METAPHILOSOPHY
Vol. 46, Nos. 4–5, October 2015
0026-1068

APPEARANCE AND REALITY IN *THE PHILOSOPHICAL GOURMET REPORT:* WHY THE DISCREPANCY MATTERS TO THE PROFESSION OF PHILOSOPHY

BRIAN BRUYA

Abstract: This article is a data-driven critique of *The Philosophical Gourmet Report* (PGR), the most institutionally influential publication in the field of Anglophone philosophy. The PGR is influential because it is perceived to be of high value. The article demonstrates that the actual value of the PGR, in its current form, is not nearly as high as it is assumed to be and that the PGR is, in fact, detrimental to the profession. The article lists and explains five objections to the methods and methodology of the report. Taken together, the objections demonstrate that the report is severely flawed, failing to provide the information it purports to and damaging the profession overall. Finally, the article explains how several modifications may improve the PGR so that it can more legitimately and equitably play the role it already plays.

Keywords: *Philosophical Gourmet Report*, philosophy, academic philosophy, rankings, philosophy Ph.D. programs, philosophy graduate programs.

The Philosophical Gourmet Report (PGR) is an opinion poll that bills itself as a ranking of philosophy graduate programs based on the quality of faculty. Brian Leiter, who earned his Ph.D. in philosophy at the University of Michigan and is now a chair professor of jurisprudence at the University of Chicago Law School as well as the director of the University of Chicago's Center for Law, Philosophy, and Human Values, handpicked his original slate of evaluators and has since asked them to recommend others. The evaluators are given faculty lists from philosophy Ph.D. programs, absent school names, and it is the task of the evaluators to rate each program on a scale of 0–5, instructed as follows:

> Please give your opinion of the attractiveness of the faculty for a prospective student, taking in to account (and weighted as you deem appropriate) the quality of philosophical work and talent on the faculty, the range of areas the faculty covers, and the availability of the faculty over the next few years. (Leiter 2011c)

> Please evaluate the following programs in terms of faculty quality, using the following scale: 5—Distinguished, 4—Strong, 3—Good, 2—Adequate, 1—Marginal, 0—Inadequate for a PhD program.
>
> You may use .5 intervals if necessary, but no scores higher than 5.0, and no smaller fractions, are permitted. **Do not check any box if you lack sufficient information to make an informed judgment about faculty quality.**
>
> **You should not evaluate either (1) your own department, or (2) the department from which you received your highest graduate degree (typically the PhD or the DPhil).** Those scores will be discounted.
>
> "Faculty quality" should be taken to encompass the quality of philosophical work and talent represented by the faculty and the range of areas they cover, with the two weighted as you think appropriate. Since the rankings are used by prospective students, about to embark on a multi-year course of study, you may also take in to account, as you see fit, considerations like the status (full-time, part-time) of the faculty; the age of the faculty (as a somewhat tenuous guide to prospective availability, not quality); and the quality of training the faculty provide, to the extent you have information about this. (Leiter 2011b)

When the PGR first appeared, it was a welcome resource for aspiring graduate students because it provides rankings not only of entire philosophy Ph.D. graduate programs but also of specialties within programs. At the time, it filled a gap in information for graduate students, purportedly giving them a better sense of where a program stands in the eyes of the profession, freeing them from inferring the reputation of a philosophy program from that of the school, and expanding their scope beyond the small set of professors available to advise them at their undergraduate institution.

But the survey's methodology has also come under severe criticism, for such things as built-in bias, a nonrepresentative sample of evaluators, and failure to consider things that graduate students most need to know (such as funding prospects going in and placement prospects coming out) (Saul 2012; Wilson 2005; Ernst 2009; Walker 2004; Wilshire 2002; Frodeman and Rowland 2009; Heck 2014; McAfee 2007, 2010a, 2010b, 2011, 2014; Wheeler 2012a, 2012b, 2012c). Despite these flaws, the report that is produced every few years has become immensely influential—to the point that top programs tout it as a measure of their success, and programs hire faculty with the intent of rising in the rankings (Saul 2012, 268; Wilson 2005; Ernst 2009; Heck 2014).

Because of the institutional status of the PGR in a significant portion of the profession, its influence must not be underestimated and is the reason I examine it. Assuming a motivation of rational self-interest on the part of philosophy Ph.D. programs, if taking a particular action by a program will have the effect of raising the program's ranking, it can be assumed that that action is more likely to be taken by a program. A ranking gives the appearance of an empirical statistic that a program can advertise to

prospective graduate students and, what is more important to administrators, to demonstrate quality nationally and internationally, which can translate into funding.[1] And the opposite applies: actions that do not lead to a rise in the rankings, or worse, that may jeopardize a current spot in the rankings, will be avoided as far as possible. In what follows, I list and explain five objections to the methods and methodology of the PGR. Taken together, they demonstrate that the report itself is severely flawed, failing to provide the information it purports to and damaging the profession overall. Finally, I explain how the PGR may be improved so that it can more legitimately and equitably play the role it already plays.

Critical Flaws in the PGR

A. Selection Bias I: Sampling Methods

In the social sciences, the usual standard for generating a survey sample from a more general population that is the target of the study is to acquire a sample that is sufficiently representative of the larger population. With a nonrepresentative sample, one risks magnifying aspects of the sample that are different from the population in general or missing aspects of the general population that are absent in the sample. Any such difference between the sample generated by the selection process and the general population is called a selection bias. For instance, in conducting a presidential poll across a state, one cannot poll the entire population, and so one focuses on a small sample. If one selects from just one geographic area of the state or from just one age group, a selection bias will be introduced, and the results will most likely differ from a more representative sample, thus threatening the soundness of any generalization made from the results.

The goal of any sampling procedure is to first mitigate the possibility of obtaining a biased sample stemming from selection bias. The way this is generally accomplished is to randomize the sample. How to randomize a sample is a topic of much research and discussion, and methods vary depending on the field, the goals of the study, and the population being studied. The best method, when feasible, is simply to survey the entire population, getting what's called a saturation sample. That way, one's sample and the general population are identical, eliminating the possibility of overt selection bias.[2] Saturation samples are possible when the population is of a limited size and the method of obtaining data across the

[1] Fully half of the twelve programs at the top of the U.S. PGR ranking either explicitly or implicitly highlight their high PGR rank on their department webpages.

[2] There are still ways of engendering covert selection bias, such as wording the questionnaire in such a way that it encourages or discourages the participation of some subgroups more than others.

population is cost-effective.[3] For instance, if a university wants to understand something about the working conditions among its faculty, it can simply create an online survey and e-mail the link to all faculty.

When evaluating a completed poll with regard to its sampling procedures, one looks first to the description of its methods. The PGR does not specify how respondents are recruited, stating only: "In October 2011, we conducted an on-line survey of approximately 500 philosophers throughout the English-speaking world; a little over 300 responded and completed some or all of the surveys" (Leiter 2011b).[4] The description then goes on to list all of the respondents, providing the name of the Ph.D.-granting institution for each, the name of the institution at which each respondent worked at the time of the survey, and the general area(s) of philosophy in which each purports to do research. No other demographic information is provided, let alone comparisons with the general population of working philosophers. Because there are 147 graduate programs and 841 undergraduate programs in philosophy in the United States (Romaniuk 2012) and probably an average of between five and twelve professors in each program, it is obvious that the PGR is not canvassing the entire population of the six thousand or so working philosophy professors in America.[5]

If we look elsewhere, however, we do find some information. On his personal blog, Leiter states: "The whole rationale for a 'snowball' sampling procedure, which is what the PGR uses, is to garner informed, expert opinion" (Leiter 2012). Here, Leiter makes it clear that he is using a specific sampling procedure and identifies it as "snowball sampling," which is also known as chain-referral sampling. Although snowball sampling is a sampling method used in sociological research, it "contradicts many of the assumptions underpinning conventional notions of sampling" and "violate[s] the principles of sampling" (Atkinson and Flint 2001, 1). The reason it is occasionally used is as an expedient way to access a hidden population, such as social deviants (drug users, pimps, and the like), populations with very rare characteristics (such as people with rare

[3] Both of these criteria, while impossible at the time of the origin of the PGR, are now practicable in the case of the PGR, especially considering, first, that online surveys are easily accomplished now and, second, that the survey is financed by a large publishing house (Wiley-Blackwell; the fee is "in the low five figures" (Wilson 2005)). A saturation sample, or census, in quantitative research is distinct from a sample that reaches data saturation in qualitative research. Even a full census does not guarantee representativeness if a significant portion of queried respondents elect not to respond or respond incompletely. Demographic analysis is always recommended.

[4] This critique of the PGR was written when the 2011 edition was the most current edition. The 2014–2015 edition came out when this article was under review. Although the 2014–2015 edition has added Berit Brogaard as coeditor, she says in personal communication that she came to the project during the data collection phase and that she endorses the report as is. The methods did not change from the earlier edition to the more recent edition.

[5] A conservative estimate could assume five permanent full-time faculty for each undergraduate program and twelve for each graduate program: $(5 \times 841) + (12 \times 147) = 5{,}969$.

diseases, interests, or associations), or subsets of populations associated in idiosyncratic ways (such as networks of friendships).[6] Leiter does not offer a rationale for using this method. Philosophers are neither social deviants nor difficult to find, as every philosophy program's faculty list is public information.

