An Alternative Ranking Method in Philosophy

Robert Pasnau, University of Colorado

February 14, 2012

1. Overview

I have long contemplated a new method of ranking philosophy departments. This fall, with a great deal of help from friends and colleagues, I began to put it into execution, ranking only U.S. departments. What follows are my results, admittedly incomplete.

The method is to ask philosophers to rank other philosophers who work within the same area of expertise, and with whose work they are familiar. These results are then aggregated to produce departmental rankings. This seems to me, in principle, a considerable improvement on the method of Brian Leiter's Philosophical Gourmet Report. The PGR asks philosophers to rank whole departments, basing their judgments on a list of all the faculty within a department. The evident weakness of this method is that no one individual is competent to evaluate all or even most of the members of any given department. The idea behind my method is to ask philosophers to evaluate only those philosophers with whose work they have expert familiarity. With enough such data, one can arrive at an overall picture of a department's strength.

It was obvious from the start that this approach would face many obstacles, and as I proceeded I discovered still more. The two main obstacles, only one of which I take myself to have surmounted, were to gather adequate data from other philosophers, and to devise a method of aggregating individual scores to capture the strength of whole departments.

It is in the domain of data gathering that I abjectly failed. It turned out that I greatly disliked asking others to play my ranking game, and that many philosophers disliked the idea of doing it. Consequently, I have not made a very great effort to solicit feedback, and many of those who were solicited did not care to participate. In the end, then, my data comes from only fourteen philosophers, half of them from CU/Boulder. Although the group is reasonably diverse in areas of research, it no doubt fails to be a representative sample in many other ways. All the remarks to follow – and of course the rankings themselves – should be read with this fact in mind.

The aggregation problem turned out to be more tractable. As will be evident below, one gets distorted results if one tries simply to take the *average* ranking of each philosopher within a department, or if one tries simply to *sum* the rankings of each philosopher within a department. But one gets intuitively plausible results if one takes the average scores of the *best* philosophers within each department. My preferred method is to take the top-12 philosophers from each department. As it happens, this yields a result that is strikingly similar to the PGR results.

2. The Phil-12 Rankings

My data yields the following rankings for U.S. departments, averaging the scores of the twelve top-ranked philosophers in 34 departments.

Phil-12 Rankings of U.S. Departments

	Department	Avg. Score	PGR Rank
1	NYU	3.52	1
2	Princeton	3.27	3
3	Rutgers	3.25	2
4	Harvard	3.15	5
5	Yale	3.06	7
6	Berkeley	3.05	14
7	Michigan	3.04	4
8	USC	2.98	11
9	Pittsburgh	2.94	5
10	UNC	2.94	9
11	MIT	2.92	7
12	Columbia	2.89	11
13	Arizona	2.88	14
14	CUNY	2.82	14
15	Stanford	2.79	9
16	UCLA	2.78	11
17	Cornell	2.75	14
18	Notre Dame	2.72	18
19	Brown	2.64	19
20	Texas	2.63	20
21	UCSD	2.61	22
22	Colorado	2.59	24
23	Wash U/St. Louis	2.59	31
24	Chicago	2.55	20
25	UMass	2.49	24
26	Wisconsin	2.48	22
27	Indiana	2.44	24
28	Maryland	2.40	31
29	Duke	2.31	24
30	Penn	2.31	29
31	Northwestern	2.27	31
32	UC/Irvine	2.25	29
33	Syracuse	2.23	37
34	Ohio State	2.22	24

The Phil-12 list comes strikingly close to the Gourmet Report. Most departments come out within a few rungs of their PGR rank – the largest exceptions being Berkeley (up 8), Stanford (down 6), UCLA (down 5), Wash U (up 8), Duke (down 5), and Ohio State (down 10).

Following the PGR, my survey included the faculty of 26 other departments not listed above. I have not included them, because my informants were insufficiently informed about their faculties to generate reliable data.

Perhaps more significant than the rankings themselves is the distribution of numerical scores. The gap between NYU at #1 and Princeton-Rutgers, effectively tied for #2, is enormous, and it is astonishing that NYU has managed to hire twelve philosophers whose average score is over 3.5, on a 4-point scale. (See the following section for the significance of the numbers.) After the top three or four schools, there is a large cluster of schools with hardly any significant differences between them, and the numbers fall off from there, very gradually. Although the difference between being ranked #5 and #26 looks very large when so expressed, in terms of raw scores the difference is not really all that great – it is equivalent to the distance between #1 and #10. There is, overall, a great deal of parity among top research departments.

