# MEASURING SUCCESS:

# LESSONS FROM THE NRC STUDY OF THE RESEARCH DOCTORATE

Joan F. Lorden
Associate Provost for Research and Dean of the Graduate School
University of Alabama at Birmingham

*Why Measure Performance?*

For a complex organization like a university, it is a formidable challenge to communicate what we do and its value. The task is made more difficult by the wide variety of audiences that must be addressed. The 2000 Merrill conference focused on making university research part of the public agenda. The public, at this conference, was broadly defined to include not just public officials, potential donors, and industry but also those that we might think of as an internal audience: alumni, students and their parents. Whether the audience is internal or external, we need tools for communication.

Universities have two basic ways of expressing what it is we do and why it is important: numbers and stories. By stories, I mean the narrative explanations that we offer about the significance of the work we perform. Our publications highlight student achievements, faculty discoveries, and the services the university provides for the community. These narratives provide the context for the statistical data that we assemble. As instruments of persuasion, the narratives are compelling and effective, but they are unwieldy when our need is to: set benchmarks, recognize excellence, and promote improvement. Quantitative performance measures can be tools for persuasion but ideally, they are also objective indices that we can use to help set and meet our goals.

The kinds of questions we are called upon to answer imply comparison, either with our own past performance or with that of others. How do we compare? Do we meet the standard? Are we getting better? What are the best practices? To be able to answer these questions, we need to measure the right things in the right way. We need to be sensitive to the limitations of what we are measuring. And when we start comparing ourselves to others, we need to recognize the very real possibility that if we tie the measures to rewards, we will change behavior and become what we measure.

I would argue that there are four basic elements to consider when we select measures for evaluating performance: goals, audience, values, and the practicality of the measures themselves. We need to be able to answer the questions:

- What are we trying to achieve?
- Who are we trying to reach? Do the questions we pose address the concerns of the audience?
- What do we consider important? What behaviors do we want to promote?
- What are the practical limits of what we are trying to do? If we say we are measuring quality or effectiveness, do the measures we have at hand really allow us to do it?

For the remainder of my talk, I would like to discuss *Research Doctorate Programs in the US: Continuity and Change*, published in 1995 as a case study. This work is usually referred to as the NRC study, since it was conducted by the National Research Council (NRC). The NRC appointed a Committee for the Study of Research Doctorate Programs that actually undertook the study. This study was the subject of a position paper by the Council on Research Policy and Graduate Education (Lorden & Martin, 1999). Some of my comments will be drawn from that paper. Following my critique, I would like to provide a few examples of the ways we have tried to measure performance at the University of Alabama at Birmingham (UAB) in research and graduate education.

*Why Examine the NRC Study?*

The NRC study is interesting for a number of reasons, not the least of which is that it produced rankings based on program reputation. Let me mention a few arguments for taking time to look carefully at the NRC study:

- Any rankings tend to be quoted and used. (Take a look at your campus website; See also Webster & Massey, 1992.) For these reasons, it is worth knowing where they come from and what they might or might not mean.
- There are many criticisms of the NRC study, but the reputational rankings, unlike other published ranks, were done by people in the field and were presented alongside objective measures of performance. The choice of the measures reveals the values of the academy.
- Intended as a study of graduate education, the NRC study is almost as much a study of research. Graduate education and research are so

closely intertwined that many of the measures selected are those that one might choose if evaluating the research programs of a university.

- Although the study was thoughtfully designed, its measures have limitations. It is one thing to measure quantity; it is another to measure the quality of what is produced. Because the NRC study is the major national study of graduate education and is soon to be repeated, its measures deserve scrutiny.

As background, let me mention a few of the main features of the NRC study. First and foremost, there was a reputational survey from which rankings of programs within disciplines were derived. The reputational survey was supplemented with objective measures. Similarities with earlier studies allowed for longitudinal comparisons within most disciplines. Basic institutional information such as enrollment, library holdings, and level of research funding was collected and reported.

*Evaluating the NRC Study*

How does the NRC study look if we apply the four questions raised above about goals, audience, values, and practicality of measures?

