

How to Interpret NRC 2009 Results

Jaxk Reeves, SCC Director
Jien Chen, SCC Associate Director

Presented at UGA Graduate School Workshop
October 27, 2009

National Research Council Surveys of Research Doctoral Programs

- Surveys conducted about once every 15 years (1983, 1995, 2009)
- One goal is to evaluate (rank) PhD-granting programs in various fields
- Current survey has more quantifiable aspects to counter previous criticisms of being 'popularity contest' that unduly favored established programs

Planning for 2009 NRC Survey

- 1995 – Previous NRC Survey Released
- 1998-2001 - Comments/suggestions on methodology obtained
- 2003 – New Methodology Proposed
- 2006-2007 – Data collected from Institutions, Programs, Faculty, and Other Sources
- July 2009 – NRC Guideline to Methodology Released (202 pages!)
- Nov. 2009 (?) – Final NRC Report to be Issued

Sources for More Information on NRC 2009

- A Guide to the Methodology of the National Research Council's Assessment of Doctorate Programs:
<http://sites.nationalacademies.org/pga/Resdoc/index.htm>
(202 page official document describing methodology)
- University of Texas at Austin Graduate School Presentation at College Dean's Workshop (8/31/2009) :
NRC Assessment of (some) US Research Doctoral Programs
<http://www.utexas.edu/ogs/nrc/assessment083102.pdf>
(56 page 'summary' for UT Deans and Administrators)

NRC 2009 Survey – Who Is Included

- Institutions Participating (N=222, including UGA)
- Broad Fields (Areas) (7 Areas; UGA has programs in 6)
- Fields (76 defined, 61 ranked; UGA has programs in 41)
- Programs (N=4965 nationally; UGA has 50 participating)
- To be eligible, a program must have produced at least 5 PhD's in the 5-year period 2001-2005.
- To be a ranked field, there must be at least 25 programs nationwide participating.

NRC 2009 Survey Who is Included at UGA by College

- Excludes Professional Schools
- Excludes Performing Arts
- Excludes Schools of Education (generally)

INCLUDED	EXCLUDED
Arts and Sciences	Business*
Agriculture & Envir. Science	Education*
Ecology	Environment & Design
Family & Consumer Science	Law
Forestry & Natural Resrc.	Pharmacy
Journalism & Mass Comm.	Social Work
Public & Intl Affairs	Veterinary Medicine

NRC Survey Composition by Broad Field Nationwide and at UGA

AREA	FIELDS	RANKED FIELDS	UGA FIELDS	UGA PROGS	UGA RNK PROGS
Agric Sci.	6	6	6	10	10
Bio&Hlt Sci.	13	13	10	12	12
Phys Sci.	9	9	7	7	7
Engineering	10	10	1	1	0*
Social Sci.	10	10	10	13	13
Humanities	14	13	7	7	6**
Emerging	14	0	0	0	0
Total	76	61	41	50	48

Differences Between NRC 2009 Survey and Previous NRC Surveys

- Previous NRC Assessments were felt to be mostly reputational.
- Previous surveys collected little data from programs, and collected data were used differently by different evaluators.
- NRC 2009 Survey collects quantifiable information on 21 different variables for each program and uses this information in a 'well-defined' way to obtain rankings.

3 Types of Data Collected (P,X,R)

- P (Program Data)

21 Quantifiable Variables for each program, obtained from Institution, Program, Faculty Member, or National Source.

- X (Perceived Importance Data)

Information obtained from each responding faculty member in a field (86% participation nationally, 78% at UGA), with each giving his/her weight for the contribution of each variable to a program's worth.

- R (Perceived Rankings Data)

Information obtained from a randomly selected subset of faculty respondents (about n=180 in large fields) asking each to evaluate 15 randomly selected programs on a 1 (awful) to 6 (great) scale. For large programs, about 45 faculty evaluated each of 50 chosen programs.