As I discuss below, other ways of aggregating the data yield results quite different from the PGR, but those results strike me as clearly less satisfactory.

3. Methodology

The departments selected for evaluation were those U.S departments evaluated by Leiter's 2011 PGR. For lists of faculty, I took the faculty lists that Leiter painstakingly assembled. To facilitate expert evaluation, it seemed crucial to divide up philosophers by area of expertise, and so each of the 1477 philosophers was assigned one or more AOS. (Here I relied heavily on some CU/Boulder graduate students – thanks!) The following AOSs were used, which again was taken from Leiter's divisions into specialization, though with some consolidation:

a. Ancient
b. Medieval
c. Modern [17th-18th-19th]
d. History of Analytic
e. Continental
f. Pragmatism
g. Language
h. Logic/Math
i. Science
j. Decision Theory
k. Epistemology
l. Mind
m. Metaphysics
n. Religion
o. Action/Will
p. Metaethics
q. Ethics
r. Applied Ethics
s. Political
t. Feminism
u. Law

v. Aesthetics
w. Non-Western

With each philosopher listed by AOS and school, on a sortable spreadsheet, I was ready to solicit rankings. Here are the instructions as they ultimately read (after various revisions):

> Rank only those individuals with whose work you are reasonably familiar and which you judge yourself to have sufficient expertise to evaluate. You may wish to limit yourself to philosophers within your area(s) of expertise, but you may stray outside that area as far as seems reasonable. You may not rank philosophers from your own department, but you may rank philosophers in departments where you have studied, if you feel you have sufficient perspective to do so. Do not assign rankings simply on the basis of second-hand reputation, no matter how well-established that reputation is. On the other hand, being "reasonably familiar" does not require anything like a comprehensive knowledge of a philosopher's work. Keep in mind that the goal is to establish a ranking for every philosopher listed. Rank in those cases where you feel well-enough informed to serve as a useful guide to others.
>
> For each philosopher you wish to rank, **assign any rational number between 0 and 4**. Think of these numerical rankings along the following lines:
>
> 4. One of the true elites; a member of the most select group of philosophers in the profession.
>
> 3. A leading contributor to the profession OR a relatively junior figure on the way to becoming one of the true elites.
>
> 2. A solid, significant contributor to the profession OR a relatively junior figure who shows signs of becoming a leading contributor to the profession.
>
> 1. A marginal contributor to the profession OR a relatively junior figure of uncertain worth.
>
> 0. Makes no scholarly contribution to the profession. Please mark as 0, rather than leaving blank, so that we can distinguish between low scores and lack of information.
>
> These guidelines encourage taking into account the future potential of junior candidates, but otherwise they intentionally refrain from giving any substantive direction as to how exactly the rankings should be applied. There will be doubtless be differences of opinion reflecting the different weight that might be put on factors such as **quantity** versus **quality,** or **quality** versus **influence.** Some may wish to focus exclusively on a scholar's written work, whereas others may weigh a scholar's overall contribution the profession. Some may think that every sub-field has its elite figures, whereas others may think that only figures working in core areas could count as category 4. Contributors are encouraged to distinguish between figures who have had a real influence on the issues and those who have merely achieved a popular reputation.
>
> Although philosophers are associated with an AOS, in order to facilitate expert evaluation, please keep in mind that **you are being asked to evaluate the philosopher as a whole, not qua a given area of expertise.** For philosophers who work in many different areas, this may require an extrapolation from merit in one domain to overall merit. If you feel unable to make the extrapolation, then of course you should not rank that individual.

Anonymity was guaranteed in three different ways:

1. Outside evaluators were asked to submit their rankings not directly to me, but to a CU/Boulder administrative assistant.
2. That assistant stripped the documents of all identifying information and forwarded them to me anonymously.

3. I am the only one who has seen the individual rankings of philosophers, and I have pledged not to share this information with anyone else. Only scores of whole departments will be made public.