- One of the declared goals of the NRC study was to assess graduate education in the United States at a time when doctoral enrollment had reached an all time high and more institutions were offering the doctoral degree. There was a perceived need for data to guide policy decisions. The study was undertaken to address the quality and quantity of research doctorates. The objectives for the study were to: update data last collected in 1982; collect new information; analyze components of the new database; and make the data available for further analysis. In addition to having people in the field rate the merits of different programs, the NRC collected quantitative measures of faculty productivity covering publications, citations, and funding. These measures were presented so that one can get a sense of their distribution across the faculty in a program. The information reported included: the percent of faculty publishing, the number of publications/faculty members, and the gini coefficient that measured the degree to which publications were concentrated in a small number of program faculty. Whatever the initial intent, once published, the study became a ranking of graduate programs and universities, and there was little discussion of other measures (e.g., Webster & Skinner, 1996).

- The study was aimed at multiple audiences. The committee that guided the study expected that it would be useful to students and their advisors making choices about graduate programs and that it would

inform the judgment of university administrators and other decision makers at the state and national levels, including those in funding agencies. Finally, the data were presumed to be of interest to scholars in graduate education and to those involved in the research enterprise. One can reasonably ask if it is possible to address an audience this broad with a single document when only limited interpretation is provided. Individual institutions have used the rankings as publicity and to argue for investment (Lorden & Martin, 1999; Mervis, 2000), but there has been no systematic "study of the study." We do not know what the overall impact has been in terms of student choice, or in terms of institutional and other investments. Much of the public debate has revolved around what the rankings mean and how much weight to give them.

- The NRC study is an example of self-examination by the academy. The committee that guided the study was drawn from the academy and the product must be assumed to reflect its values. The choice of variables was based on assumptions about the features of doctoral training environments that facilitate the preparation of research scholars and scientists, including: the reputation of the faculty, their publications, and their funding and awards. Measures related to the subjects of graduate education, the students, were limited to: the number enrolled, the number of women, and the number of degrees reported. The information secured on graduates was: sex, minority status, the percent supported as research or teaching assistants, and the time to degree.

- Because the NRC study was national in scope, it included a broad coverage of disciplines and institutions and a uniform method of data collection. The breadth of coverage was one of the features that made the study interesting and useful, but it also meant that some measures were too costly to undertake, particularly those related to students and alumni. As a practical matter, the study was also too big to be done very often. Data points ten years apart will not track the movement of faculty in and out of programs that can result in significant changes in program profiles.

Given the broad goals and wide audience that the NRC study aimed to reach, it is not difficult to enumerate omissions or aspects of the study that could have been done differently. As a study of graduate education, the most obvious omission was a valid measure of program effectiveness. Program effectiveness was presented in a measure called 93E and was obtained through the same survey of reputation that produced 93Q, the measure of program quality used for the rankings. These two measures, 93Q and 93E, were highly correlated, leading to the criticism that the raters had no real knowledge of program

effectiveness and so made the assumption that reputation and effectiveness were the same. The measure was inadequate to the task.

The study also lacked any measure of student outcomes. This is an admittedly difficult area to tackle on a national scale, but we would all agree that it is an important measure of the success of a program. For prospective students and their advisors, it may be **the** measure of interest. The absence of measures for student outcomes does not imply that the study committee disregarded this valuable information, but it does illustrate that we tend to measure what we can and not necessarily what we need or value.

A similar point can be made about other aspects of graduate education that have become increasingly important. Interdisciplinary programs and research have grown, and many emerging areas are inherently interdisciplinary. Others, such as the disciplines encompassed by the biomedical sciences, have evolved into interdisciplinary fields. It is not a simple task to measure how interdisciplinary a program is and whether it produces students with breadth of training. If we value interdisciplinarity and want to promote it, we need to find a way to capture it in studies of graduate education.

