Comment on Bruya

David Wallace, 15/12/2015 (v3 20/12/2015)

1. Introduction

Brian Bruya's recent article, "Appearance and Reality in The Philosophical Gourmet Report: Why the Discrepancy Matters to the Profession of Philosophy" (Metaphilosophy 46 (2015) pp.657-690) presents a range of criticisms and proposals for reform for the Philosophical Gourmet Report (PGR).

I don't find the qualitative parts of the criticisms persuasive (for the most part they seem to assume a straw man – that the PGR is intended as an opinion poll rather than an expert review – and then draw conclusions from that). But that's outside the scope of this comment. For my purposes the salient point is that Bruya describes his article as "a data-driven critique". This is problematic since the use of data in the article is severely methodologically and statistically flawed in several places. In this note I briefly (and non-exhaustively) sketch what I think are the most severe quantitative problems with the analysis in Bruya 2015. (They appear in no particular order.)

2. Hidden reallocation of subject speciality areas to groups

PGR 2011 categorises speciality areas into 5 categories: M&E, Value, History, Science (i.e., philosophy of the sciences and mathematics), Other.[1] Bruya combines the Science and M&E categories together and calls them M&E. All of his category based analysis – notably (i) his "egregious explosion of M&E specialties" (p.674) and his "area dilution" argument that M&E is more independent than other categories, and that evaluators in other categories disproportionately co-evaluate in M&E (pp.668-9) – rely on this recategorisation.

There may be an academic case for the reclassification. (As a fairly metaphysics-sceptical philosopher of science I don't find that case convincing, but that's a matter of legitimate disagreement.) However, if so, it really ought to have been front and centre in the paper's methodology. Instead, it appears nowhere in the main paper, which is *entirely silent* on the reclassification. It is noted only in Appendix 2 (p.686). Here Bruya actually acknowledges that some areas in the Science category are "not M&E specifically, but methodologically and topically closely allied". Again, I don't agree with this, but again, the point could be debated; the real problem is that the main paper proceeds from the premise that these areas *are M&E*, not just that they're somehow related to it.

Most concerningly, Bruya's appendix comment goes on as follows:

"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 … would not alter the conclusions of the arguments made in this critique."

But in fact several of Bruya's arguments rely centrally on this reclassification. Firstly, Bruya's "egregious explosion" critique of M&E is based on the fact that there are 15 speciality areas in M&E, as against 6 in Value, 9 in History, and 3 in Other. Once the PGR's original classification is restored, there are only 7 areas in M&E, 8 in Science, 6 in Value, 9 in History, and 3 in Other.

[1] This categorisation appears very consistently throughout the 2011 PGR; the only deviation from it is that of the 303 listed assessors, 2 of them are specifically listed as assessing "Logic" and one of them as assessing "Chinese Philosophy". The other 300 listed assessors fall into one of the 5 categories, as do all of the subjects listed in the "breakdown of speciality ratings", where they are identified as "reflecting conventional demarcations".

Secondly, Bruya's "dependence ratios" (table 1) purport to show that evaluation of non-M&E areas is dependent on M&E, in that a high fraction of assessors in other areas co-assess in M&E. This effect is drastically reduced when the PGR's classification is restored:

Table 1: recalculated degrees of dependence between areas (cf Bruya table 1)					
Field	M&E	Value	History	Science	Other
Degree of dependence on M&E		0.20	0.31	0.58	0.19
Degree of dependence on Value	0.12		0.2	0.11	0.63
Degree of dependence on History	0.20	0.21		0.18	0.19
Degree of dependence on Science	0.24	0.074	0.12		0.25
Degree of dependence on Other	0.019	0.11	0.03	0.061	
Overall degree of dependence	0.51	0.51	0.55	0.72	0.93

The overall levels of dependence for M&E, Value, and History are now about the same, all around 50%. The only subject-specific level of dependences above 50% are now the dependence of Science on M&E and the dependence of Other on Value.