P Data – 21 Key Variables

- Many variables collected by Institutions in 2005-06
- Many others provided by Programs in Fall 2006
- Some provided by Faculty during 2006-07 AY
- Data on Publications, Grants, and Awards for 2001-2006 obtained by NRC from national databases, where possible
- 21 variables divided into 3 groups of 7 variables
- 3 groups correspond to Faculty characteristics, Student characteristics, & Program characteristics

P Data – Faculty Characteristic Variables

- Publications per faculty member per year [2000-2006] **R**
- Avg. citations per publication [2000-2006] **R**
- % Faculty with Grant [2005-06 AY] **R**
- % Interdisciplinary Faculty [2005-06 AY]
- % non-Asian Minority Faculty [2005-06 AY] **D**
- % Female Faculty [2005-06 AY] **D**
- Awards/Faculty Member [2001-2006] **R**

P Data – Student Characteristic Variables

- Median GRE Q Score of Entering Students [2004-2006]
- % 1st-yr Receiving Full Support [Fall 2005] **S**
- % 1st-yr With External Support [Fall 2005]
- Average # of student pubs/presentations [not collected]
- % non-Asian Minority Students [Fall 2005] **D**
- % Female Students [Fall 2005] **D**
- % International Students [Fall 2005] **D**

P Data – Program Characteristic Variables

- Average # completing PhD/year [2004-2006] S
- % PhDs graduating within 6 years [F96 to F01 cohorts] S
- % Median Time to PhD Degree [F96 to F01 cohorts] S
- % PhD's with Academic Positions [2001-2005] S
- PhD Student Work Space [-1, +1] [Fall 2005]
- PhD Student Health Insurance [-1, +1] [Fall 2005]
- Average Number of Student Support Mechanisms [F 2005]

Combining P Data – (Z-scores)

- The 20 variables measured for each program are in different units and scales, making it hard to compare them. We desire to combine them into one overall score.
- Standardize using $Z(i,j) = (X(i,j) - \text{Avg}(j)) / \text{SD}(j)$, where
 $X(i,j)$ = Value recorded for Variable j by Program i in Field
 $\text{Avg}(j)$ = Average for Variable j over all Programs in Field
 $\text{SD}(j)$ = Standard Deviation for Variable j over Progs in Fld
- Positive Z-scores are above average, Negative below avg.
- If the overall distribution of scores for a variable (over all Programs in a Field) is roughly mound-shaped, then about 2/3 of Z-scores are in the range $[-1, +1]$, about 95% are within $[-2, +2]$, and values outside of $[-3, +3]$ are very rare.

How to Weight Z-scores

- One could average Z-scores for Program i in Field, yielding: $S(i) = (Z(i,1) + Z(i,2) + \dots + Z(i,20)) / 20$ as the 'Score' for Program i, and then rank programs (higher score is better), but ...
- The above weighting assumes that 'positive Z is good' for all variables, and, more crucially, that all 20 program variables should be counted equally.
- A more general weighted score would be given by :
$$S(i) = w_1 * Z(i,1) + w_2 * Z(i,2) + \dots + w_{20} * Z(i,20),$$
with 'weights' summing to 1. ($w_1 + w_2 + \dots + w_{20} = 1$).
- Key question is who determines these relative weights.

Using X Data to Determine Weights-1

- All faculty in a field (86% participated overall) were asked to give 3 weights summing to 100% for the groups of Program variables (Faculty Characteristics, Student Characteristics, Program Characteristics), representing the respondent's opinion of relative importance of each of these characteristic factors on overall program quality.
- Among the 7 variables within each characteristic factor, responding faculty were asked to name up to 4 variables which were 'important' and up to 2 which were 'very important' with respect to the overall characteristic.
- The above information can be combined to determine an individual faculty member's relative weighting of the 20 Program variables, and averaged over all faculty members within a field to obtain weights w_1, w_2, \dots, w_{20} .

Using X Data to Determine Weights - 2

- Within a factor, 2 'votes' were given to 'very important' variables, 1 point to 'important, but not very important'.
- Weights within a group assigned proportionally to 'votes' assigned, such as $\{2/6, 0/6, 1/6, 2/6, 0/6, 0/6, 1/6\}$ for the 7 variables within a factor.
- Weight for an individual variable is given by product of Factor Weight and Within Factor Weight. For example, if the Factor above had been assigned 30% weight, the 3rd variable in the factor would have $Wt=30\%*(1/6)=5\%=.05$.
- The weights for the 20 variables for each individual are averaged over all responding faculty in that field to obtain the final X-weights $\{ \bar{X}_1, \bar{X}_2, \dots, \bar{X}_{20} \}$ for that field.