Without question the variable that has been the most frequent object of discussion is 93Q. Defined as the mean scholarly quality of program faculty, 93Q was the variable on which programs were ranked and was intended to capture the perceived intellectual resources in a general field. The measure does not, however, tell us about the experience of students. Nor does it capture the quality of faculty performance or mentoring in graduate education. In fact, we know little about what 93Q actually measures. Reputations are slow to change and may be influenced by a variety of factors, including halo effects of the institution or one prominent faculty member. We could ask whether older programs or larger programs with many graduates are more likely to be familiar to the raters. Translated into rankings, 93Q was reported to two decimal places, and many programs differed only in the third decimal place. The confidence intervals presented in the appendices revealed that the quality of programs as established by this measure could be distinguished only in rather broad terms.

*What Did We Learn?*

Given the limitations of the NRC study, we still learned several things. There were interesting findings about change in programs over time. We also gained information about: the impact of program size; faculty involvement in research and its relationship to rankings; and the relationship of program ranking to student variables. For example, when comparisons with the 1982 study were possible, they indicated that there is great stability at the top and the bottom of the rankings. Differences in fields over time could also be discerned. Biology underwent radical change during the 1980's and the taxonomy of programs in the 1995 study bore little resemblance to the earlier study. In the sciences and engineering, the greatest growth in programs occurred at the top of the rankings.

In the arts and humanities, the largest decreases in program size occurred in the top quarter. In general, new programs tended to be ranked low.

Looking at the measures of research and scholarship, several points stand out:

- Federal funding in science and engineering was highest in programs in the top quarter.
- The percentage of faculty who are publishing varied little across the sciences and engineering.
- Per capita publications correlated significantly with ranks.
- Citations/faculty were much higher in the top quarter in the sciences.
- Numbers of awards and honors were much higher in programs ranked in the top quarter in the arts and humanities.
- Citations tend to be more concentrated in a few faculty members in the top ranked programs, rather than being broadly distributed, indicating the presence of a few highly influential individuals.

Examining the relationship of program size to ranking reveals that top ranked programs in all fields tended to have larger numbers of faculty and larger numbers of students than lower ranked programs, although as noted above, the top ranked programs in the humanities tended to decrease in size during the 1980's. The relationship of size to ranking has raised questions about how we evaluate the overall quality of niche programs. Do programs that are more narrowly defined necessarily offer a weaker intellectual experience? The NRC data also allow one to ask how a particular program does, given its size.

The NRC study confirmed what most institutions know—that time to degree increased across all ranks and all disciplines when compared with the 1982 data—but we also learned that the increase was greatest in the lower ranked programs. This may be related to another observation: students from lower ranked programs were more often supported by teaching assistantships. Those in more highly ranked programs were more often supported as research assistants, coincident with the greater availability of research funding in those programs.

Stepping back from the data, one can build an image of what it would take to construct a top ranked program. Some of the elements would certainly be: "star" faculty, a wide range of faculty representing all aspects of the discipline, and resources sufficient to support a large number of students without heavy reliance on teaching assistantships. Having done that, it is worth stopping to reflect on the study by Cerny and Nerad, known as the "Ten Years Later" study,

in which the emphasis was on student outcomes. In a presentation to the National Association of State Universities and Land Colleges (November 1999), Cerny presented data from a national dataset in which graduates who had received their degrees ten years before and who were employed, were asked about their graduate experience (see also Mervis, 2000). The response to questions about whether or not a person would repeat the degree program and, if so, if he or she would do so at the same institution, was remarkable when put in the context of the NRC study. A high rank was clearly no guarantee of a positive experience. The relationship between program rank and the willingness of graduates to repeat the experience at the same institution was weak to nonexistent. This is not to say that faculty reputation and productivity are unimportant in graduate education, but clearly, we need to know more about what it means to the education and experience of the student. If our goal is to sustain and improve graduate education, the question of how to rate a student's experience or how to define the effectiveness of a program deserves an answer. These are topics that we must tackle both at an institutional and a national level.