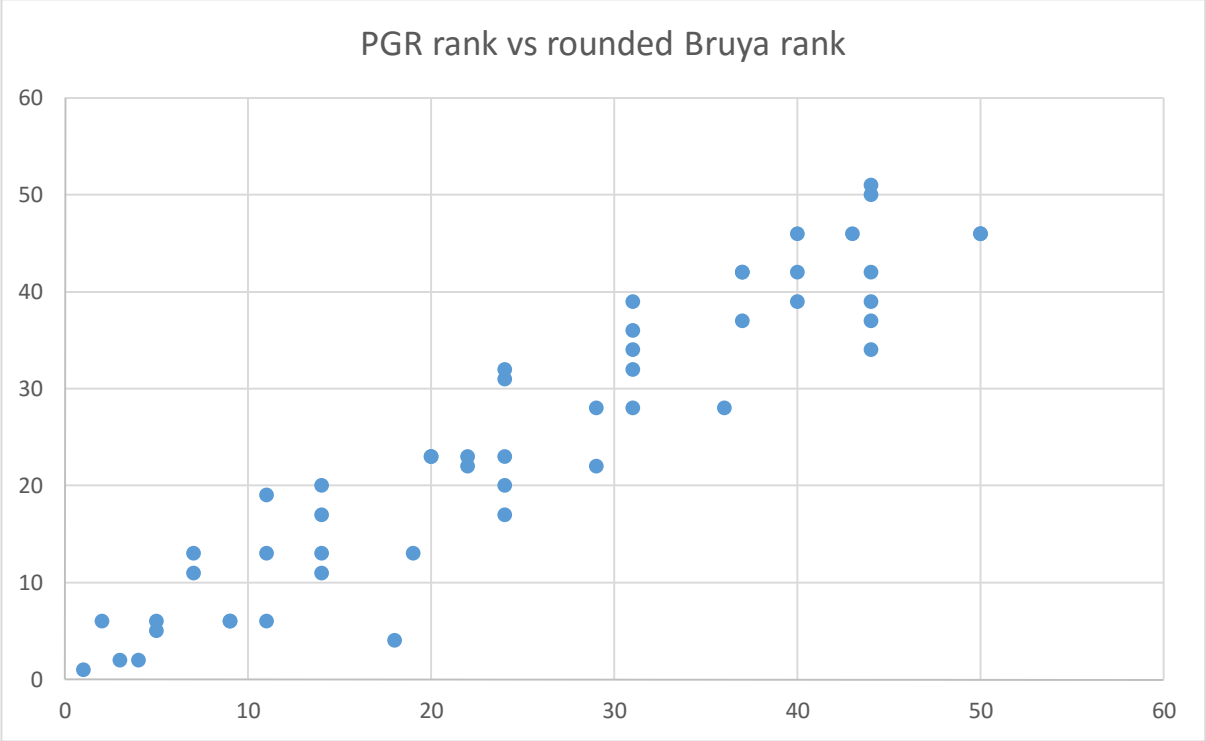
To repeat: it's not *crazy* to imagine a critique of the PGR based on the fact that it gives much too much attention to philosophy of science, and/or that philosophy of science is disguised M&E. But to relegate all this to an appendix and to bake that critique silently into the main analysis is very troubling.

3. "Aggregated rank" for departments actually tracks PGR rank closely

Bruya calculates an "aggregated rank" for each department by simply summing their ranks in the speciality ratings. The method of aggregation is simply to sum the marks in each speciality area. This is problematic for three reasons:

1) it assumes that the PGR categories ought to be seen as of equal weight, something the PGR itself never claims. (This is a recurring issue in Bruya 2015.) Philosophy of physics, for instance, is a niche subject in almost all institutions, and a fairly high rank can be achieved by just having a couple of well-regarded people in the field; regarding it as of equal weight to (say) Ethics is hard to motivate. (This seems to be driving Notre Dame's very high rank on Bruya's analysis).
2) It confines the strength of an area to 5/165 of the total rank. The only way to get a high score on Bruya's analysis is to be strong across a very large number of disciplines; exceptional strength in a smaller number of large subject areas contributes little.
3) It assumes that the PGR marks are aggregative, so that having three speciality areas ranked 2.0 or two ranked 3.0 is better than having one speciality area ranked 5.0. On the PGR scale 5.0 means "distinguished" while 2.0 means "adequate" and 3.0 "good", all of which are qualitative – and the scale is capped at 5 even for large and super-distinguished research groups, so this looks prima facie implausible. (This, and the previous point, seems to be driving Rutgers' comparatively low rank on Bruya's analysis.)