It must be, then, that Leiter has refined the general population of philosophers down to only those at an "expert" level: only expert philosophers would be competent to judge other philosophers. If that is the case, then Leiter has chosen the wrong nonprobabilistic sampling technique. For expert sampling, one would want to use a method known as purposive sampling, or judgment sampling.[7] In this method, a researcher establishes the criteria that are descriptive of all potential experts in a pool and then seeks people who fit these specific criteria. Chain referral may be used as part of this process, but the purposive sampling procedure is larger than just that and necessarily involves screening to meet specified criteria. It is possible that Leiter used purposive sampling implicitly. If he did, it would behoove him to divulge the criteria he used to include some philosophers in the pool of experts and, more important, to exclude others. After all, the very earning of a Ph.D. is the academic standard for expertise. If Leiter thinks that some experts are more expert than others, then he should make his criteria of special expertise clear in his methods section. What's more, purposive sampling, if it purports to represent a larger population, is still flawed. Maria Tongco offers the following advice to researchers who choose to use purposive sampling: "In analyzing data and interpreting results, remember that purposive sampling is an inherently biased method. Document the bias. Do not apply interpretations beyond the sampled population" (2007, 151).

[6] Drăgan and Isaic-Maniu 2012 (provides extensive citations of studies that have used snowball sampling); Atkinson and Flint 2001 (explains that snowball sampling is used primarily for qualitative research [e.g., interviews] and for research on the sample population itself); Biernacki and Waldorf 1981 (provides a case study of snowball sampling and the methodological issues encountered); Erickson 1979 (discusses the benefits and limits of snowball sampling; distinguishes it from other chain sampling methods); Coleman 1958 (examines snowball sampling and networks). Reading these articles, one realizes that the PGR actually does not use chain-referral sampling in the standard way. To imagine the use of chain-referral sampling in the standard way, you have to imagine a population hidden to you—you want to survey the members of the population, but you can't find them. First, you find one or two, then you ask them to identify more, then you ask the new ones to identify more, and so on. Since philosophers are easy to find, the only way Leiter did anything like snowball sampling is if he looked for, as he says, "research-active" philosophers (see my Appendix 1). He would have identified a few on his own (using what selection criteria, we can only guess), and then he would have asked those "research-active faculty" to identify others (again, by unstated criteria), and so on. But are research-active philosophers really so hard to find? Of course not—they are, by definition, published. One could conclude that Leiter's snowball is not about finding a hidden population but about excluding a large portion of an otherwise prominent population, as we shall see.

[7] For an excellent overview of purposive sampling as a technique, including numerous examples from prior literature, see Tongco 2007.

It is important to keep in mind that this talk of snowball sampling and experts does not appear in the PGR itself. In the PGR's "Methods & Criteria" section, there is, in fact, no mention of selection criteria at all beyond the vague "research-active" faculty, which would describe the majority of all working philosophers, for the simple reason that it is part of the job of being a professor that one must maintain an active research agenda.[8] Instead of detailing clear criteria of selection, the survey is described as "an on-line survey of approximately 500 philosophers throughout the English-speaking world." Any reasonable reader would understand this to imply an attempt at garnering a sample that is representative of the opinions of a larger population of philosophers, most likely achieved through a general call for participation. As we see from the analysis above, however, this simply is not the case. The PGR sample suffers from selection bias right from the beginning, and instead of representing the profession at large, it is merely a reflection of one person's overriding opinion—the creator of the snowball.

The effect of snowball sampling on the sample is that any bias in the creator may be propagated throughout the sample, to the point that even though evaluators are evaluating from their own informed opinions, their presence is a function of the original biases of the creator. Only evaluators judged consistent with these biases will be present, effectively making each evaluator a statistical clone, and tool, of the original chooser.

So far, we have discussed ways to assess the possibility of sample bias in the PGR by considering the sampling method that was used, and the conclusion up to here is that serious and damaging bias is not only likely but virtually unavoidable. Another way to assess the possibility of sample bias is to look at the sample itself with respect to the results. In a well-executed poll of this sort, one would expect that there would be no correlation between the number of evaluators hailing from a particular school and that school's rank. We wouldn't want, for instance, the school with the second-highest number of evaluators to rank second and the school with the seventh-highest number of evaluators to rank seventh, and so on. Any such correlation would immediately call into question the validity of the poll's conclusions—and the stronger the correlation, the worse the poll would look.

In Figure 1, we see just such a correlation, and it is impressively high—61 percent (R^2) of the variance in program scores is accounted for merely by the number of evaluators hailing from each program. This surprisingly large correlation is exactly why snowball sampling is problematic. What's worse is the fact that the PGR instructs evaluators not to evaluate their own current program or their Ph.D.-granting program, claiming to "discount" such votes (see the PGR instructions above). Either the discounting is not happening, and the survey is a victim of

[8] For further discussion of the PGR's "Methods & Criteria" section, see my Appendix 1.

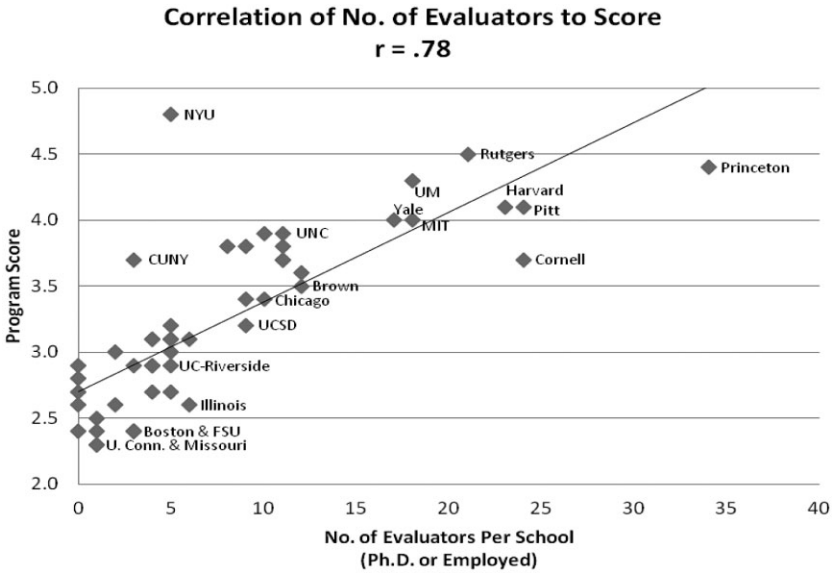

FIGURE 1. Correlation of number of evaluators associated with a school and the overall PGR score of that school.

self-centered voting, or the correlation illustrates Atkinson and Flint's remark that snowball sampling "violate[s] the principles of sampling."[9] Either way, it is difficult to take the results seriously.[10]

Figure 1 reveals another, related, aspect of the sampling that is equally troubling. Notice that the small group at the top right, bounded by Yale, Rutgers, Princeton, and Cornell, accounts for approximately half of all votes in the PGR. These eight programs in a tight geographical area are in effect driving the ratings of the PGR. Thus, the PGR is not a survey of philosophers generally about the quality of programs generally but a survey of a small, select group of programs about each other and about what they think of other Ph.D. programs.

As we see, then, this kind of selection bias is lethal to any poll that either purports to be representative of any larger population or to have

[9] A defender of the PGR could claim that this result simply confirms that the report has the best philosophers as evaluators, but that is, of course, circular reasoning: we know the best programs because the best evaluators tell us which are best, and we know the best evaluators because they come from the best programs as determined by the self-same evaluators.

[10] Leiter (2011b) observes that the results of the PGR are stable from iteration to iteration, despite the differences in evaluators from iteration to iteration, as if that might lend it some validity. In a snowball sample, it should come as no surprise that the creator's biases are propagated from one iteration of the sample to the next.

any special expert status. What purports to be a solid measure of faculty quality and reputation broadly is really an idiosyncratic collection of opinions about faculty and reputation, from which no reliable conclusions can be drawn.

The limits of the PGR methodology do not end there.

B. Selection Bias II: Expert Opinion and Overall Ranking

The attentive reader may already have noticed an obvious contradiction in the PGR's methods. The survey asks philosophers who are experts in specific specialties to rank entire programs, not just those faculty within specific specialties. Leiter defends the PGR against such a criticism by claiming that his process is an aggregation of information—that he is obtaining the multiple perspectives of experts in narrow fields and that when combined, these different perspectives give an accurate overall assessment.[11] In fact, however, he has the proper sequence of aggregating his data exactly backward. If one wants to aggregate expert opinion, employing diverse experts on diverse topics, one *first* asks the experts to give their opinions within their area of expertise, *then* one takes these separate results and aggregates them numerically. If one asks all experts to evaluate all areas of expertise, they are no longer experts for the vast majority of cases in which they are making evaluations.[12]

The result of Leiter's method is *not just* an amorphous mixture of opinions, however. Given that the "expert" evaluators are competent to evaluate only those philosophers working in their own specialty and yet are evaluating entire departments, it follows that evaluators will favor those departments that have more faculty working in their own specialty. Thus, the results are not about the quality of departments but about the number of evaluators in any particular specialty. I examine concrete ramifications of this below, in section F.

Leiter uses the term "expert" four times in his short defense of the PGR methods on his blog (but not at all in his "Methods & Criteria" section of the report), saying that they are "experts in many different fields [of

[11] Leiter says, "No kidding! That's why we do a survey of hundreds of experts in many different fields. A good survey aggregates a lot of partial knowledge to give us a more complete picture. If any one individual could know as much as the 300 philosophers who complete the PGR surveys, then we could just ask that person, and be done" (Leiter 2012).

[12] One might claim in Leiter's defense that I have misconstrued his use of the term "expert," and that he did not intend it to mean expert in a particular philosophical specialty. But there are only two other possible construals. First, the meaning of "expert" could be that the survey's philosophers are all experts in philosophy broadly, making them perfectly suitable for performing the overall ranking. But then the ranking would not properly be called an expert ranking and should instead be called a *self-ranking*, and, as discussed, there is no rationale for excluding other such experts from the evaluation pool. Second, "expert" could mean superior to other specialists. If that is the case, then, as I explained above, specific criteria for such expertise should be detailed, and it should be made clear how these particular evaluators meet those criteria while other potential evaluators would not.

philosophy]" (Leiter 2012). The only criteria of expertise that can be gleaned from his statements is where he says, "Research universities—in their hiring and tenure decisions—are based on the premise that the opinion of experts is what matters. We have nothing else to go on." He seems to be saying that the hiring and tenure evaluation committees of research universities use experts in the same way that he is using them. This supposition is false, and we shall see exactly why in section D below. For now, let us run a scenario to see how absurd this claim actually is.