Using R Data to Determine Weights - 1

- The X Data represent what faculty say is important. The R Data purport to measure what faculty really value.
- At least 25 and up to 50 randomly chosen programs in a field are each given scores on a 1-6 scale by randomly selected raters. The average of these scores over those who rated that program ($n \approx 40+$) is $R(i)$, the perceived quality of Program i .
- A regression equation of the form:

$$R(i) = m_1*Z(i,1) + m_2*Z(i,2) + \dots + m_{20}*Z(i,20) + \text{error}$$

could be fit to the R data to obtain the best estimates $\{m_1, m_2, \dots, m_{20}\}$, and these could be used as weights.

Using R Data to Determine Weights - 2

- Actual regression fit was made with these modifications:
 - Regression was run on Principal Components of Z-scores, rather than Z-scores themselves.
 - Statistically insignificant Principal Components were eliminated by backwards elimination procedure and back-transformed to give regression weights for 20 variables.
 - Final Weights $\{m_1, m_2, \dots, m_{20}\}$ were normalized so that sum of absolute values is 1. ($|m_1| + |m_2| + \dots + |m_{20}| = 1.00$.

The 20 weights are mostly positive, but could be negative if a high Z-score is actually detrimental (such as for 'Median time to graduate') or if the variable's contribution to program quality is really random noise close to zero.

Final Weight Variable Determination

- We now have 2 sets of possible weights for each Z-score, $\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{20}\}$ and $\{m_1, m_2, \dots, m_{20}\}$.
- The first set of weights directly measures what faculty say is important while the second purports to measure what faculty really value.
- A reasonable compromise is to use the average of the 2: $w_1 = (\bar{x}_1 + m_1)/2$, $w_2 = (\bar{x}_2 + m_2)/2$, ..., and to rank program i within its field on the basis of its score, $S(i)$:

$$S(i) = w_1 * Z(i,1) + w_2 * Z(i,2) + \dots + w_{20} * Z(i,20) .$$

If the NRC had stopped here, most people would generally understand the process, a single rank for each program within a field would be obtained, and we'd be through!

Let Variability in X and R Affect Weights

- Since both the X Data and R Data are subject to sampling variability, it was decided to perform 'half-sampling' re-sampling of both data sets 500 times, calculating 500 new sets of X weights, m weights, and, hence, w weights.
- The interquartile range (middle 50% of responses) for each of the 20 $w(i)$ variables was examined. If all 250 of these weights were positive (or, rarely, all 250 negative), the variable was considered significant and left in the model. If not, the variable was considered insignificant, its weight was set =0, and the remaining $w(i)$'s renormalized.
- This process was done separately for each field, and was continued until a set of 500 possible weight vectors, identical at the 0 points, but varying elsewhere was found.

Accounting for Variability in P Data

- In addition to accounting for variability in X and R data, NRC wished to account for variability in the P data. This was done by generating 500 'perturbed' P data sets, where the 20 variables reported by each program were randomly perturbed around the value reported. This would allow 500 perturbed sets of $Z(i,j)$ scores to be calculated in each field.
- For the 5 program variables which were measured annually by a program, the SD of the perturbation process was set equal to the SE(mean) of the observed values.
- For most of the other 15 variables, the SD of the perturbation process was set to 10% of the observed value. For '% faculty holding grants', 20% was used, while no variation was used for Student Workspace and Health Ins.

Final Ranking Procedure - IQR

- Within each field, each of the 500 sets of $Z(i,j)$ scores obtained from a perturbed data set was randomly matched to one of the 500 randomly generated sets of weight vectors for that field, and a score for each program for that perturbation/weight was calculated as in slide 20. The programs in the field were then ranked from 1st to last on the basis of this score, with high score given rank 1, etc.
- These 500 possible ranks for a program were sorted from least to greatest, and the Interquartile Range (the range containing the middle 250 ranks) will be reported as a program's 'overall ranking'. For example, a program might be told that its overall ranking is 27th to 45th among 83 programs in the field.

