proportion of faculty supported on extramural funds, or even number of degrees conferred—reflect the standard academic **values** of peer review and publication as the appropriate measure of research productivity. To the extent that some portion of our external non-academic audience is composed of aspirants to the academic roles, values, and community (i.e., prospective graduate students), the language of rankings and prestige will be compelling and influential for them, as well.

Although our non-academic external audiences probably share in the same general goals for assessment, i.e. resource allocation and program improvement, it is quite likely that business, governmental, and non-profit leaders will have different performance standards. We are in a situation where appropriate indicators of productivity and quality are still contested within the academy, and we have yet to consider how we might develop productivity and quality measures that address the core values of industry or the public at large. If we consider job placement a measure of student learning, how do we apply this to non-academic placements? By the size of the firm? By the firm's profitability measures? By dollars spent by the firm on R&D? Would placement in a federal agency be ranked higher in quality than placement within a county or municipal social service agency? If the productivity of our graduates is an indicator of graduate program and individual graduate student research quality, what kinds of non-academic research productivity measures speak to the core values of our non-academic audiences? Are patents valid indicators of research productivity? Is quality then measured by patents that lead to the development of start-up companies and by the profits they derive? Is there a metric by which we can gauge the impact and quality of scholarship that leads to new public policy or law? As universities move in the direction of increased collaboration with industry, with increased public accountability, and respect for the wide range of career opportunities for our doctoral degree recipients, it will become more important to develop assessment and evaluation strategies that align with the values and goals of our non-academic audiences. To date I have heard no considered and sustained discussion of the measures we should use.

More than anything else, I believe, the **practicality** factor has led most institutional research offices, accrediting agencies, and organizations involved in educational ranking to use productivity indicators as their best, and often only, measures of research quality. Whether measured in absolute numbers or per capita, indices based on the number of refereed publications, the number of awards, and the amount of extramural research funding have the real advantages of being routinely collected as part of other faculty and student evaluation processes. Because they are numeric, they have the added advantage of being standardized and easily summarized. In its last iteration, for example, the NRC basically relied on the faculty productivity measures identified above to measure graduate faculty quality. Faculty quality, in turn, was used as the indicator of graduate program quality. Although the research protocol is not yet set, a shift toward inclusion of more student outcomes in the next NRC study will likely parallel indicators of research productivity for the faculty.

Clearly, research productivity bears an important relationship to research quality. At the individual level, however, productivity is a necessary but not a sufficient condition for research and graduate program quality. Here, I would simply reiterate that we must begin turning our attention toward the development of easily collected quality measures, appropriate at both the individual and the program/institutional levels, and pertinent to both academia and the broader community.

The practicality of an assessment and evaluation strategy depends as much on the process we use to collect data as it does upon the simplicity, reliability, and validity of the indicators we choose to collect. We will be well served, then, if we can embed the assessment of research quality into a common data collection process that has the capacity to address a variety of institutional needs. This process should recognize the differing values and priorities of our various audiences. At the University of Missouri - Columbia, for example, we are trying to create an integrated assessment process that inputs data from the annual reviews of individual graduate students and merges it with routinely collected institutional measures. Institutional measures typically include: proportion of students supported on assistantships or fellowships; part-time/full-time enrollments; number of degree recipients, and average time to degree. Where there are sufficient numbers of graduate degree recipients to do so, we also utilize summary reports from the NSF Survey of Earned Doctorates. This database, in turn, will provide much of the information about graduate education necessary for state- and institutionally-mandated five-year academic program reviews. By reducing the number of unique reports that departmental chairs and directors of graduate studies must provide, we are optimistic that one of the big stumbling blocks to meaningful assessment will be removed. I would caution, however, that efforts to use student learning outcomes and graduate program quality for thoughtful self-improvement often run at cross purposes to the academic review process, which is fundamentally about resource allocation. We will have to continue to monitor whether the savings in faculty time and the possibilities of creating a truly useful body of information for program development can offset this potential “danger.”

In the end, our core **values** should guide assessment, and not simply issues related to expediency or audience. It seems to me that one of the core values we need to resolve is the question of measuring quality or productivity, per se. If it is quality we want to measure, we must determine how to differentiate it from productivity. In general, I suspect, we are talking about the impact of research when we assess quality. How, then, do we measure impact? Once again, we may find that different audiences will be convinced by different measures. It is also important to keep in mind the distinction between the impact of individuals and the impact of programs.

Within the academy there may already be a fair amount of consensus about how to measure the impact/quality of an individual's research and scholarship. One standard indicator is publication in peer-reviewed high visibility journals with high rejection rates. Citations and the amount of extramural support for research are other standard measures. I would note, however, that these typical measures of scholarly impact work much better for the sciences than they do for the arts and the humanities. Earlier today, Dr. Lorden mentioned the adage: "We are what we measure." We must be careful to develop appropriate measures of quality and impact in the arts and humanities or we may erode the position of these disciplines at our institutions, especially when measures of "impact" drive resource allocation models in the future. By intention or happenstance, our support of the arts and humanities will be an important statement about our institutional values.

If ambiguities remain in the assessment of individual research, this is even more true of efforts to assess quality at the program level. The "value-added" dimensions of a high quality graduate research program will likely be its defining characteristics. Although we may not yet have the measures, I suspect that two important value-added indicators will be: the capacity of programs to foster interdisciplinary research skills and agendas, and the capacity to provide professional development opportunities for the next generation of scientists and scholars, namely, teamwork, sensitivity to issues of diversity and internationalization, communication skills, etc.

In summary, I find that we do not have a measure of impact that is relevant to audiences outside of academia. To do this successfully may entail tackling prejudices about applied research. We almost certainly will have to move beyond a hierarchy that gives preference to basic over applied research. We may need a separate metric appropriate to each kind of research. However, practicality will likely force us to compare the two and ask questions such as: how many refereed articles in what tier of journal does it take to equal the impact of one patent or five technical reports?

Whether we tackle these questions effectively, or indeed at all, will reflect on another basic value issue that universities are now facing—the extent to which we choose to be internally or externally focused. In the next several years we will learn something very important if our measures of research quality remain simply new and improved measures of traditional academic productivity rather than evolving to meet the challenges we have before us.