Leiter's claim is that one self-referred person is qualified to select a slate of referees and that a portion of that slate (the "Advisory Board") then recommends other referees. Then, for all universities, this slate acts as a group to judge every hiring and tenure candidate. So, if I were the candidate for tenure, neither I nor my university would be allowed to voluntarily provide input to the evaluation committee. Even though Leiter's slate includes only three referees who work in Chinese philosophy (my field) and excludes many qualified others, his entire slate would provide input on my candidacy, and no one else would be allowed to. That is, of course, not at all how hiring and tenure committees at research universities work. Each slate of referees is selected with particular relevance to the program or the candidate. To extend the hiring and tenure analogy to the PGR in a more appropriate way, one would allow each program evaluated in the PGR to select its own potential slate of evaluators, at least as a starting point. There would be no prime referee or board of referees above all others, excluding all others, and judging all others.

To summarize the conclusion of this section, either the PGR's experts are not actually acting as experts in doing their evaluations (because they are working outside their specialties in evaluating entire programs) or the pool of experts is artificially narrowed by unannounced criteria, thereby excluding many other potentially qualified experts who may hold different, but equally legitimate, expert opinions.

C. Selection Bias III: Underrepresentation of Methodological Continentalists

The PGR has consistently been criticized for a bias toward Analytic philosophy and against Continental philosophy. Leiter admits a bias toward Analytic philosophy but not against Continental philosophy (Leiter 2011a). This is because he defines "Analytic philosophy" as a style (with certain presuppositions, preferences, inspirations, and models) and "Continental philosophy" as "demarcating a group of . . . philosophers" (Leiter 2011a). Following some prominent views among Continental philosophers, such as a coauthor of Leiter's,[13] let us, for purposes of analysis,

[13] In contrast to Leiter's definition of "Continental philosophy" in the PGR, Michael Rosen (1998), coeditor with Leiter of the *Oxford Handbook of Continental Philosophy* (2007),

define a new term: "methodological Continental philosophy," or MC for short. This brand of doing Continental philosophy also has its own style of doing philosophy that has its distinct presuppositions, preferences, inspirations, and models. In contradiction to his blog, Leiter refers to this brand of philosophy in his own published work (Leiter and Rosen 2007). If the PGR is a comprehensive ranking of graduate programs, as it purports to be, then it will include MC programs as well. Traditionally, some of the most prominent MC programs have been DePaul, Duquesne, and Emory, none of which ranks in the PGR.

Looking at the list of evaluators is also helpful in clarifying this bias. Of the twenty programs that are ranked specifically within the two specialties of Continental philosophy, ten (50 percent) do not have evaluators representing them in the PGR. This is a very high ratio compared to other programs that make the overall ranking, in which only four programs (out of fifty-one; 8 percent) don't have votes. I criticize the PGR above for its lack of a representative sample in that it leans very far toward ranked schools—the more evaluators, the higher the rank. For Continental programs, it leans too far in the other direction—not enough ranked schools have evaluators. Both situations point toward a serious sampling bias that compromises the legitimacy of the results.[14]

There is yet another way to get at this particular bias. There are eleven programs that make both the overall ranking and the Continental specialty ranking, meaning that there are nine programs ranked in one of the specialties of Continental philosophy but not ranked overall. Of the eleven that are ranked overall, although five of them are at the very top of the Continental ranking, none breaks the top ten of the overall ranking. The

describes the tradition explicitly in terms of methodology. In the first few pages of Rosen 1998, he highlights four of what he calls "recurrent issues" that define the field, each of which has a core methodological component: (1) the method of philosophy; (2) the limits of science and reason; (3) the influence of historical change on philosophy; and (4) the unity of theory and practice. A quote from Leiter and Rosen's Introduction to their handbook states the point clearly: "Where most of the Continental traditions differ is in their *attitude* towards science and scientific methods. While forms of philosophical naturalism have been dominant in Anglophone [Analytic] philosophy, the vast majority of authors within the Continental traditions insist on the distinctiveness of philosophical methods and their priority to those of natural sciences" (2007, 4). This is in contrast to Analytic philosophy, which often sees its methods as consistent with, and on the same level as, those of the natural sciences. Notice that there is no mention of the Continental tradition being defined in relation to a particular set of authors.

[14] In the PGR's defense, it may be that Leiter did include MC programs in his originally invited list of five hundred philosophers, and that the MC philosophers elected not to participate. Leiter surely can't be blamed for that. Actually, it is the responsibility of the creator of the survey to ensure a representative sample, not the responsibility of the respondents. The creator of the survey should always, on publishing the results, include a discussion of the limitations of the methods, the results, and the conclusions that can be drawn from them. Leiter includes no such discussion in the report. For a discussion of the bias even in the makeup of the evaluators of the PGR's Continental specialties, see Protevi 2011.

reason for this can be traced to a bias against the area of history in the overall ranking. The PGR groups all philosophical specialties into general areas (see my Appendix 2) and classifies Continental philosophy specialties under the area of history, which, as we shall see in section F.1 below, has less influence than other areas in a program's overall ranking.

The takeaway from this section is that there is no explicit mention in the PGR's methods that Leiter excludes certain kinds of philosophy or ways of doing philosophy or discounts them in any way. And yet this section demonstrates that his methods are structured such that he does both—he excludes MC evaluators, and by so doing discounts MC programs in the overall ranking. And this is just one subdiscipline that is underrepresented. There are many others, as I discuss below.

D. Methodological Flaw I: Misapplication of the Expert Committee

Up to now, we have considered bias only in the process of selecting the sample of evaluators for the PGR, and we have found three clear and devastating angles from which to understand why the selection method used in the PGR cannot possibly yield the kind of results that the PGR purports to offer—"measures of faculty quality and reputation" (Leiter 2011d), full stop.

Let us presume, for the sake of argument, that an expert committee is warranted, and let us examine the PGR's use of such an "expert committee." One of the purported strengths of the PGR is that it relies on experts for opinions. This kind of introduced bias might be valuable and even recommended—for instance, if one wants to judge whether a medical procedure is safe. We want experts to provide their opinions when expertise is required for a sound assessment, and we would not insist on getting a representative sample of all such experts; instead, we would settle for a small number of experts. We see this all the time in academia. We have Ph.D. committees, tenure review committees, grant committees, and so on, which are formed for the purpose of providing expert evaluation. And for none of these do we insist on getting a representative sample. Sometimes a sample of just two is enough, as in the case of some peer-reviewed publications.

So, when the PGR draws up a slate of more than five hundred specialists, some three hundred of whom respond, why should we not consider it another example of an academic expert committee—a large and, seemingly, diverse one at that? We have already covered part of the reason—namely, the introduction of bias into the selection process. But why is risk of bias unacceptable in the PGR and not on committees that are so much smaller (and thus even more subject to bias)? First, we have to distinguish between the two different kinds of committee just mentioned. One was the medical-expert kind of committee that is evaluating empirical evidence to offer recommendations according to stipulated criteria. That, of course, is

not happening in the case of the PGR. There are no stipulated criteria, so one cannot regard a committee, however large, as offering any sort of valid empirical evaluation. Thus, the PGR expert committee is not comparable to a medical-expert committee.

The second kind of expert committee is the referee kind, which involves judging the academic merit of a scholar or a scholarly piece of work. We all know that such judgments are naturally biased and that a submission that is accepted by one journal or press could have been rejected by another of equal standing. The simple fact is that in the world of academic publishing there is no better alternative to this type of committee. One can't send every article or book manuscript on epistemology to all, or even to a statistically significant random sample of all, working epistemologists. The logistics and the workload would be impossible. We rely, instead, on ad hoc arrangements as a necessary expedient. If we accept bias in academic committees because there is no better alternative, why not do the same for the PGR? The reason is that the logistics are entirely different. The PGR survey is undertaken only once every few years, and the online survey already exists. There is no practical impediment to moving to a valid sampling procedure.

Perhaps that point came too quickly. The reason that the PGR should not use an ad hoc committee of expert evaluators, even though such committees are often used in academia, is that it does not need to. It could just as easily use a valid sampling procedure. Using a nonrepresentative sample and then generalizing from it is misleading. As quoted above, the PGR says: "This report ranks graduate programs primarily on the basis of the quality of faculty. In October 2011, we conducted an on-line survey of approximately 500 philosophers throughout the English-speaking world" (Leiter 2011b). There is no reason for anyone reading this claim to suspect that the sample is not representative of the entire population of working philosophers or therefore to suspect that the conclusions drawn from the sample cannot be generalized across the entire population of philosophers. And yet such a supposition would be flatly wrong. One must again attend to the fact that the sample used by the PGR is as notable for those that it excludes as for those that it includes. The simplest thing for the PGR to do to improve its validity would be to open up the evaluation pool to anyone listed on a philosophy program's faculty webpage. Given the electronic resources that Leiter has already mastered, getting the word out would be neither difficult nor time-consuming.

E. Methodological Flaw II: Area Dilution

Even if the sample used by the PGR were representative of the larger population of philosophers, it could still be used in a way that would introduce bias. The PGR does this as well, compounding bias by adding bias on top of bias.

