

Executive Summary

EXECUTIVE SUMMARY

The Committee to Examine the Methodology to Assess Research-Doctorate Programs was presented with the task of looking at the methodology used in the 1995 National Research Council (NRC) Study, *Research-Doctorate Programs in the United States: Continuity and Change* (referred to hereafter as the “1995 Study”). The Committee was asked to identify and comment on both its strengths and its weaknesses. Where weaknesses were found, it was asked to suggest methods to remedy them.

The strengths of the 1995 Study identified by the Committee were:

- *Wide acceptance.* It was widely accepted, quoted, and utilized as an authoritative source of information on the quality of doctoral programs.
- *Comprehensiveness.* It covered 41 of the largest fields of doctoral study
- *Transparency.* Its methodology was clearly stated.
- *Temporal continuity.* For most programs, it maintained continuity with the NRC study carried out 10 years earlier.

The weaknesses were:

- *Data presentation.* The emphasis on exact numerical rankings encouraged study users to draw a spurious inference of precision.
- *Flawed measurement of educational quality.* The reputational measure of program effectiveness in graduate education, derived from a question asked of faculty raters, confounded research reputation and educational quality.
- *Emphasis on the reputational measure of scholarly quality.* This emphasis gave users the impression that a “soft” criterion, subject to “halo” and “size effects,” was being overemphasized for the assessment of programs.

- *Obsolescence of data.* The period of 10 years between studies was viewed as too long.

- *Poor dissemination of results.* The presentation of the study data was in a form that was difficult for potential students to access and to use. Data were presented but were neither interpreted nor analyzed.

- *Use of an outdated or inappropriate taxonomy of fields.* Particularly for the biological sciences, the taxonomy did not reflect the organization of graduate programs in many institutions.

- *Inadequate validation of data.* Data were not sent back to providers for a check of accuracy.

The Committee recommends that the NRC conduct a new assessment of research-doctorate programs. This study will be conducted by a committee appointed once funding for the new assessment has been assured. The membership for this future committee may well overlap to some degree the membership of the current committee, but that is a matter to be decided by the NRC President. The recommendations that appear below should be carefully considered by that committee along with other viable alternatives before final decisions are made. In particular, in the report that follows, some recommendations are explicitly left to the successor committee. The taxonomy and the list of subfields, as well as details of data presentation, should be carefully reviewed before the full study is undertaken.

The 1995 Study amassed a vast amount of data, both reputational and quantitative, about doctoral programs in the United States. Its data were published as a 700-page book with downloadable Excel table files from the NRC website. Later, in 1997, it became available on CD-ROM. Because the study was underfunded, however, very little analysis of the data could be conducted by the NRC committee. Thus, the current Committee was asked not only to consider the rationale for the study, the kind of data that should be col-

lected, and how the data should be presented but also to recommend what data analyses should be conducted in order to make the report more useful and to consider new, electronic means of report dissemination.

Before the study was begun, the presidents of organizations forming the Conference Board of Associated Research Councils and the presidents of three organizations representing graduate schools and research universities¹ met and discussed whether another assessment of research doctoral programs should be conducted at all. They agreed to the following statement of purpose:

The purpose of an assessment is to provide common data, collected under common definitions, which permit comparisons among doctoral programs. Such comparisons assist funders and university administrators in program evaluation and are useful to students in graduate program selection. They also provide evidence to external constituencies that graduate programs value excellence and assist in efforts to assess it.

In order to fulfill that purpose, the NRC obtained funding and formed a committee,² whose statement of task was as follows:

The methodology used to assess the quality and effectiveness of research doctoral programs will be examined and new approaches and new sources of information identified. The findings from this methodology study will be published in a report, which will include a recommendation concerning whether to conduct such an assessment using a revised methodology.

The Committee conducted the study as a whole, informed through the deliberations of panels in each of four areas:

- Taxonomy and Interdisciplinarity

The task of this panel was to examine the taxonomies used to identify and classify academic programs in past studies, to identify fields that should be incorporated into the next study, and to determine ways to describe programs across the spectrum of academic institutions. It was asked to develop field definitions and procedures to assist institutions in fitting their programs into the taxonomy. In addition, it was to devise approaches intended to characterize interdisciplinary programs.