But for all that, the striking fact about Bruya's alternative ranking is that it tracks the PGR rank very closely. Rounding Bruya's ranking to the same accuracy as the PGR (which only distinguishes c. 30 possible scores, from 2.0 to 4.8) we get:



The correlation between PGR score and Bruya score is 94%; the standard error (average difference) is 4 ranks. Bruya describes this as "quite large" but it's pretty trivial given the spread of ranks of the PGR overall. That would seem to render moot most of Bruya's complaints about the aggregate rank, *even if* Bruya rank was methodologically sound. In other words, the effects of supposed assessor bias really aren't doing much to the overall rank.

Though to repeat: Bruya rank *isn't* methodologically sound. A possible reply is that PGR rank isn't methodologically sound either. That isn't a legitimate response. "My interlocutor's method is unsound, so I will use an unsound method too" isn't okay. (In addition, the question of PGR soundness is a matter of judgement and argument; the problems with Bruya rank are straightforwardly mathematical.)

4. Concern about distribution of speciality areas

Bruya uses the APA's division of specialities as a target level for the spread of PGR evaluators. This seems to mischaracterise the rationale for categories both in the APA and in the PGR. In each case (one assumes) the reason is because that division is useful for whatever purpose is being served by the categoriser, and for neither the APA nor the PGR is the purpose to divide the philosophy demographic evenly.

Philosophy of physics, for instance, is clearly smaller than ethics, but the PGR treats each as a special case because for prospective graduate students (the clearly-identified target of the speciality ratings), if they want to specialise in philosophy of physics they have fairly bespoke faculty needs. That carries no implication that philosophy of physics is comparable in size to ethics!

Bruya writes that "[t]here is no way to say for sure which way of slicing up specialities 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 [sic] – has it more correct."

The APA division strikes me as so unreliable a guide to the demographics of the profession (does Bruya *really* think as many people do philosophy of biology as do Ethics? That only one philosopher in 60 has

a Metaphysics primary AOS?) that if it was the best we could do, we'd just have to accept that we can't know the demographics of the profession. But in fact, we can do better. Passing over the fact that the PGR is not confined to the US (I write from Oxford!) I note that rough data on the spread of philosophers is fairly readily available, so there is no need to fall back on the flawed method of categories. For instance, the Philosophy Documentation Center collects data from university websites and departments, and from individuals, and on that basis compiles a list of how many philosophers are in each AOS. This is clearly a very crude indicator, as people can and often do have more than one AOS, but it at least bears some methodological relation to actual demographics.

On that basis (excluding the catch-all category of "modern philosophy") we get the following:

Category	% of AOS in PDC data	% of PGR categories	% of APA categories
M&E	27%	21%	Est. 9%[2]
Value	30%	18%	18%
Science	9%	24%	Est. 9%
History	27%	27%	33%
Other	8%	9%	30%

As it happens, the PGR categories track the demographics moderately well, much better than the APA categories (not that either is *intended* to track the demographics, and not that the fine details of the PDC-derived demographics should be taken too seriously).

5. Concerns about differing mean in speciality areas

Bruya's Table 4 (p.675) and associated discussion compares the "speciality ratios using actual PGR scores" to the "speciality ratings by highest possible scores", and adduces bias from them ("a compelling result" – p.675). It's a little difficult to interpret what is going on mathematically here, but I think what it comes down to is that the mean score in M&E (including Science) is higher than in other areas. Recovering mean scores (as ratios of the average) by reverse-engineering Bruya's data, we get:

Category	**Mean speciality score compared to mean across all subjects**
M&E (including science)	+17%
Value	+ 5%
History	-15%
Other	-45%

Bruya claims, without further argument, that "[i]f 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." This doesn't follow from the data. Possibilities include (I list them in what seems subjectively to be an increasing order of likelihood):

1) Bruya's suggestion: bias in favour of M&E/science
2) Work in M&E/science, on average, is somewhat better than in other areas in contemporary philosophy. (It is not an *a priori* truth that at each instant in time, each area in philosophy is doing equally well.)