Consider the areas listed by the evaluators as their own areas of research.[15] Of the 305 individual evaluators, approximately two-thirds evaluated in one area only, and one-third evaluated in more than one (predominantly two). Because assumptions and methods of one area can be distinct from those of other areas (for instance, in the area of history, more emphasis may be placed on the hermeneutic process than in metaphysics and epistemology [M&E]), if one area is represented significantly more and another significantly less in the evaluator pool, it can mean that the methods and assumptions of the dominant area can eclipse those of the other.[16] This would especially be the case if a person who works predominantly in one area (say, M&E) also lists another area (say, history), in which case the assumptions and methods of the first would likely be brought to bear on the second.[17] If area crossovers were evenly spread out across all areas, no particular area would stand at a disadvantage. If there were an imbalance, however, minority areas would be diluted in their influence in the overall ranking.

We can calculate a *degree of dependence* for each of the four individual areas. The more often an area is evaluated by evaluators who list only that area as their area of expertise, the lower that area's degree of dependence and the higher the *degree of independence*. The more often an area is evaluated by evaluators who list more than just that area as within their expertise, the higher the degree of dependence and the lower the degree of independence. In Table 1, we see (column D) that value and history are nearly twice as likely as M&E to be influenced by another area. Furthermore, we see (highlighted in column E and comparing the numbers in E with the numbers in columns F, G, and H) that the single area that is most likely to do the influencing in both cases is M&E. Statistically, this is not surprising, because M&E is by far the most well-represented area. This fact, however, does not lesson the effect of its dilution of the other areas with which it is found in combination.

These results demonstrate that for the most part M&E functions as an independent area while also influencing the areas of value and history, thereby implicitly compromising the independence of value and history in the rankings. Overall, this means that M&E has a disproportionate influence on the results. We see, then, that not only do we begin with a bias in

[15] For an explanation of areas and specialties, see my Appendix 2.

[16] See Dotson 2012 for a detailed argument against the exceptionalism of Analytic methodology and a "culture of justification" that prioritizes it over other kinds of philosophical praxis.

[17] If one wonders whether this really happens, Leiter confirms it: "Most evaluators are asked to evaluate more than one area, and inevitably that means they evaluate areas in which they don't necessarily work primarily" (Protevi 2011, first outside comment; it is not obvious at first that this "Brian" is Brian Leiter, but reading through all of the comments further down, it is easy to confirm that it is.)

TABLE 1. Dependent and independent areas.

A: Area	B: No. of Times Area Listed with Another Area	C: No. of Times Area Listed in Total	D: Degree of Dependence: Ratio of A : B	E: No. of Times Area Listed with M&E	F: No. of Times Area Listed with Value	G: No. of Times Area Listed with History	H: No. of Times Area Listed with Other
M&E	61	185	0.33		25	37	5
Value	48	94	0.51	25		20	10
History	55	100	0.55	37	20		3
Other	15	17	0.88	5	10	3	

Degree of dependence refers to a particular ratio (represented as a decimal in column D). The numerator of the ratio is the number of evaluators who list that particular area plus at least one more area (column B; for M&E, it is sixty-one). The denominator of the ratio is the total number of times that a specific area is listed by the evaluators (column C; for M&E, that number is 185). Columns E through H show the number of times that a particular area is listed in combination with the area in column A (value, for example, is listed in combination with M&E twenty-five times). If the degree of dependence for an area were 1, that would indicate that every evaluator who lists that area also lists another area. If the degree of dependence were 0, that would indicate that the area is never listed by any evaluator in combination with another area, and is thus entirely independent of other areas. (The sum of columns E through H for each row does not exactly equal the corresponding number in column B because a small number of evaluators list three areas, the accounting of which is necessary to achieve accurate numbers in E through H but not necessary in achieving the numbers in column B.)

which MC is underrepresented, as I demonstrated in section C, but now it is compounded by being diluted by M&E. Likewise, value is being diluted, and the poor area of other has virtually no independence at all.

F. Evidence of Bias: Assessing the Results

Here, we move from examining flaws in the PGR's overall methods, which have been shown to be entirely unreliable, to demonstrating that the results are, themselves, flawed. Both approaches can then be understood as reinforcing the conclusions of the other.

F.1. PGR results demonstrate a bias against history. It has been demonstrated above that the PGR uses the dubious method of asking experts in narrow fields to evaluate the overall quality of programs. As I mentioned, there is a valid way to aggregate such information, which is to take the specialty scores that are done individually for each program by small panels of experts within specific specialties and then simply add them up. The program that scores highest for the sum of all ranked specialties gets the highest overall score.[18] I undertook such a mathematical aggregation, taking all the specialty scores for each program, as provided by the PGR, summing them for each program, and then ranking the programs accordingly. The difference between the overall ranking and the mathematically aggregated ranking is quite large, with the average change in rank being four spots (Table 2).

In a method as ill defined as the overall ranking, such a difference between it and the mathematical aggregation is to be expected, and with the data of specialty scores at hand, one wonders why Leiter would persist in using the overall ranking.[19] Still, let me state clearly what is wrong. The rankings of the PGR give the illusion of a kind of numerical precision, an empirical toehold in a subjective world of judgment. But is MIT ranked seventh or fourteenth? Is Boston University thirty-seventh or forty-fourth? Is UCLA eleventh or nineteenth? Is the University of Pennsylvania twenty-third or twenty-ninth? Is Notre Dame eighteenth or fourth? There is no precision to the overall ranking and, therefore, they are of

[18] I haven't yet explained a second way that Leiter does the rankings, partly because he doesn't explain it. In his results, he provides the names of evaluators assigned to specialty panels, and he provides rankings of departments according to each specialty. For instance, a program that ranks sixteenth overall may rank first for general philosophy of science and ninth in ancient philosophy. The specialty ranking appears to be an entirely distinct process and to have no mathematical relationship to the overall ranking. In response to multiple queries from me about this issue, Leiter has not provided an explanation beyond referring me to the report itself.

[19] See Appendix 2 for a potential rationale for preferring the overall ranking.

TABLE 2. Difference in rank for each program when shifting from overall scoring to mathematically aggregated scoring

A: PGR Overall Rank	B: Rank by Mathematical Aggregation of Specialty Score	C: Difference in Rank from A to B	D: School	E: Mathematically Aggregated Specialty Score
1	1	0	New York University	91
4	2	2	University of Michigan, Ann Arbor	89
3	3	0	Princeton University	87
18	4	14	University of Notre Dame	82
5	5	0	Harvard University	79
11	6	5	Columbia University	71.5
5	7	–2	University of Pittsburgh	71
2	8	–6	Rutgers University, New Brunswick	70.5
9	8	1	University of North Carolina, Chapel Hill	70.5
9	10	–1	Stanford University	69
7	11	–4	Yale University	68
14	12	2	University of California, Berkeley	66
11	13	–2	University of Southern California	59
7	14	–7	Massachusetts Institute of Technology	58.5
14	14	0	City University of New York Graduate Center	58.5
19	16	3	Brown University	57
14	17	–3	University of Arizona	55.5
24	18	6	Indiana University, Bloomington	55
11	19	–8	University of California, Los Angeles	53.5
14	20	–6	Cornell University	50
24	20	4	Ohio State University	50
22	22	0	University of California, San Diego	47
29	23	6	University of Pennsylvania	45.5
20	24	–4	University of Chicago	44
22	25	–3	University of Wisconsin, Madison	43
24	26	–2	University of Colorado, Boulder	42.5
20	27	–7	University of Texas, Austin	42
29	28	1	University of California, Irvine	40.5
31	28	3	University of California, Riverside	40.5
36	28	8	Georgetown University	40.5
24	31	–7	Duke University	36
24	32	–8	University of Massachusetts, Amherst	35.5
31	33	–2	University of Maryland, College Park	34
31	34	–3	Northwestern University	32.5
44	35	9	University of Minnesota, Minneapolis-St. Paul	30.5
31	36	–5	University of Miami	29
44	37	7	Boston University	25.5
37	38	–1	Syracuse University	24.5
40	39	1	Carnegie-Mellon University	23.5
31	40	–9	Washington University, St. Louis	22.5
44	41	3	University of California, Davis	22
37	42	–5	University of Virginia	19.5
37	43	–6	Johns Hopkins University	19
44	43	1	Rice University	19
40	45	–5	University of California, Santa Barbara	18.5
50	46	4	University of Connecticut, Storrs	17.5
43	47	–4	University of Washington, Seattle	16.5
40	48	–8	University of Illinois, Chicago	16
50	49	1	University of Missouri, Columbia	15
44	50	–6	Florida State University	11
44	51	–7	University of Rochester	6

TABLE 3. Correlation of area score to change in rank

Area	r value
M&E	0.175
Value	0.1
History	0.405
Other	0.032

little of value.[20] Also, keep in mind that we are still working within the arbitrarily circumscribed world of the PGR categories. Imagine how different the landscape would look if the ranking were done by specialty scores and if other specialties, such as environmental ethics, existentialism, or Indian philosophy, were allowed a justly inclusive role.

If one were to object and say, well, the mathematical aggregation is so crude, it doesn't account for the size of departments, for focused strengths, and so on. Well, neither does the overall ranking, which has no modalities at all and is just a black box that spits out a number with no rhyme or reason.

One can now ask if there is anything that accounts for the movement of schools up or down the ranking when shifting from the overall score to the mathematically aggregated score. In other words, are the overall scores favoring or disfavoring any dimension of a program in particular? There is evidence that history is being systematically discounted.

If one takes the change in rank from overall rank to mathematically aggregated rank and runs a statistical regression for each area score for each program, the result is that there is no sizable correlation between change in rank and area score—except in the area of history (Table 3). It turns out that statistically the higher a program's score in history, the higher its mathematically aggregated rank compared to its overall rank. In other words, specialists in the area of history, compared to specialists in other areas, are not being sufficiently recognized by the PGR evaluators in the overall ranking. The size of the history correlation, while quite a bit larger than the others and, therefore, indicative of a difference, is not itself impressive. However, when combined with the realizations that (1) MC evaluators are underrepresented, as I showed in section C, and (2) history is diluted as an area, as I demonstrated in section E, it becomes that much more impressive.