Various further tricky questions needed to be adjudicated, in particular:

a. What about scholars from other units who make a contribution to philosophy? Leiter lists these folk separately, and leaves it to individual evaluators to decide what they count for. I too included them on my list, and they occasionally got ranked by my evaluators, but I decided not to count this information except in a few cases where the contribution seemed especially direct and important.

b. What about part-time faculty? Again, Leiter lists these folk, and leaves it to individual evaluators to decide how much it counts. I decided to count part-time faculty except in cases where Leiter lists them as having less than a half-time appointment.

c. What about folk who are close to retirement? Again, Leiter lists this information. I followed suit, but instructed evaluators to ignore it. (Why then did I bother to include it? Dunno.)

4. Discussion

Why on earth have I done all of this? I don't really know. It was just an idea that I had, that wouldn't go away – a bit like wanting to learn how to whitewater kayak. Unlike kayaking, though, I feel pretty confident that I've now gotten this out of my system.

There is a view out there that there's something bad about these reputational rankings, and in particular that the PGR has been bad for philosophy. I think this is quite wrong. A reliable ranking of departments allows excellence in the field to be rewarded. Departments doing things well can count on recruiting better faculty and graduate students, and receiving more resources from their own administration. Departments doing things poorly will accordingly have considerable incentive to do better. Without some such ranking, the predictable old hierarchies of the Ivy League etc. would continue to dominate the profession, with little regard for merit or industry.

If there is something bad about the PGR, the problem has always seemed to me to be that it is *the only thing out there*. So it seemed to me quite worthwhile to attempt an alternative ranking system. The fact that the Phil-12 Rankings arrives at results almost identical to the PGR is, it seems to me, extremely encouraging for the field. Certainly, it was very good news for me personally. If my limited data had yielded results quite different from the PGR, I would have felt obligated to pursue the project in much greater deal, little as I might have wanted to. But once I realized how similar the results were, I took this as an excuse to stop.

Next to Leiter's 300 or so participants, my paltry data look laughable. Even so, there is a case to be made that my results are the more reliable. One can gather the input of hundreds of people, but if those people are not well informed, the large numbers may not help. When I began soliciting evaluations from my colleagues, the most striking thing I learned is how little each of us knows about the profession as a whole. I knew this about myself, but I somehow supposed that others were wiser. Over and over, though, I got the response from other scholars that they wouldn't be of any help, because they don't know enough about the work of other philosophers. Moreover, when I did succeed in getting data, the data was strikingly limited. My sources were good in their own little areas, but rarely ventured much outside that area. And of course, on reflection, this is not surprising. How much have any of us actually read of the 1477 philosophers listed in my survey? (And this of course covers just the best research departments in the U.S. alone.) Those who rank for the PGR are required to look over lists of faculty – most of whose work they do not know – and arrive at some sort of overall score. How does one do that with any reliability? My results, though based on much smaller numbers, reflect an informed, expert assessment of individual philosophers.

And yet the results are nearly the same! What is going on? My results suggest an answer to the question of how the PGR rankings work. Evaluators, it seems, do not attempt to take into account whole faculties – most of whose work they know nothing of – but instead they evaluate departments based on the leading figures in each department. This is a task that many of us are capable of, for we are all largely familiar, at least by reputation, with the work of the great names in the field.

It is particularly encouraging that there is no obvious pattern to the differences that exist betweenthe Phil-12 Rankings and the PGR. One might have expected a general downward trend from schools with old and venerable reputations, and an upward trend from relative upstarts in the field. So far as I can see, this has not occurred.

5. Alternative Aggregation Methods

Even with information in hand about individual philosophers, it is not at all obvious how to turn that information into a measure of departmental quality. Let it be granted that the quality of a department supervenes on the individual worth of its faculty members. (This, as everyone knows, is very far from being the case. But I'll leave the study of those further ingredients to others.) Even so, what is the correct algorithm?