[2] Bruya combines the Science and M&E categories for the APA and I don't have access to the raw data

3) The PGR only names departments that score at least 3+ on the speciality ranking. If there are a large number of reasonably good Ethics departments (scoring 3-4) but comparatively fewer in Philosophy of Language, philosophy of language could obtain a *higher* average speciality score as a consequence, but that spread doesn't obviously mean that Ethics is doing worse (there may well be a number of departments rated 2 or 1 in philosophy of language that didn't show up; equally there might just be more Ethics concentrations period). This distribution within subject areas could readily have been analysed; not doing so seems a serious methodological lacuna in the critique.
4) Different groups of assessors are interpreting the criteria somewhat differently, but harmlessly so. There is no prima facie reason to expect different groups of assessors to be systematically scaling in the same way – and it doesn't really matter if they are, given that the point of the speciality rankings is to rank-order *within a speciality*. Only if the PGR used an aggregative formula to work out overall rankings would it necessary to assume comparability of numerical rankings. (For that reason, an aggregative approach probably should aggregate ranks, not scores.)

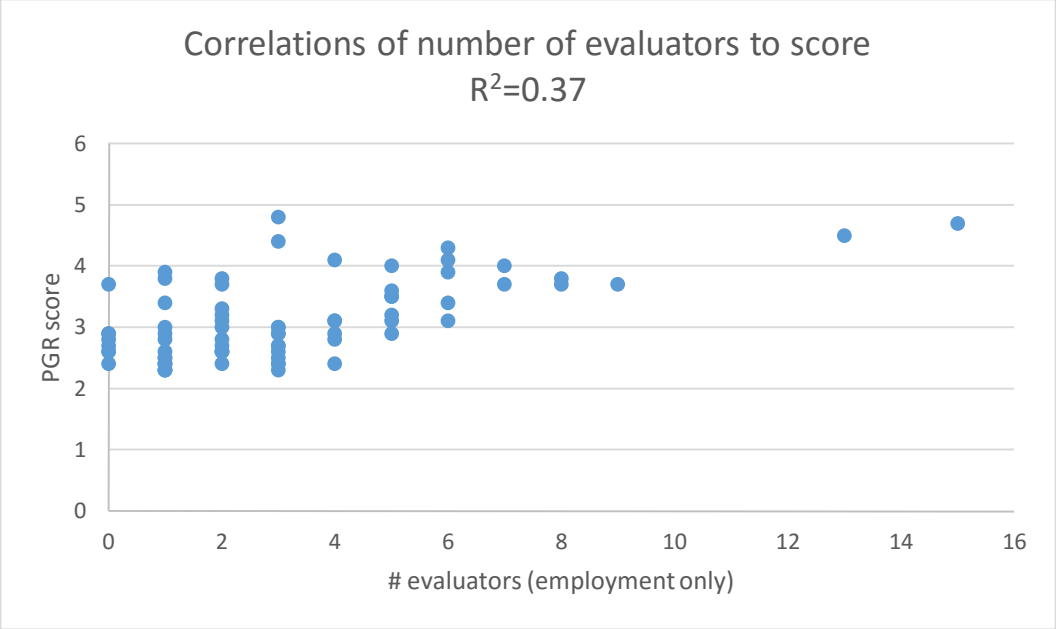
6. Discussion of skewed distribution of evaluators conflates two meanings of "evaluator's institution"

On pages 662-664, Bruya criticises the PGR's selection of assessors by comparing the number of assessors "hailing from" a given institution with the PGR score obtained by that institution. He claims that (i) there is a strong and worrying correlation between number of assessors "hailing from" a given institution and the rank that institution obtains in the PGR; (ii) in particular, about half of all PGR evaluators come from "eight programs in a tight geographical area", which obtain most of the highest PGR scores, and that as such "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".

Let me begin by noting one oddity of Bruya's analysis here. Without comment, he excludes from the survey all features of the PGR concerning the world outside the USA. Since 87 of the PGR's 303 evaluators had current affiliations at non-US institutions, this discards a non-trivial amount of data.

More significantly, Bruya speaks of evaluators being "from" a given school, or "hailing from" a school, or of "the school with" a given number of evaluators. At least to this reader, the clear impression is that what is meant is the *current affiliation* of the evaluator. In fact, apparently what is meant is that an evaluator is associated with a particular school *either* if they are currently affiliated at that school, *or* if their PhD was granted by that school. This fact is not stated *anywhere* in the main text, so far as I can see, but I was able to deduce it from the x-axis label in Figure 1 (p.663). It's problematic that this isn't clear and explicit in the text, and that vague terms like "from" are used in lieu.