Regressions were also run between area score and overall rank and area score and aggregated specialty rank, then compared. The only area that

[20] Kieran Healy (2012a) did an analysis of the overall ranking of programs by breaking the evaluators into categories according to specialty. He found wide variation from one specialty to another in their rankings for most of the programs. See his third and fourth figures.

showed a significant change between overall and specialty rankings was the area of history, in which R-squared rose from 0.36 for the overall ranking to 0.51 for the specialty ranking—yet another way of showing that history is devalued in the overall ranking.

F.2. PGR Demonstrates a Bias Against Other. The structural biases in the PGR against the area other are obvious. There are only three specialties in other, and there are a mere seventeen evaluators, for only two of whom is it their sole area. So not only are they badly outnumbered overall, they don't even have an independent area, so dominated are they by the other areas.

It is worth looking a little more closely at the division of specialties in the PGR. Under M&E, not only is there a specialty called "General Philosophy of Science," there are also specialties identified as "Philosophy of Physics," "Philosophy of Biology," "Philosophy of Social Science," and "Philosophy of Cognitive Science"—all of equal standing in the ontology. Similarly, there are not just "Philosophical Logic" and "Philosophy of Mathematics" but also "Mathematical Logic." These seem to indicate an egregious explosion of M&E specialties, given that there is not even a distinct category for bioethics, environmental ethics, philosophy of education, existentialism, hermeneutics, Indian philosophy, Buddhist philosophy, Islamic philosophy, African philosophy, or Latin American philosophy, among many other possibilities. All of these latter are either lumped under other categories, such as twentieth-century Continental, or not given any recognition at all.

Since we cannot change the specialty breakdown in the PGR as it stands, let us at least examine whether there is a balance across these already unbalanced categories. That is to say, let's look at the scores that evaluators give to programs within specific areas and see if the proportional area scores match the proportions built into the structure of the PGR. If the scores are skewed toward additional imbalance, this will indicate further evidence of bias in the survey.

Under M&E, the PGR recognizes fifteen specialties. Under value, it recognizes six. Under history, it recognizes nine. Under other, it recognizes three. This breakdown already shows us that M&E (the meat and potatoes of Analytic philosophy) is perceived by the PGR to be as important as value and history combined.[21]

[21] It is worth comparing the PGR's list of philosophical specialties to those put out in a survey from the American Philosophical Association (2013), the largest society of philosophers in the United States. As I've already remarked, the PGR has the following number of specialties in each area: M&E—15, value—6, history—9, other—3. The survey by the APA was sent out by the executive director (Amy Ferrer) following the Eastern Division annual meeting (the largest annual meeting of philosophers in the United States) in order to evaluate the success of the meeting and how welcoming the climate was for underrepresented groups. In the demographic section of the survey, sixty philosophical specialties are listed. Compare

TABLE 4. Ratios of specialties within each area to all specialties, comparing various ways of identifying and quantifying specialties.

	M&E:All	Value:All	History:All	Other:All
1. Specialty ratios of "ideally balanced" program, by PGR specialties	0.45 (15:33)	0.18 (6:33)	0.27 (9:33)	0.09 (3:33)
2. Specialty ratios of "ideally balanced" program, by highest possible PGR scores	0.45 (75:165)	0.18 (30:165)	0.27 (45:165)	0.09 (15:165)
3. Specialty ratios using actual PGR scores, averaged across all programs	0.53	0.19	0.23	0.05
4. Specialty ratios of "ideally balanced" program, by APA specialties	0.18 (11:60)	0.18 (11:60)	0.33 (20:60)	0.30 (18:60)

Row 1 establishes an "ideally balanced" program according to the PGR specialty breakdown (standardized in my Appendix 2). For example, the PGR identifies fifteen specialties placed into the area of M&E, which compares to thirty-three PGR specialties overall, meaning that 45 percent of all PGR specialties are M&E specialties. Row 2 establishes an "ideally balanced" program according to highest possible PGR scores. Row 3 shows the same ratio as in row 2 but using actual PGR scores, averaging the ratio of the sum of each program's PGR specialty scores in each of the four areas to that program's overall score. Row 4 shows the same ratio as in row 1, but using the American Philosophy Association's list of specialties (see my footnote 21)

Under this schema, a program that was both complete and perfectly balanced (with one expert in each specialty) would have a total of thirty-three faculty members: fifteen in M&E, six in value, nine in history, and three in other.[22] The proportion of M&E faculty to all faculty would be 15:33, or 45 percent. For value, the same proportion would be 6:33, or 18 percent; for history it would be 9:33, or 27 percent; and for other it would be 3:33, or 9 percent (see the first row of Table 4).

We can create the very same ratio in a different way that will eventually yield a compelling result. Recall that the highest possible rating score for any specialty for every program is 5. If you take all the specialty scores for a program and add them up, grouping them by area, the highest possible program score in the area of M&E would be 5 (the highest possible individual specialty rating) $\times$ 15 (the number of specialties in M&E) = 75. The highest possible score for value would be $5 \times 6 = 30$; for history it would be $5 \times 9 = 45$; and for other it would be $5 \times 3 = 15$. So, again, we can

this to the PGR's thirty-three and you begin to see indications of exclusivity in the PGR. Using the PGR's own way of grouping specialties into areas, and standardized as described in Appendix 2, the APA's grouping would look like this: M&E—11, value—11, history—20, other—18. The differences are dramatic. No longer is M&E the dominant area; instead, history and other dominate, while M&E and value are equally sized minorities.

[22] Healy (2012b) presents an instructive way to visualize this for the 2006 PGR, categorizing the various specialties and areas into twelve what he calls "specialty areas."

create an area:all ratio. For M&E, the highest possible ratio is 75:165, or 45 percent; for value, the highest possible ratio is 30:165, or 18 percent; for history 45:165, or 27 percent; and for other 15:165, or 9 percent (see the second row of Table 4). The results are the same as in the first row of Table 4. These ratios become instructive when we look at the actual scores of programs.

No program scored the maximum of 165 across all areas. The highest actual score was 91 (New York University). But in which areas did that program score? In an evenly balanced program (by the standards of the PGR), 45 percent of that score would be accounted for by M&E specialties, 18 percent by value specialties, 27 percent by history specialties, and 15 percent by other specialties. The extent to which the actual scores of programs in the PGR deviate from this "ideal balance" is the extent to which the PGR results are biased away from that (already flawed) ideal. If programs on average have higher scores in a particular area, that means that evaluators are recognizing scholars in that area more often than in other areas.

Let's take NYU as an example. The total score, as I mentioned, was 91. The score for M&E specialties was 49; for value it was 20; for history it was 22; and for other it was 0. The ratios, then, are M&E 49:91, or 54 percent; value 22:91, or 22 percent; history 22:45, or 24 percent; and other 0:91, or 0 percent. Thus, 54 percent of NYU's total PGR score is accounted for by M&E specialties; 22 percent is accounted for by value specialties; 24 percent is accounted for by history specialties; and none of it is accounted for by specialties in other. How do these numbers compare to the "ideal balance"? A score of 54 in M&E is 120 percent of the "ideal" M&E ratio of 45; 22 in value is 116 percent of the "ideal" value ratio of 19; 24 in history is a mere 82 percent of the "ideal" history ratio of 27; and 0 in other is of course as far below the "ideal" other score of 9 as any program can get. We see, then, that NYU is scoring disproportionately high in M&E and in value and disproportionately low in history and especially in other.

What if we calculate this same set of ratios for each program in the PGR and create an average over all the programs for each area? Doing that, we get the results in the third row of Table 4, which show us that on average programs ranked in the PGR are scoring disproportionately high in M&E, slightly disproportionately high in value, disproportionately low in history, and very disproportionately low in other.

In other words, M&E, the most independent of all areas (as demonstrated above), is overrepresented by eight percentage points in its total score with respect to an "ideally balanced" score, whereas the comparatively dependent area of history is underrepresented by four percentage points—amounting to a twelve-percentage-point difference with respect to those two areas. Value looks to be slightly well-off, being overrepresented by one percentage point, but this is deceptive because it is a

dependent area, like history, and tied to M&E evaluators. Most salient is that the already scarce and dependent area of other drops 45 percent from the "ideal balance," to account for a mere 5 percent of all specialty scores.

The takeaway from this is that in programs that make the rankings of the PGR, there is an imbalance in program scores tilted toward M&E and away from history, and especially away from other. If evaluators are recognizing history disproportionately less than other fields and if history is more dependent as an area, then it is corroborating evidence of the relative lack of importance of history specialties in the PGR. The same goes even more so for other.

Let's look more closely at the area of other, which is so marginalized in the PGR, by comparing the above numbers to numbers from another source. Under the PGR's "ideally balanced" regime in which there is one rank-5 professor in each of the PGR specialties, specialties in the area of other account for a mere 9 percent of a program's total score. In actuality, for ranked PGR programs, specialties in the area of other account for a paltry 5 percent of a program's score. This means that evaluators are very rarely recognizing expertise in other. Does this mean that in philosophy programs across the country there are few experts working in the specialties of other? Not necessarily. It means that the PGR evaluators are rarely identifying them. For example, suppose that there were no PGR evaluators who had any specialty of other as a field of expertise. Suppose further that there were ten Ph.D. programs consisting of fifteen professors each with extremely high expertise only in specialties that the PGR classifies as other (or doesn't classify at all) and with no expertise in the other three PGR areas. None of these programs would even make the PGR rankings—not because they didn't have philosophical expertise but because the slate of evaluators in the PGR wouldn't be diverse enough to be able to identify them. This extreme hypothetical demonstrates the importance of having a slate of evaluators that is both diverse enough and balanced enough to equitably rate philosophers in all specialties of philosophy. Any imbalance in evaluator specialties with respect to the broader population will necessarily be reflected in the rankings, undercounting the specialties that are not adequately represented.