With my data in hand, it is a simple matter to try out other approaches. Suppose one wants to rank according to the average score of every philosopher in a department. Then one gets the following:

Ranking by Average Score of Whole Faculty

	Department	**Avg. Score**	**PGR Rank**	**Phil-12 Rank**
1	NYU	2.97	1	1
2	Princeton	2.85	3	2
3	MIT	2.77	7	11
4	Rutgers	2.70	2	3
5	Yale	2.59	7	5
6	Harvard	2.58	5	4
7	Berkeley	2.57	14	6
8	Michigan	2.53	4	7
9	Brown	2.48	19	19
10	UCLA	2.45	11	16
11	UNC	2.38	9	10

12	UMass	2.35	24	25
13	Columbia	2.33	11	12
14	UCSD	2.32	22	21
15	Pittsburgh	2.32	5	9
16	Arizona	2.30	14	13
17	USC	2.29	11	8
18	Stanford	2.27	9	15
19	Cornell	2.27	14	17
20	Texas	2.23	20	20
21	Duke	2.17	24	29
22	Wash U/St. Louis	2.17	31	23
23	Colorado	2.13	24	22
24	Indiana	2.13	24	27
25	Wisconsin	2.08	22	26
26	Penn	2.03	29	30
28	Northwestern	2.00	31	31
29	Maryland	1.93	31	28
30	Notre Dame	1.92	18	18
31	UC/Irvine	1.90	29	32
32	Syracuse	1.88	37	33
33	Ohio State	1.79	24	34
	Chicago[1]		20	24
	CUNY[1]		14	14

This does not on its face look crazy, but some strange things have happened. MIT, Brown, and UMass, with their small and uniformly strong faculties, have gone way up. Pittsburgh, USC, and Notre Dame, with their large and diverse faculties, go significantly down.

Is this as it should be? Here the question arises of what makes for a good department. Is it better to have a small and uniformly excellent program, or a larger but more inconsistent program? How would we even answer such questions?

I myself think the whole-faculty average method gets the wrong results, for two reasons in particular. First, if a department has one or more senior scholars who contribute nothing, and bring down the average ranking, this seems to me not very important to measuring the quality of that department. It is better simply to set these figures aside in evaluating the program. Second, if a department has made a number of recent junior appointments, that too will bring down the whole-faculty average rating. But clearly it is good to make junior appointments. If two departments have equally good senior faculties, and one of those departments makes an additional junior hire, this should make that department stronger. On the whole-faculty average rating method, it would make that department look weaker.

[1] I am unable to rank Chicago and CUNY using this method: Chicago because much of its diverse faculty could not be evaluated by my group; CUNY because its faculty list includes all the various city college departments, most of whose faculty were not known to my group of evaluators.

What if we take the sum of the whole faculty's scores?

Ranking by Summed Score of Whole Faculty

	Department	Avg. Score	PGR Rank	Phil-12 Rank
1	Rutgers	78.42	2	3
2	Notre Dame	72.79	18	18
3	Princeton	71.30	3	2
4	NYU	71.19	1	1
5	Pittsburgh	67.28	5	9
6	Michigan	65.77	4	7
7	Arizona	62.09	14	13
8	UNC	59.61	9	10
9	USC	57.23	11	8
10	Harvard	56.72	5	4
11	Columbia	55.96	11	12
12	Colorado	53.33	24	22
13	Berkeley	51.47	14	6
14	Yale	49.13	7	5
15	Texas	49.06	20	20
16	Stanford	45.38	9	15
17	UCLA	44.17	11	16
18	UCSD	44.13	22	21
19	Cornell	43.06	14	17
20	Maryland	40.49	31	28
21	Ohio State	37.66	24	34
22	Wisconsin	37.38	22	26
23	Wash U/St. Louis	36.87	31	23
24	Brown	36.49	19	19
25	Indiana	36.18	24	27
26	UC/Irvine	36.10	29	32
27	MIT	36.06	7	11
28	Syracuse	33.52	37	33
29	Northwestern	32.01	31	31
30	UMass	30.57	24	25
31	Duke	30.45	24	29
32	Penn	28.70	29	30
	Chicago[2]		20	24
	CUNY[2]		14	14

[2] See previous note.

Here departments that benefited from the whole-average method get hammered, with MIT falling to an absurd #27. Conversely, Notre Dame vaults to #2. These results speak for themselves, I should think.

It has been suggested to me, however, that there is something appealing about this method of measurement. After all, doesn't size matter? Do we not value quantity as well as quality? If what we want is a raw measure of philosophical goodness as aggregated on one college campus, then this whole-faculty sum is perhaps the way to go. But even if this is so in some ideal sense, it seems clearly wrong for any practical purpose. These rankings matter in the real world when it comes to things like choosing a graduate program, choosing where to accept a job or to spend a year on sabbatical, and choosing how much of a university's money to spend on its philosophers. Although there are clearly limits to how *small* a department can be and still be effective, there seems a point at which increasing size has nothing to do with excellence.