*Measurement of Performance in an Institutional Context*

Moving from a national study of research and graduate education to assessment at the institutional level allows one to revisit the question of goals, audience, values, and practicality on familiar turf. I would like to give you a few examples of our efforts to assess the performance of programs at UAB. As background to this discussion, let me point out that UAB is a relatively new institution. We have only existed as an independent institution since 1970. The growth of the research enterprise and graduate education during this short history has been substantial. Although it offers a comprehensive education with undergraduate, graduate, and professional degrees in a wide variety of areas, the university has focused its development around its strengths in medicine and health.

The easiest measure to present when discussing research is funding. As we have seen in the NRC study, the ease with which data can be collected and presented is not necessarily an indication of its importance or relevance. Just as the rankings of the NRC study were the easiest piece to communicate, an institution's extramural research funding is the easiest way to express the level of research activity.

In communicating with both internal and external audiences, all institutions produce graphs that show the changes in research dollars over time. One that we use frequently at UAB is a graph that shows extramural grants and contracts awarded over the past decade. This clearly puts UAB in a positive light, since we have grown from about $100 million to just over $300 million in awards. Like other institutions, we compare ourselves to various peer groups, depending on the audience we are trying to reach. A table that I have often shared with trustees and visitors shows how UAB compares in federal research and

development expenditures with other institutions, nationally and in the south. Nationally, UAB ranked 27th in 1999, and among southern institutions, we ranked third, just behind Duke and the University of North Carolina at Chapel Hill. In funding from the National Institutes of Health, our single largest sponsor, UAB ranked 20th nationally in 2000 and fourth in the south, following North Carolina, Duke, and Baylor. This is a good showing for a relatively new institution and the achievement has garnered attention nationally, but it is not the whole story of our research enterprise.

It is clear from examining research rankings based on funding that if you evaluate your institution by funding alone, it is difficult to move up and keep up. In 1997, UAB ranked 30th in federal research and development expenditures with $125 million. To move up to 27th by 1999, we had to increase our federal expenditures by $40 million. Although our numbers were increasing, we were not alone. This level of growth is difficult to sustain. Over the same time period, Duke had an increase of $36 million in funding and moved up from 24th to 23rd. The University of North Carolina at Chapel Hill had a $25 million dollar increase and decreased in the rankings from 20th to 25th. Do these changes in rank translate into significant changes in the quality of the institutions or the worth of the research supported? I would never argue that, but federal research funding is important because it is awarded based on what most investigators would agree is a fairly rigorous peer review process. While the process is not without its faults, the award of millions of dollars of research support annually from a variety of agencies represents a significant consensus on the value of the work produced by the faculty of the institution. Ultimately, the merits will be assessed by the impact of the findings on problems the research addresses as they are published and translated into products.

Another measure that indicates the success of the faculty at UAB is the percent of proposals to federal agencies that are funded. The fact that we exceed the national average is an indication of the quality of the proposals that are subjected to the peer review process. As a public institution, we also consider in various ways the return on the state's investment in the institution. Appropriations to support public universities differ substantially from state to state. Compared with other research-intensive universities, UAB's state appropriation is modest. The current appropriation is about equal to the university's expenditures in federal research and development dollars. It is important to note the impact of the federal dollars in jobs, income to local government, and spending for goods and services in the city and state because this can be compared to the impact of other state investments and industries.

Limited state funding has led UAB to focus on biomedical research with the goal of being preeminent in this area. The rest of the institution does not flounder as a result, because our emphasis is on building depth and making connections in related fields. We can measure the impact of the investment by looking at the number of students in the undergraduate programs doing research

or by the number of participants in the research enterprise from schools outside medicine. Our approach has been to stay focused, set priorities, fund according to opportunities for leveraging university funds, and then monitor success. The questions we ask are: do we have the basic resources to make an impact? Will we be able to muster enough external resources to sustain a program?

As part of a strategic planning exercise that the university undertook in 1994-95, we affirmed the importance of collaborative, interdisciplinary research to the institution. This was an important factor in the success of the university in research, and it was agreed that as the university matured, we needed to have structures in place to sustain and foster a collaborative environment. As part of this effort, we developed a mechanism for funding our major University-wide Interdisciplinary Research Centers (UWIRC) based on evaluations conducted every three years.