I point out in footnote 21 that including more specialties than the PGR does in its arbitrary exclusion of specialties could radically change the balance across areas. What if, for example, the PGR's slate of evaluators had a distribution of specialties that more closely matched the sixty specialties put out by the American Philosophical Association (APA)? Instead of the ratios in the first row of Table 4, we would have the ratios in the fourth row. The differences between M&E in the two rows and other in the two rows are large enough to take your breath away. There is no way to say for sure which way of slicing up specialties is most representative of philosophy in the United States, but let's just say that the APA—

the largest body of philosophers in the United State—has it more correct. Granting this, the only area in which the PGR gets the ratio right is value. The PGR far overcounts M&E, significantly undercounts history, and far, far undercounts other.

The only conclusion to be drawn from the analysis of actual results in this section is that in the eyes of the handpicked PGR evaluators, M&E is far and away the dominant area of contemporary philosophy. If a program is not strong in M&E, it has little hope of ranking well in the PGR, and if a program is strong in M&E but weak in other areas, it can still do well.[23] But is that a reflection of the field of philosophy more generally? We'll never know until we open up the process to the whole field instead of arbitrarily excluding 99.5 percent of all qualified evalutators.

Summary

Before moving on to recommendations for how the PGR might be improved, let us recap the flaws detailed so far and consider the effects on the field of philosophy. First, there is a selection bias from the beginning in Leiter's method of selecting evaluators. Leiter uses no acceptable sampling procedure that could lead to generalizable conclusions beyond the opinions of the evaluators of his survey. In other words, one cannot conclude from the results of the PGR that Notre Dame has the program with the eighteenth-highest reputation of all philosophy programs in the United States, as the PGR purports one can. Instead, one can only conclude that it has the eighteenth-highest reputation among the select group of evaluators that Leiter and his handpicked group of advisers have deemed worthy, which make up a mere one-half of 1 percent of all working philosophers, while systematically excluding all others (except those two hundred unnamed philosophers who were invited but did not participate).

[23] This can be seen clearly in Healy's (2012b) visualization for the 2006 PGR mentioned in the previous footnote. Each program is represented by a variable-size pie chart, with each wedge representing a category of philosophy (groupings of the thirty-three PGR specialties). Five wedges represent M&E specialties, five history, and two value. Scanning the programs from top to bottom in the figure, at least four out of the five M&E wedges for the top programs are near the maximum size, until one gets to #10 (not counting numerical ties), Harvard. The most revealing is Australian National University (ANU; the PGR has an international ranking as well as a national ranking), which ranks above Harvard, and has sizable wedges in the five M&E categories, sizable wedges in ethics and political philosophy, and no visible wedges at all in the five history categories—proof that one can do well in the rankings relying on M&E and absent history. Georgetown is nearly a mirror image of ANU, with particular strengths in four of the five history categories, along with ethics and political philosophy, but weak in all five M&E categories. Georgetown winds up much farther down the list—# 57 (again, not counting ties)—evidence that one cannot do well in the rankings without strengths in M&E, and evidence that strengths in history guarantee nothing. Healy comments, "MIT and ANU had the narrowest range, relatively speaking, but their strength was concentrated in the areas that are strongly associated with overall reputation—in particular, Metaphysics, Epistemology, Language, and Philosophy of Mind."

Second, while Leiter seems to suggest that this one-half of 1 percent of philosophers represent the cream of the crop of all philosophers and so are most worthy to undertake such evaluations, in the way he executes his survey, all such experts are mostly working outside their own areas of expertise, and so the rationale of exclusivity, such as it is, crumbles.

Third, the exclusivity is not innocent. There is an unstated assumption (or set of assumptions) driving the selection of evaluators that systematically excludes certain portions of the community of working philosophers. What are those assumptions, and why are they so central to the PGR's methodology?

Fourth, and a possible answer to the question of what the underlying assumptions are, the selection biases are manifested in the results in the form of undercounting the area of history, resulting in lower scores for programs that have strengths in specialties that Leiter categorizes under the area of history.

Finally, and a further answer to the question of assumptions, the selection bias is also manifested in the results in the form of undercounting the area of other, which plays a negligible role in the overall ranking and for this reason provides a negative rationale to any program wishing to hire in any specialty of other, and in any specialty not encompassed by the PGR's list of specialties.

A stark conclusion can be drawn from these five flaws. The PGR is structured to marginalize and/or exclude experts working in specialties that the PGR places under the areas of value, history, and other—82 percent of all specialties according to the APA's accounting. This practice of marginalization and exclusion begins to affect the profession as soon as any university takes the PGR seriously enough to make personnel decisions in order to affect a program's ranking. If any school sets out to raise its philosophy program's ranking in the PGR, it will purposely marginalize specialties in the areas of value and history and outright exclude specialties in the area of other and specialties that do not even make the PGR slate. The more programs do this, the more Ph.D. programs reflect the biases built into the PGR, and as graduates from these programs take jobs at non-Ph.D.-granting colleges and universities, the more the field of philosophy overall begins to resemble the biases implicit in the PGR's methodology. In this way, the PGR becomes a self-fulfilling prophecy, projecting its own biases about the right way to do philosophy onto the rest of the field, thereby molding the field in its own image.

What about an answer to the second question above: Why are the assumptions that are built into to the PGR's selection process of evaluators so central to the PGR's methodology? When we notice that the APA specialties that the PGR would, or does, categorize under other are often associated with feminism and non-Western ethnicities and cultures, one cannot help but wonder whether the PGR's hidden biases are based in sexism, racism, ethnocentrism, and xenophobia.

It is often recommended by defenders of the PGR that the PGR rankings be taken with a grain of salt, but because of its status in the profession, which actually does take account of this flawed instrument in making such important decisions as hiring, the PGR is having an unwarranted and negative effect on the profession. The harm can be seen most saliently in the way that non-Western philosophy is treated. Despite the growth in multiculturalism across all levels of education and despite calls for diversity and globalization in all corners of academia (Bruya 2015), any philosophy Ph.D. program that considers hiring in any branch of non-Western philosophy, and that strives to achieve or maintain a high rank in the PGR, need only look at the above biases in the PGR to be convinced that it would be an infinitely bad idea to make such a hire. That post could be used instead to hire in an area that would have an impact on a program's rank. For instance, even if a program hired the most distinguished scholar working in Indian philosophy, the likelihood of the general slate of PGR evaluators recognizing this person's name for the overall ranking would be little to none. Thus, this person, prominent in her or his own field, would do nothing to raise the program's rankings overall and would instead waste a slot that could be filled by someone who could raise the program's rank. This is why the PGR is having a deleterious impact on the profession through its deep-seated methods of exclusivity and why it is worth being examined in detail in this article.

How the PGR Can Be Improved

I pointed out at the start of the article that when the PGR was first introduced, it was a welcome resource. It has achieved its status in the profession for its perceived utility, a perception that continues today. Although we have seen how the biases inherent in its methodology compromise its actual utility to devastating effect, it is not unsalvageable as a useful instrument, and can gain genuine validity with targeted structural changes. Below I outline how the PGR can be changed so that it can gain a legitimate status as an instrument that provides genuinely useful information.

1. Use a Random Sample for the Evaluator Pool

The core problem of the PGR is the biased sample of respondents. There are a number of ways that a random sample could be achieved. Since the poll is already conducted online, the obvious way would be to simply open it up to all faculty employed in philosophy programs or, better, all those with a Ph.D. in philosophy.[24] Expanding the pool would have the added

[24] Would the PGR, then, simply turn into a popularity contest? Because of the way the survey is conducted, it is already highly reliant on name recognition, which is what I take

benefit, beyond engendering basic validity, of getting the opinions of those who will actually do the hiring of prospective graduate students when the time comes for them to apply for jobs. Most graduates of Ph.D. programs do not get hired into Ph.D. programs, let alone into the even smaller set of PGR-ranked Ph.D. programs, and so restricting the pool of evaluators largely to those in PGR-ranked Ph.D. programs artificially reduces the PGR's utility to prospective graduate students who will eventually be on the job market and need to know the reputations of their prospective program among those schools that are doing the hiring, the vast majority of which have no current role in the PGR.[25]

2. Use Mathematically Aggregated Specialty Scores to Calculate Overall Ranking

We saw above how the overall ranking is not an expert ranking at all and is mathematically divorced from the specialty rankings, which really are expert rankings (or, more properly, self-rankings). There are any number of ways that the specialty rankings could be aggregated, but the obvious and simplest way would be to simply sum them for each program.

A related improvement would be to rank individual professors, rather than entire programs. If a department has three philosophers working in epistemology who all deserve top scores in the specialty of epistemology, that program's score in epistemology should be three times higher than the score of a program that has only one epistemologist who deserves a top score. Granted, this would give a numerical advantage to larger programs, but this is a genuine advantage of larger programs, not an artifact of the poll's methodology.[26] It seems obvious that a set of students can learn much more from three specialists in a field than they can from just one, all else being equal. There are two added benefits to this method. The first is that it would encourage growth in programs without discouraging comprehensiveness. Second, if a program could approach its university administration with a data-driven argument that growing the program would improve the program (in substance and in its ranking), that would be to the good.[27]

"popularity" in large part to mean. I don't see any reason that expanding the evaluator pool would make it more or less about name recognition; expansion could make it less so if evaluators were ranking in their specialties instead of across whole programs, per my second recommendation.

[25] Recall from section A that there are 841 undergraduate philosophy programs in the United States to 147 graduate programs, and only fifty of the graduate programs make the PGR overall ranking.

[26] It is not so clear, however, that two specialists who rate a score of 3 are better than one specialist who rates a score of 5. This problem could be remedied by adjusting the evaluation scale. On the other hand, given that professors have a tendency to go on leave, having one 3 and one absent 3 would be better for a student than having an absent 5.