Total Procedure (See Figure A-1, Page 3)

- For those who want to see a complete flow diagram of the process used to generate the rankings, please refer to Figure A-1 of the NRC Methodology report, which is reproduced as Page 3 of today's purple handout.
- It took us several days and very careful re-readings of the NRC Methodology report to understand what is displayed in Figure A-1.
- The 'true' process is slightly more complicated than we have summarized here!

Dimensional Rankings (See App. F, Page 4)

- In addition to the overall ranking of a program, the NRC also generated rankings in 3 separate dimensions, called Research Activity, Student Support & Outcomes, and Diversity of the Academic Environment. The components of these measures are the same for all fields, but the relative weights of the components are different by field.
- These weights are obtained from the X data without re-sampling, but the Perturbed P data are used to generate 500 different rankings within a dimension, and IQR is again used to obtain a ranking range for each of the dimensions.
- Average weights for the components used for each of the 3 dimensions over 7 broad fields (Areas) are shown in Tables 3a-3c in Appendix F (Page 4 of purple handout).

Economics Program Ex. (See Table 5.3, Pg. 5)

- There are n=117 Economics Programs rated by NRC 2009.
- Some descriptive statistics, from the P Data, for the 20 Program variables over all 117 programs, are shown in Table 5.3 of the NRC report, reproduced as Page 5 of today's purple handout.
- These descriptive statistics, which will be produced for each field, provide the complete range and IQR for all variables. This information would allow a program to see approximately how its fares on each program variable.
- Why Table 5.3 doesn't list the most common centrality measures (mean and median) is not clear. The mean can be recreated from other information given. The median is usually about halfway between the 1st and 3rd quartiles.

Economics Program Ex. (See Table 5.1, Pg. 6)

- This table is for Program #62 of the Economics Field. The program value listed in column 3 contains the actual non-perturbed program values recorded for Program #62 for each of the 20 variables.
- Column 4 contains the original $Z(i,j)$ -values. For example, $Z(62,3) = (25.50\% - 37.10\%) / 19.9\% = -0.583$, since this Program's % of faculty with grants (25.5%) was slightly more than one-half of an SD below the Econ mean (37.1%).
- Columns 5-6 give Weight +/- 1 SD over the 500 vectors.
- The largest weights are for Cites, Publications, & # of PhDs.
- Note that 8 of the 20 weights are set=0 in all weight vectors, since they were deemed insignificant. The insignificant variables vary by Field.

Economics Program Ex. (See Table 5.2a, Pg. 7)

- Table 5.2a displays the particular Weight*P-perturbation combination (among the 500 run for Program #62 in Econ) which yielded the 125th worst (or 375th best) ranking among the 500 ranks which the program received.
- The weighted score associated with that rank was -0.054, and associated rank was 56th among the 117 programs.
- 'On the average', the perturbed program values were lower than observed for program 62, and the weighting was heavier on 'bad' aspects than 'good' for this run, but exceptions occur.

Economics Program Ex. (See Table 5.2b, Pg. 8)

- Table 5.2b displays the particular Weight*P-perturbation combination (among the 500 run for Program #62 in Econ) which yielded the 375th worst (or 125th best) ranking among the 500 ranks which the program received.
- The weighted score associated with that rank was +0.085, and associated rank was 45th among the 117 programs.
- 'On the average', the perturbed program values were higher than observed for program 62, and the weighting was heavier on 'good' aspects than 'bad' for this run, but exceptions occur.

Economics Program Ex. (See App. G, Pgs 9-12)

- Appendix G of the NRC Report, shown as pages 9-12 of the purple handout, contain the complete rankings for all 117 Economics programs.
- Program #62, highlighted in the previous 3 slides, has its overall rank measure reported as [45 to 56]. Its IQR ranks for Research, Student Support, and Diversity are [21,31], [74,87], and [64,77], respectively.
- Program #93 is best overall ([1,1]), and on Research [1,1], excellent [4,10] on Student Support, and rather dismal [88,98] on Diversity.
- For most fields, the Research dimension correlates highly with overall ranking, Student Support is weakly positively associated, and Diversity is weakly negatively related.