The Phil-12 Rankings capture something from both of these two flawed approaches. Departments get rewarded for uniform quality, as with the whole-average method, but without paying an undue price for junior faculty or unproductive senior faculty. Large departments benefit from the Phil-12 method too, because a department of twenty-some has the luxury of selecting its top group from among a much larger team. MIT, the smallest department being assessed (13 faculty), gets to leave only one player on the bench.

Obviously, the Phil-12 Rankings cohere much more closely to the PGR. On one standard statistical measure, the sum of the squared differences, the deviation from the PGR runs as follows:

Phil-12 avg:	465
Whole-Faculty avg:	1030
Whole-Faculty sum:	1502

The Phil-12 avg. also comes closer to the PGR than taking the top-8 faculty or the top-16 faculty:

Top-8 avg:	569
Top-16 avg:	503

If you're still reading, you may want to know about the top-10 and top-14, etc., but there are limits to my obsessiveness.

6. Pitfalls

i. It has been argued to me that the idea of a reputational survey, even one limited to expert evaluations, is misguided. Instead, I have been told, we should pursue an evaluation based on more objective data like publication record and citation frequency. I am not persuaded. There seems to me little correlation between the amount scholars publish (even in very good venues) and the quality of their work. It also seems to me that the frequency with which work is cited depends as much on academic trends and fashions as it does on philosophical quality. In contrast, there seems to me no better method of evaluating philosophical merit than to ask other experts in the same field. This is one of the main methods we use to make the most important academic decision of all, granting tenure, and I think we are right to rely on it so heavily.

ii. By ranking philosophers individually, one arrives at a very clear sense of the incommensurability of different programs. My team of evaluators had reasonably broad interests, but had almost nothing to say about the faculty of some of the schools on the PGR list. Rather than put these schools at the bottom of some comprehensive list, it is better not to rank them at all. One might think that better still would be to gather a still more diverse team of evaluators, capable of evaluating all kinds of different programs. This, however, would not work. My team was able to rank only one person, for instance, from among Penn State's faculty of sixteen. This, evidently, is a failing in my data. But suppose I *had* collected data from people enthusiastic about the Penn State faculty. Penn State might easily be a top-20 department. To take another example, Fordham (not on the PGR survey, and so not on my survey) has a wonderful group of medievalists whom I myself would score very highly, if anyone asked me. How high should Fordham go, just in virtue of being strong in medieval?

The point is that any department will have its aficionados who are keen, in all sincerity, to rank it highly. Rankings of the sort attempted here and by the PGR make sense only when limited to a homogeneous group of departments and evaluators. As it happens, most research departments of philosophy in the English-speaking world have very similar perspectives on the field, and have faculties distributed in much the same way over the same specializations. One can, then, readily compare them against each other. For those departments that have a different vision of the field, a wholly different survey would be required. Such a survey could of course be done according to the Phil-12 methods. But it would be pointless to use these methods to compare programs of wholly different kinds. My evaluators were as little able to evaluate Penn State's philosophy department as they would have been able to evaluate Penn State's physics department. Indeed, they could have done better with the physicists.

iii. From the start of this project, I was warned that it would face resistance from philosophers who were unwilling to evaluate their peers individually. Since I myself evidently do not share such reluctance, I did not fully credit these warnings, but they turned out to be quite right. One scholar who declined to participate sent me this articulate response, which I quote with permission:

> I find it very difficult to get myself to record rankings of individual philosophers in my field -- despite all the good that I know it can do. Not that I have no opinions on the ranking issue: of course I have opinions, as I'm sure we all do! It's just that, for some reason that I do not fully understand, the act of recording these opinions feels shameful to me.

My respondent does not doubt the usefulness of rankings in general, or of this one in particular, and does not deny having views about how individuals should be ranked. Even so, the sense of something "shameful" remained. This struck me as quite apt, because although I evidently do not share these scruples in the same way, I think it is this same obscure sense of shame that has made me reluctant to try harder to gather more data.

One of my colleagues kindly remarked to me, "You have better things you could be doing with your time, Bob." And so I do. Let me therefore bring this to a close.