[27] Jennifer Saul (2012, 269–70) also recommends that the PGR drop overall rankings and use an algorithm to establish rankings from specialty scores.

3. Allow Evaluators to Evaluate Only One Specialty

Allowing evaluators to evaluate only one specialty would reduce the possibility of the biases of one area or specialty diluting another area or specialty, as happens under the current method.

Instituting this step and steps 1 and 2, involving three simple and straightforward changes to the PGR methodology, would be an outstanding first step in bringing it a warranted legitimacy. There are two further steps that would make it a genuinely representative and pluralist instrument relevant to all philosophy Ph.D. programs.

4. Revise the List of Specialties

Before executing a randomly sampled poll, the PGR should institute a randomly sampled questionnaire asking for the main field in which a philosopher works. Again, only one selection should be allowed. No prompts should be offered, except perhaps that the field should be something like a hiring area of specialization (since one of the key purposes of the PGR is to provide guidance to future graduate students, ideally eventuating in employment). The results should be systematically analyzed by a representative committee for the purpose of creating a set of specialties, each of which is narrow enough to represent specialized research and broad enough to constitute a populated field of research, but which ideally does not necessarily entail inclusion in another specialty (as philosophy of biology in general philosophy of science).[28] For fields that are currently demographically rare, such as Japanese philosophy and philosophy of education, a small population may be acceptable in order to allow the field to grow. This work should be repeated regularly, perhaps once every five to ten years in order to keep the list current.

5. Offer a Special Score to Indicate Comprehensive Balance Within Programs

One wouldn't want to effectively force programs to pluralize by incorporating a comprehensive balance score directly into the PGR's equation for calculating overall scores. It would, however, be desirable to encourage comprehensive balance within programs by providing such a score in addition to the regular scores. A *comprehensive balance score* could be arrived at in a variety of mathematical ways that would indicate how many specialties are represented with respect to all specialties in a program, or how many areas are represented with respect to all areas in a program, or how deep a program is in specific specialties in relation to how broad it is overall.

[28] Creating such demarcations would be controversial, and that's why it would require a diverse committee.

❖

In my offering the above advice, there is one possibility that has been overlooked. It may be that Brian Leiter prefers to still view the PGR as it was originally intended—exclusively for the use of graduate students (particularly those interested in the Analytic style of doing philosophy)—and that any other uses are illegitimate and/or not his problem. If that is the case, then he should remove the thin veneer of scientific validity (for example, "Methods & Criteria") in his report and post a disclaimer in multiple prominent places in the PGR, stating something like the following:

> This report is an ad hoc instrument created solely for the use of prospective graduate students in evaluating Analytically oriented philosophy Ph.D. programs and is of limited utility in that it uses a nonrepresentative sampling procedure that selects evaluators according to vague and unannounced criteria, marginalizing some disciplines of philosophy while excluding others, and in that it lacks important measures, such as a program's graduation rate, its placement record, and the amount of financial resources offered to students. Any institutional use beyond the stated purpose, such as in hiring or in demonstrating program excellence, is illegitimate, ill-advised, beyond both the stated and the actual scope of the report's demonstrable results, and potentially damaging to the profession of academic philosophy.

Department of History and Philosophy
Pray-Harrold 701
Eastern Michigan University
Ypsilanti, MI 48197
USA
bbruya@emich.edu

Appendix 1: A Note on the "Methods & Criteria" Section of the PGR

The "Methods & Criteria" section of the PGR contains very little in the way of either methods or criteria. It is composed of three parts, "Description of the Report," "Faculty Lists Used by Evaluators," and "'Analytic' and 'Continental' Philosophy." The "Faculty Lists" section is a blank page (but would presumably contain the faculty lists available for download on the "Overall Rankings" page [Leiter 2011c]), and the third section has nothing to do with methods or criteria. Here is a summary of the main methods described in the "Description" (Leiter 2011b):

- Online survey given to five hundred evaluators. Just over three hundred responded. (No demographic information or analysis is offered.)
- Faculty lists are provided to evaluators from eighty-eight programs, minus the name of each school.

- Ranked programs from previous years plus a few more programs make up the faculty lists. (No mention of the criteria for including the extra programs.)
- Verbatim instructions given to evaluators.

Here is a summary of the criteria listed for selecting evaluators (Leiter 2011b):

- Selected "with an eye to balance in terms of area, age and educational background." (No specifications are given, and there is a caveat that "since, in all cases, the opinions of research-active faculty were sought, there was, necessarily, a large number of alumni of the top programs represented," thus flatly contradicting any sort of balance with regard to area or educational background and presupposing which programs are "top programs." No parameters for, or definition of, "research-active" are offered. Gregory Wheeler [2012b] demonstrates the educational imbalance in the PGR and how it correlates with the rankings.)
- "Approximately half of those surveyed were from previous years; the other half were nominated by members of the Advisory Board, who picked research-active faculty in their fields." (There is no mention of whether all nominations were accepted, of how the members of the Advisory Board were selected, of demographic information about the advisory board, or of how prior-year evaluators were selected.)

A list of all evaluators is then provided, including the current institution and Ph.D.-granting institution of each. Almost all are based at doctoral institutions. The obvious assumption is that only professors at Ph.D.-granting institutions are "research-active." This assumption is patently false. One can easily see why when considering that the tenure and promotion criteria at all colleges and universities require active research on the part of all professors. But we don't have to rely on this observation alone. James S. Fairweather did a study that included a measure of the level of productivity of faculty across institutional types and found that 49 percent of faculty at research universities are "highly productive" in research, 47 percent at doctoral universities, 42 percent at comprehensive universities, and 35 percent at liberal arts institutions (Fairweather 2002, 40, Table 3). So if the PGR is looking for "research-active" faculty, it has counterproductively narrowed the pool by a large margin—approximately 80 percent of all "highly productive" philosophers are systematically excluded. Using the percentages above and the numbers in footnote 5, there are approximately 847 (1,764 × .48, splitting the difference between research and doctoral schools) highly productive faculty in graduate programs and 1,639 (4,205 × .39, splitting the difference between

comprehensive universities and liberal arts schools) in undergraduate programs. With only five hundred of these twenty-five hundred tapped, 80 percent are excluded.

The fundamental flaw that undermines every claim of the PGR is that it uses a qualitative research sampling method from which it infers general conclusions more appropriate to quantitative research methods. Qualitative research methods, such as chain-referral sampling that Leiter purports to employ, are used in the social sciences in order to study the opinions, motivations, and behaviors of certain small populations, and because the samples are nonrepresentative, no conclusions can be inferred about larger populations. In the way that Leiter does his sampling, he is effectively studying the evaluators rather than the actual philosophy programs. In qualitative research, one uses nonprobabilistic sampling techniques, such as chain-referral sampling, in order to study the sample. In quantitative research, one uses random sampling techniques and statistical analysis to draw general conclusions beyond the sample. The best that Leiter can do from his sampling is to say that the three hundred or so philosophy professors linked together by a network of mutual high regard think X about the philosophy programs in question, which might be interesting in terms of understanding more about the folks in the sample, but it does not constitute information about the philosophy programs themselves. In order to provide genuine information about the philosophy programs, the PGR needs to use a statistically relevant sampling technique, which it purports to use ("with an eye to balance in terms of area, age and educational background") but clearly does not. With a statistically relevant sampling technique, Leiter could say, "These fifty programs are ranked in the way described by working philosophers in the United States." Right now, he can only say, "There is a small group of working philosophers, constituting 0.5% of all working philosophers in the U.S., whom I hold in high regard, who generally hold each other in high regard, and who rank these programs in the way described." This is hardly a sound basis for readers of the report to build philosophy programs, let alone to build a whole profession.

Appendix 2: A Note on Philosophical Areas and Specialties Used in the PGR

An *area* in the PGR is a general category under which various specialties are grouped. In his "Description of the Report" (2011b), Leiter allows seven distinct areas for evaluators: "Metaphysics and Epistemology," "Science," "History," "Value," "Logic," "Chinese Philosophy," and "Other." In his "Breakdown of Programs by Specialties," Leiter (2011e) lists five distinct areas for programs: "Metaphysics and Epistemology," "Philosophy of the Sciences and Mathematics," "Theory of Value," "History of Philosophy" and "Other." For the purpose of statistical

analysis of evaluator area and programs, I have standardized the areas, merging all into the following four PGR areas: metaphysics and epistemology (now including specialties in philosophy of science, mathematics, and logic, which, if not M&E specifically, are methodologically and topically closely allied), value, history, and other (including Chinese philosophy). One could argue that specialties in philosophy of science, mathematics, and logic should not fall under M&E. There is no reason to think, however, that logic, for example, should necessarily be grouped with general philosophy of science into a separate area. It is uncontroversial that many of the specialties of M&E, philosophy of science, philosophy of mathematics, and logic are core specialties of Analytic philosophy. Breaking them out into several more separate groups (as, for example, Kieran Healy [2012b] does) would not alter the conclusions of the arguments made in this critique.