¹These were: John D'Arms, president, American Council of Learned Societies; Stanley Ikenberry, president, American Council on Education; Craig Calhoun, president, Social Science Research Council; and William Wulf, vice-president, National Research Council. They were joined by: Jules LaPidus, president, Council of Graduate Schools; Nils Hasselmo, president, Association of American Universities; and Peter McGrath, president, National Association of State Universities and Land Grant Colleges.

²The study was funded by the National Institutes of Health, the National Science Foundation, the United States Department of Agriculture, and the Alfred P. Sloan Foundation.

- Quantitative Measures

This panel was charged with the identification of measures of scholarly productivity, educational environment, student and faculty characteristics, and with finding effective methods for collecting data for these measures. In particular, it was asked to identify measures of scholarly productivity, funding, and research infrastructure, which could be field-specific if necessary, as well as demographic information about faculty and students, and characteristics of the educational environment—such as graduate student support, completion rates, time to degree, and attrition. It was asked specifically to examine measures of scholarly productivity in the arts and humanities.

- Student Processes and Outcomes

The panel was asked to investigate possible measures of student outcomes and the environment of graduate education. It was to determine what data could be collected about students and program graduates that would be comparable across programs, at what point or points in their education students should be surveyed, and whether existing surveys could be adapted to the purpose of the study.

- Reputational Assessment and Data Presentation

The task of this panel was to critique the method of measuring reputation used in the 1995 Study, to consider whether reputational measures should be presented at all, and to examine alternative ways of measuring and presenting scholarly reputation. It was to consider the possible incorporation of industrial, governmental, and international respondents into the reputational assessment process. Finally, it was to decide on new methods for presenting reputational survey results so as to indicate appropriately the statistical uncertainty of the ratings.

The panels made recommendations to the full committee, which then accepted or modified them as recommendations for this report.

The Panel on Quantitative Measures and the Panel on Student Processes and Outcomes developed questionnaires for institutions, programs, faculty, and students. Eight diverse institutions volunteered to serve as pilot sites.³ Their graduate deans or provosts, with the help of their faculties, critiqued the questionnaires and, in most cases, assisted the NRC in their administration. Their feedback was important in helping the Committee ascertain the feasibility of its data requests.

³These were: Florida State University, Michigan State University, Rensselaer Polytechnic Institute, University of California-San Francisco, University of Maryland, University of Southern California, University of Wisconsin-Milwaukee, and Yale University. The type of participation varied from institution to institution, from questionnaire review to administration as well as review of questionnaires.

Because of the transparent way in which NRC studies present their data, the extensive coverage of fields other than those of professional schools, their focus on peer ratings, and the relatively high response rates they obtain, the Committee concluded that there is clearly value added in once again undertaking the NRC assessment. The question remains whether reputational ratings do more harm than good to the enterprise that they seek to assess.

Ratings would be harmful if, in giving a seriously or even somewhat distorted view of the graduate enterprise, they were to encourage behavior inimical to improving its quality. The Committee believes that a number of steps recommended in this report will minimize these risks. Presenting ratings as ranges will diminish the focus of some administrators on hiring decisions designed purely to “move up in the rankings.” Ascertaining whether programs track student outcomes will encourage programs to pay more attention to improving those outcomes. Asking students about the education they have received will encourage a greater focus by programs on education in addition to research. Expanding the set of quantitative measures will permit deeper investigations into the components of a program that contribute to a reputation for quality. A careful analysis of the correlates of reputation will improve public understanding of the factors that contribute to a highly regarded graduate program.

Given its investigations, the Committee arrived at the following recommendations:

Recommendation 1: The assessment of both the scholarly quality of doctoral programs and the educational practices of these programs is important to higher education, its funders, its students, and to society. The National Research Council should continue to conduct such assessments on a regular basis.

Recommendation 2: Although scholarly reputation and the composition of program faculty change slowly and can be assessed over a decade, quantitative indicators that are related to quality may change more rapidly and should be updated on a regular and more frequent basis than scholarly reputation. The Committee recommends investigation of the construction of a synthetic measure of reputation for each field, based on statistically derived combinations of quantitative measures. This synthetic measure could be recalculated periodically and, if possible, annually.

Recommendation 3: The presentation of reputational ratings should be modified so as to minimize the drawing of a spurious inference of precision in program ranking.