The success of a center is judged on the extramural research funding associated with it, but a number of other qualitative and quantitative criteria are also important in determining whether a center will receive funding and how much it will receive. A defining characteristic of a UWIRC is substantive involvement of members from at least three of the university's twelve schools. Some, like the Center for Aging, have participation from all twelve. The other factors on which these centers are judged include: the resources that they provide for the campus; their outreach efforts; their contribution to education and the intellectual environment through courses and seminars; their contribution to translational research; and the extent to which they serve as a resource for the state and region. Evaluating the centers on these criteria has led to changes in the way they operate. Many have started outreach programs that they would not otherwise have had. Others have initiated post baccalaureate certificate programs or specialized courses. Over the six years that the program has been in operation, we have experienced increased participation as the centers develop pilot grant programs and encourage new interdisciplinary activities.

In graduate education, programs are evaluated on the basis of internal training grants for the award of fellowship and assistantship positions. The programs define their mission, describe the curriculum and degree requirements, and then provide data on the following measures, some of which overlap with those in the NRC study:

- Applicant population (e.g., size, number of international and minority applicants);
- Characteristics of students matriculating (e.g., grades, test scores, percentage of minority students, percentage of international students);
- Funding levels and sources for research and student support;
- Training experience of mentors;
- Publications of mentors and students;
- Intellectual environment, facilities;

- Time to degree;
- Attrition;
- Placement of graduates.

As is the case with the UWIRC funding program, we evaluate graduate programs on the basis of defined criteria and tie resources to changes in behavior that are demonstrated by performance. For example, enrollment of African American doctoral students has more than tripled since 1989, increasing from 53 to 178. Programs with high attrition rates have been forced to examine the reasons and take corrective actions. In both cases, putting a spotlight on an issue made a difference in the attention that programs gave it.

*Conclusions*

Historically, universities have been among the most stable institutions in our society and their contributions are numerous. The performance of universities as the purveyors of knowledge and the creators of new knowledge has been a great success story. University research has been the basis for improvement in the lives of all citizens and has been an economic engine for the nation. Universities have served as the entryway for new citizens seeking full participation in the cultural and economic life of the country. Graduate education draws students to the U.S. from around the world. Despite these and other successes, universities, particularly public universities, are increasingly being held accountable by a multiplicity of audiences. In this environment, measurement of performance is inevitable.

As members of the academy, we need to develop ways to capture accurately the performance of our institutions and their programs. In choosing measures for the future, we need to bear in mind our goals. Why are we engaged in a measurement process? Are we asking how to move up in the ranks? Or, do we want to know how we have served the state or advanced our mission? We need to ask whom we are trying to influence: faculty and administration, prospective students, donors, the legislature? These groups have different questions for us, and measures designed to answer those questions will be more effective. Most importantly, we need to acknowledge the power that performance measures have to change the behavior of individuals. We will not always agree on what to measure and not everything that we value will be easily captured in quantitative measurements. But as members of the academy, we are in the best position to develop valid measures that will promote our values and apply them in ways that sustain and enhance our mission.

*References*

Goldberger, M.L., Maher, B.A., and Flattau, P.E. (Eds.) (1995). *Research Doctorate Programs in the United States: Continuity and Change*. Washington D.C.: National Academy Press.

Lorden, J.F., and Martin, L.B. Towards a Better Way to Rate Graduate Programs. (www.nasulgc.org/councils_research.htm)

Mervis, J. (2000). Graduate Educators Struggle to Grade Themselves. *Science, 287*, 268-570.

Webster, D.S., and Massey, S.W. (1992). Exclusive: The Complete Rankings from the *U.S. News & World Report* 1992 Survey of Doctoral Programs in Six Liberal Arts Disciplines. *Change, 24,* 20-45.

Webster, D.S., and Skinner, T. (1996). Rating PhD Programs. What the NRC Report Says…and Doesn't Say. *Change, 28,* 22-44.