Is working from mathematically aggregated specialty scores, as I do in section F.1, the best way of ranking the programs? In defense of the PGR, one may say that because of the fluidity of specialties in the profession, it is best to not rely on specialty scores and instead simply give an overall ranking of each program, independent of specialty categorizations. Very likely, this is Leiter's rationale for retaining the overall ranking. Keep in mind that the way Leiter does his specialty rankings is to allow one ranking from each evaluator for each specialty for every program. If one were to create a mathematical aggregate from these scores for their umbrella areas, as I recommend in section F.1, the number of specialties would take on a special significance. The more specialties a program has in an area, the higher the program's score in that general area could be. For instance, suppose we collapsed all M&E specialties into one and diversified history specialties into thirty more fine-grained specialties. The highest a program could score in M&E would be 5 (1 specialty $\times$ 5 [the highest possible rating]), whereas the aggregate history score could conceivably be as high as 150 (30 specialties $\times$ 5). Therefore, creating an overall ranking based on mathematical aggregation is highly dependent on how the profession is divided into specialties and areas. A solution would be to forgo such divisions altogether and give one ranking across entire programs, as Leiter does. As I demonstrate in section D, however, ranking across entire programs has its own, more intractable, problem: evaluators are no longer ranking in their areas of expertise, which means that the PGR is a measure of how many evaluators there are in any one area or specialty (for example, M&E specialists are more likely to recognize the quality of M&E specialists in a program, and so the more M&E specialists there are in the evaluator pool, the higher M&E-heavy programs will rank). A better solution, in addition to creating a representative sample of evaluators, would be to rate individual professors instead of entire programs (a program with ten philosophers all rated 5, and no other rated philosophers, would have a score of 50). This would dilute (but not

entirely resolve) the problem with the number of specialties. The problem could be further ameliorated by honing the number of specialties such that expertise in one would not entail expertise in another. For instance, the PGR has both philosophy of biology and general philosophy of science, and expertise in the former generally entails a certain level of expertise in the latter, resulting in double counting (by the aggregate method).

Acknowledgments

I owe a large debt of gratitude to the following for help analyzing various parts of the data here and for help putting it into precise language: Hamony Lu, Tai Chang, Jeanne Nakamura, and Andrew Abbott (any residual errors are my own). Thanks also to John Koolage for allowing me to bounce arguments off him. This article was written in part while on a teaching Fulbright at National Taiwan University. Thanks to Eastern Michigan University for giving me leave of absence for the Fulbright, to the U.S. Fulbright Program and to Fulbright Taiwan for funding the opportunity, and to the Department of Philosophy at NTU for welcoming me and providing office space and other assistance.

References

American Philosophical Association. 2013. "APA Eastern Division Meeting Evaluation and Climate Survey." Survey circulated by e-mail on January 6.

Atkinson, Rowland, and John Flint. 2001. "Accessing Hidden and Hard-to-Reach Populations: Snowball Research Strategies." *Social Research Update* 33 (Summer): 1–4.

Biernacki, Patrick, and Dan Waldorf. 1981. "Snowball Sampling: Problems and Techniques of Chain Referral Sampling." *Sociological Methods & Research* 10, no. 2 (November): 141–63.

Bruya, Brian. 2015. "The Tacit Rejection of Multiculturalism in American Philosophy Ph.D. Programs: The Case of Chinese Philosophy." *Dao* 14, no. 3 (September): 369–89.

Coleman, James S. 1958. "Relational Analysis: The Study of Social Organizations with Survey Methods." *Human Organization* 17, no. 4 (Winter): 28–36.

Dotson, Kristie. 2012. "How Is This Paper Philosophy?" *Comparative Philosophy* 3, no. 1:3–29.

Drăgan, Irina-Maria, and Alexandru Isaic-Maniu. 2012. "Snowball Sampling Developments Used in Marketing Research." *International Journal of Arts and Commerce* 1, no. 6 (November): 214–23.

Erickson, Bonnie H. 1979. "Some Problems of Inference from Chain Data." In *Sociological Methodology*, volume 10, edited by Karl F. Schuessler, 276–302. San Francisco: Jossey-Bass.

Ernst, Zachary. 2009. "Our Naked Emperor: The Philosophical Gourmet Report." http://www.dropbox.com/s/qd9gdl7ozofhit0/emperor-1.pdf?dl=0APS_DEA_Science_Final.pdf. Accessed October 1, 2015.

Fairweather, James S. 2002. "The Mythologies of Faculty Productivity: Implications for Institutional Policy and Decision Making." *Journal of Higher Education* 73, no. 1 (January/February): 26–48.

Frodeman, Robert, and Jennifer Rowland. 2009. "De-Disciplining the Humanities." *Alif: Journal of Comparative Poetics* 29, special issue entitled "The University and Its Discontents: Egyptian and Global Perspectives," 62–72.

Healy, Kieran. 2012a. "Ratings and Specialties." *Kieran Healy* (blog). March 21. http://kieranhealy.org/blog/archives/2012/03/21/rating-and-specialties. Accessed February 27, 2014.

———. 2012b. "A Not Quite Satisfactory Way of Looking at Departments and Specialties." *Kieran Healy* (blog). March 22. http://kieranhealy.org/blog/archives/2012/03/22/a-not-quite-satisfactory-way-of-looking-at-departments-and-specialties. Accessed February 27, 2014.

Heck, Richard. 2014. "About the Philosophical Gourmet Report." *Richard Heck: Philosophy, Linux, Feminism, etc.* (blog). January 27. http://rgheck.frege.org/philosophy/aboutpgr.php. Accessed February 27, 2014.

Leiter, Brian. 2011a. "'Analytic' and 'Continental' Philosophy." *Philosophical Gourmet Report*. http://www.philosophicalgourmet.com//2011/analytic.asp. Accessed October 1, 2015.

———. 2011b. "Description of the Report." *Philosophical Gourmet Report*. http://www.philosophicalgourmet.com/2011/reportdesc.asp. Accessed October 1, 2015.

———. 2011c. "Overall Rankings." *Philosophical Gourmet Report*. http://www.philosophicalgourmet.com/2011/overall.asp. Accessed October 1, 2015.

———. 2011d. "What the Rankings Mean." *Philosophical Gourmet Report*. http://www.philosophicalgourmet.com/2011/meaningof.asp. Accessed October 1, 2015.

———. 2011e. "Breakdown of Programs by Specialties." *Philosophical Gourmet Report*. http://www.philosophicalgourmet.com/2011/breakdown.asp. Accessed October 1, 2015.

———. 2012. "The Five Most Common Objections to the PGR." *Leiter Reports: A Philosophy Blog*. February 1. http://leiterreports.typepad.com/blog/2012/02/the-five-most-common-objections-to-the-pgr.html. Accessed February 27, 2014.

Leiter, Brian, and Michael Rosen, eds. 2007. *Oxford Handbook of Continental Philosophy*. New York: Oxford University Press.

McAfee, Noëlle. 2007. "Philosophy Rankings." *GonePublic: Philosophy, Politics, & Public Life* (blog). November 24. http://gonepublic.net/2007/

11/24/philosophy-rankings/?relatedposts_exclude=443. Accessed February 27, 2014.

———. 2010a. "Ranking Continental Philosophy Programs." *GonePublic: Philosophy, Politics, & Public Life* (blog). October 21. http://gonepublic.net/2010/10/21/ranking-continental-philosophy-programs. Accessed February 27, 2014.

———. 2010b. "Shadow of a Phantom or How to Do a Survey." *GonePublic: Philosophy, Politics, & Public Life* (blog). November 8. http://gonepublic.net/2010/11/08/the-shadow-of-a-phantom-or-how-to-do-a-survey. Accessed February 27, 2014.

———. 2011. "The Favorites' Favorites: Another Round of PGR Rankings of Continental Philosophy." *GonePublic: Philosophy, Politics, & Public Life* (blog). November 15. http://gonepublic.net/2011/11/15/the-favorites-favorites-another-round-of-pgr-rankings-of-continental-philosophy. Accessed February 27, 2014.

———. 2014. "Is the PGR Sexist?" *GonePublic: Philosophy, Politics, & Public Life* (blog). February 12. http://gonepublic.net/2014/02/12/is-the-pgr-sexist. Accessed February 27, 2014.

Protevi, John. 2011. "A Brief Look at 2011 PGR 20th-Century Continental Philosophy Evaluators." *New APPS: Art, Politics, Philosophy, Science* (blog). http://www.newappsblog.com/2011/12/2011-pgr-20th-c-cp-board.html. Accessed February 27, 2014.

Romaniuk, Bohdan, ed. 2012. *The College Blue Book*, vol. 3: *Degrees Offered by College and Subject*. Detroit: Macmillan Reference.

Rosen, Michael. 1998. "Continental Philosophy from Hegel." In *Philosophy 2: Further Through the Subject*, edited by A. C. Grayling, 663–704. New York: Oxford University Press.

Saul, Jennifer. 2012. "Ranking Exercises in Philosophy and Implicit Bias." *Journal of Social Philosophy* 43, no. 3 (September): 256–73.

Tongco, Maria Dolores C. 2007. "Purposive Sampling as a Tool for Informant Selection." *Ethnobotany Research and Applications* 5:147–58.

Walker, Margaret Urban. 2004. "Waiter, There's a Fly in My Soup! Reflections on the Philosophical Gourmet Report." *Hypatia* 19, no. 3 (Summer): 235–9.

Wheeler, Gregory. 2012a. "Manufacturing Assent: The Philosophical Gourmet Report's Sampling Problem." *Choice & Inference* (blog). April 17. http://choiceandinference.com/2012/04/17/manufactured-assent-the-philosophical-gourmet-reports-sampling-problem. Accessed February 27, 2014.

———. 2012b. "More on the Educational Imbalance Within the PGR Evaluator Pool." *Choice & Inference* (blog). April 19. http://choiceandinference.com/2012/04/19/more-on-the-educational-imbalance-within-the-pgr-evaluator-pool. Accessed February 27, 2014.

———. 2012c. "Two Reasons for Abolishing the PGR." *Choice & Inference* (blog). April 24. http://choiceandinference.com/2012/04/24/two-reasons-for-abolishing-the-pgr. Accessed February 27, 2014.
Wilshire, Bruce. 2002. *Fashionable Nihilism: A Critique of Analytic Philosophy*. Albany: State University of New York Press.
Wilson, Robin. 2005. "Deep Thought, Quantified." *Chronicle of Higher Education*. Faculty. May 20.