Recommendation 4: Data for quantitative measures should be collected regularly and made accessible in a Web-readable format. These measures should be reported

whenever significantly updated data are available. (See Recommendation 4.1 for details.)

Recommendation 5: Comparable information on educational processes should be collected directly from advanced-to-candidacy students in selected programs and reported. Whether or not individual programs monitor outcomes for their graduates should be reported.

Recommendation 6: The taxonomy of fields should be changed from that used in the 1995 Study to incorporate additional fields with large Ph.D. production. The agricultural sciences should be added to the taxonomy and efforts should be made to include basic biomedical fields in medical schools. A new category, “emerging fields,” should be included.

Recommendation 7: All data that are collected should be validated by the providers.

Recommendation 8: If the recommendation of the Canadian Research-Doctorate Quality Assessment Study, which is currently underway, is to participate in the proposed NRC study, Canadian doctoral programs should be included in the next NRC assessment.

Recommendation 9: Extensive use of electronic Web-based means of dissemination should be utilized for both the initial report and periodic updates (cf. Recommendations 2 and 4).

DETAILED RECOMMENDATIONS

Taxonomy and Interdisciplinarity

The recommendations concern the issue of which fields and which programs within fields should be included in the study. Generally, the Committee thought that the numeric guidelines used in the 1995 Study were adequate. Although the distribution of Ph.D. degrees across fields has changed somewhat in the past 10 years, total Ph.D. production has remained relatively constant. Thus, it was concluded that there is no argument for changing the numeric guidelines for inclusion unless a field that had been included in past studies has significantly declined in size.

Recommendation 3.1: The quantitative criterion for inclusion of a field used in the preceding study should be, for the most part, retained—i.e., 500 degrees granted in the last 5 years.

Recommendation 3.2: Only those programs that have produced five or more Ph.D.s in the last 5 years should be evaluated.

Recommendation 3.3: Some fields should be included that do not meet the quantitative criteria, if they had been included in earlier studies.

Doctoral programs in agriculture are in many ways similar to programs in the basic biological sciences that have always been included. Recognizing this fact, schools of agriculture convinced the Committee that their research-doctorate programs should be included in the study along with the traditionally covered programs in schools of arts and sciences and schools of engineering. In addition, programs in the basic biomedical sciences may be in either arts and science schools or in medical schools. A special effort should be made to assure that these programs are covered regardless of administrative location.

Recommendation 3.4: The proposed study should add research-doctorate programs in agriculture to the fields in engineering and the arts and sciences that have been assessed in the past. In addition, it should make a special effort to include programs in the basic biomedical sciences that are housed in medical schools.

A list of the fields recommended for inclusion is given in Table ES-1, at the end of the Executive Summary.

Recommendation 3.5: The number of fields should be increased, from 41 to 57.

The Committee considered the naming of broad categories of fields and made recommendations on changes in nomenclature for the next report.

Recommendation 3.6: Fields should be organized into four major groupings rather than the five in the previous NRC study. Mathematics/Physical Sciences are merged into one major group along with Engineering.

Recommendation 3.7: Biological Sciences, one of the four major groupings, should be renamed "Life Sciences."

The actual names of programs vary across universities. The Committee agreed that, especially for diverse fields, the names of subfields should be provided to assist institutions in assigning their diversely named fields to categories in the NRC taxonomy and to aid in an eventual analysis of factors that contribute to reputational ratings.

Recommendation 3.8: Subfields should be listed for many of the fields.

Although there is general agreement that interdisciplinary research is widespread, doctoral programs often retain their traditional names. In addition, interdisciplinary programs will vary from university to university in whether their status

is stand-alone or whether they are a specialization in a broader traditional program. The Committee believes that it would assist potential students in identifying these programs, regardless of location, if it introduced a new category: *emerging field(s)*. The existence of these fields should be noted and, whenever possible, data about them should be collected and reported, but their heterogeneity, relatively brief historical records, and small size would rule out conducting reputational ratings since they are not established programs.

Recommendation 3.9: Emerging fields should be identified, based on their increased scholarly and training activity (e.g., race, ethnicity, and post-Colonial studies; feminist, gender, and sexuality studies; nanoscience; computational biology). The number of programs and degrees, however, is insufficient to warrant full-scale evaluation at this time. Where possible, they should be included as subfields. In other cases, they should be listed separately.

The Committee wished to recognize a particular class of interdisciplinary program, "global area studies." These are programs that study a particular region of the world and include faculty and scholars from a variety of disciplines.

Recommendation 3.10: A new broad field, "Global Area Studies," should be included in the taxonomy and include as subfields: Near Eastern, East Asian, South Asian, Latin American, African, and Slavic Studies.

Quantitative Measures

Data collection technology and information systems have vastly improved since the 1995 Study. Although the Committee wishes to minimize respondent burden, it concluded that collecting additional quantitative measures would assist users in characterizing programs and in understanding the correlates of reputation.

Recommendation 4.1. The Committee recommends that, in addition to data collected for the 1995 Study, new data be collected from institutions, programs, and faculty. These data are listed in Table 4-1 in Chapter 4.

Student Processes and Outcomes

The Committee concluded that all programs should periodically survey their students about their experiences and perceptions of their doctoral programs at different stages during and after completing their doctoral studies, and that programs in different universities should be able to compare the results of such surveys. It also recognized that to conduct these surveys and to achieve response rates that would permit program comparability for 57 fields would be pro-

hibitively expensive. Thus, it recommended that a questionnaire for graduates be designed and made available for program use (Appendix D) but that the proposed NRC study should only administer a questionnaire, targeting students admitted to candidacy in selected fields.

Recommendation 5.1: The proposed NRC study of research-doctorate programs should conduct a survey of enrolled students in selected fields who have advanced to candidacy for the doctoral degree regarding their assessment of their educational experience, their research productivity, program practices, and institutional and program environment.

Although potential doctoral students are intensely interested in the career outcomes of recent graduates of programs that they are considering and although professional schools routinely track and report such outcomes, such reporting is not usual for research-doctorate programs. The Committee concluded that such information, if available, would provide a useful way of distinguishing among programs and be helpful to comparative studies that wish to group programs that prepare students for similar kinds of employment. The Committee also concluded that *whether* a program collects and makes available employment outcomes data useful to potential students would be an indicator of responsible educational practice.

Recommendation 5.2: Universities should track the career outcomes of Ph.D. recipients both directly upon program completion and at least 5-7 years following degree completion in preparation for a future NRC doctoral assessment. A measure of whether a program carries out and publishes outcomes information for the benefit of prospective students and as a means of monitoring program effectiveness should be included in the next NRC assessment of research-doctorate programs.

Reputational Measures and Data Presentation

The part of the NRC assessment of research-doctorate programs that receives a lion's share of attention, both from the general public and within academia, is the presentation of survey results of scholarly quality of programs. Often these results are viewed as simply a "horse race" to determine which programs come in first or are in the "top 10." In truth, many factors contribute to program reputation, and earlier studies have failed to identify what they might be. What the Committee views as the overemphasis on ranking has encouraged the pursuit of strategies that will "raise a program in the rankings" rather than encourage an investigation of the determinants of high-quality scholarship and how that should be preserved or improved. Toward this end, the Committee recommends that the next report emphasize rating rather than ranking and include explicit measurement

of the variability across raters as well as analyses of the factors that contribute to scholarly quality of doctoral programs. Furthermore, in reporting ranking, appropriate attention should be paid to statistical uncertainties. This recommendation, however, rejects the suggestion that reputational ratings should be totally discarded.

Recommendation 6.1: The next NRC survey should include measures of scholarly reputation of programs based on the ratings by peer researchers in relevant fields of study.

The Committee applied and developed two statistical techniques that yield similar results to ascertain the variability in ratings of scholarly quality.

Recommendation 6.2: Resampling methods should be applied to ratings to give ranges of rankings for each program that reflect the variability of ratings by peer raters. The panel investigated two related methods, one based on Bootstrap resampling and another closely related method based on Random Halves, and found that either method would be appropriate.

The Committee concluded that the study could be made more useful to both general users and scholars of higher education if it provided examples of analytical ways in which the study data could be used.

Recommendation 6.3: The next study should have sufficient resources to collect and analyze auxiliary information from peer raters and the programs being rated to give meaning and context to the rating ranges that are obtained for the programs. Obtaining the resources to collect such data and to carry out such analyses should be a high priority.

After examining how closely the measure of effectiveness in doctoral education ("E") correlates with the measure of scholarly quality of program faculty ("Q") in the 1995 Study, the Committee agreed that "E" should be dropped from the next study. Another qualitative measure, the change in program quality in the last 5 years ("C") should be replaced by the change in "Q" between studies for those programs and fields that were included in both studies.

Recommendation 6.4: The proposed survey should not use the two reputational questions on educational effectiveness (E) and change in program quality over the past 5 years (C). Information about changes in program quality can be found from comparisons with the previous survey analyzed in the manner we propose for the next survey.

Although in some fields the traditional role of doctoral programs as trainers of the professoriate continues, in many

other fields a growing proportion of doctorates takes up positions in government, industry and in academic institutions that are not research universities. The Committee was undecided whether and how information from these sectors might be obtained and incorporated into the next study and leaves it as an issue for the successor committee.

Recommendation 6.5: Expanding the pool of peer raters to include scholars and researchers employed outside of research universities should be investigated with the understanding that it may be useful and feasible only for particular fields.

There are very few doctoral programs that will admit that their mission is anything other than to train “world-class scholars.” Yet it is clear that different programs prepare their graduates to teach and conduct research in a variety of settings. Programs know who their peer programs are. Thus, rather than ask programs to declare their mission, the Committee concluded that it would be most useful to provide the programs themselves with the capability to select their own peers and carry out their own comparisons.

Recommendation 6.6: The ratings should not be conditioned on the mission of the programs, but data to conduct such analyses should be made available to those interested in using them.

The Committee wondered whether raters would rate programs differently if they had more information about the program faculty members and their productivity. The Committee recommends an investigation of this question.

Recommendation 6.7: Serious consideration should be given to the cues that are given to peer raters. The possibility of embedding experiments using different sets of cues given to random subsets of peer raters should be seriously considered in order to increase the understanding of the effects of cues.

Different raters have different degrees of information about the programs that they are asked to rate, even if all they are given is a list of faculty names. The Committee would like to see an investigation of the nature and effects of familiarity on reputational ratings.

Recommendation 6.8: Raters should be asked how familiar they are with the programs they rate and this information should be used both to measure the visibility of the programs and, possibly, to weight differentially the ratings of raters who are more familiar with the program.

EXECUTIVE SUMMARY

TABLE ES-1 Recommended Fields for Inclusion

Life Sciences

Biochemistry, Biophysics, and Structural Biology
Molecular Biology
Developmental Biology
Cell Biology
Ecology and Evolutionary Biology
Microbiology
Genetics, Genomics, and Bioinformatics
Immunology and Infectious Disease
Neuroscience and Neurobiology
Pharmacology, Toxicology, and Environmental Health
Physiology
Plant Sciences
Food Science and Food Engineering
Nutrition
Entomology
Animal Sciences

Emerging Fields

Biotechnology
Systems Biology

Physical Sciences, Mathematics, and Engineering

Aerospace Engineering
Biological and Agricultural Engineering
Biomedical Engineering
Chemical Engineering
Civil and Environmental Engineering
Electrical and Computer Engineering
Operations Research, Systems Engineering, and Industrial Engineering
Materials Science and Engineering
Mechanical Engineering
Astrophysics and Astronomy
Chemistry
Computer and Information Science
Earth Sciences
Mathematics
Applied Mathematics
Oceanography, Atmospheric Sciences, and Meteorology
Physics
Statistics and Probability

Emerging Fields

Nanoscience and Nanotechnology
Information Science

Arts and Humanities

American Studies
History of Art, Architecture, and Archaeology
Classics
Comparative Literature
English Language and Literature
French Language and Literature
German Language and Literature
History
(Linguistics moved to Social and Behavioral Sciences)
Music
Philosophy
Religion
Spanish and Portuguese Language and Literature
Theatre and Performance Studies
Global Area Studies

Emerging Fields:

Race, Ethnicity, and Post-Colonial Studies
Feminist, Gender, and Sexuality Studies
Film Studies

Social and Behavioral Sciences

Anthropology
Communication
Economics
Agricultural and Resource Economics
Geography
(History moved to Arts and Humanities)
Linguistics
Political Science
Psychology
Sociology

Emerging Field

Science and Technology Studies