In fact, the correlation between number of evaluators and PGR score is almost entirely due to evaluators' PhD-granting institutions, and has almost nothing to do with their current affiliation. Here's the equivalent of Bruya's Figure 1, comparing the two, when only current affiliation is taken into account (and including UK/Canada/Australia):



Note that:

- The graph is much flatter than Bruya's graph: increase in number of evaluators makes only a very minor difference to score.
- The correlation coefficient is much smaller than in Bruya's analysis (R^2=0.37 compared to 0.61)
- Both the slope and the correlation (such as they are) are quite driven by the two outliers, which are Rutgers and Oxford. (R2 values are sensitive to outliers given that they try to minimise a sum of *squares*). If these are taken out the correlation coefficient drops to 0.26.
- Bruya's "top 8" institutions don't show up. The nearest analogue is Rutgers/Oxford, but those two institutions' 28 evaluators are still less than 10% of the total

So the results Bruya demonstrates in his Figure 1 are almost entirely due to a strong correlation between number of evaluators whose PhD is from an institution, and that institution's rank. With that realised, we can re-evaluate Bruya's claim that "the correlation would immediately call into question the validity of the [PGR]'s conclusions".

Bruya's argument presupposes (as a null hypothesis, to be undermined by the correlation) that the PGR evaluators are drawn randomly from the pool of research-active philosophers. (This is itself an odd assumption as the PGR is reasonably explicit that it's intended to survey leading experts and not to be an opinion poll, but put that aside). Suppose that were the case; what would we expect?

Well, suppose that (i) faculty quality correlates with placement record (at least for research-active positions); (ii) faculty quality changes slowly. Then we would expect that those faculties that have the strongest faculties have the strongest placement record, and so a random sample of research-active faculty would expect to disproportionately represent the strongest schools. That is, a strong correlation between PGR score and number of assessors is *exactly what would be predicted* if the PGR indeed succeeds in measuring something that tracks placement success and if the assessors are randomly selected from the placed. Bruya's claim (p.662) that a correlation would undermine the validity of the PGR *only makes sense if it is assumed in advance that the PGR fails to measure what it purports to measure*.

(As a small illustration of how Bruya's Figure 1 works, notice two outliers: NYU and CUNY, both up in the top left. Both have high PGR scores – in NYU's case, the outright highest score – but very low levels

of evaluators. This represents the fact that they've risen to prominence recently, so that very few of their graduates have reached the point of being asked to evaluate.)

What about the dominance of the Bruya Top 8? Firstly, and trivially, inclusion of non-US evaluators reduces the fraction of evaluators in the Top 8 to 40%, which in itself suggests that Bruya's description of these 8 programs as "driving the PGR" overreaches; it also suggests adding Oxford to the Top 8, which would stretch Bruya's "tight geographical area" to include a few hundred thousand square miles of the North Atlantic. But more seriously, once we recognise that the dominance of these programs is driven by former graduates and not by current faculty, it's not obvious that a 40% representation by these programs in PGR assessors is particularly surprising. These are for the most part schools which have had excellent placement records in research universities, and large graduate programs, for a long time. I'm not knowledgeable enough about US departments to make a reliable estimate of what fraction of research-active US academics studied at one of these departments, but 40% doesn't strike me as implausible even using a relatively minimal criterion for expertise, and becomes even more plausible as a more demanding criterion for expertise is applied. To take an admittedly extreme illustration of expertise, I looked at the PhD-granting institutions for the (alphabetically) first 25 philosophers in the American Academy of Arts and Sciences. 15 of them (60%) got their PhD at a Bruya Top 8 institution, compared to only 6 (24%) who got it from another US university. (And of the other 4, 3 went to Oxford).

Bruya doesn't make any analysis at all of what fraction of research-active US academics would be randomly expected to have a PhD from a Bruya Top 8 school, and without any such analysis his description of these programs as a small group driving the PGR is unsupported. This serious weakness in the argument is unfortunately not obvious to the reader given that it is easy to read Figure 1 as referring to current affiliations.

To summarise:

- This section of Bruya's paper is opaque about its methodology, making it easy for a reader to conflate two readings of an evaluator being "from" an institution: current affiliation, and PhD institution
- The quantitative statistical results Bruya points to rest mostly on the PhD-institution reading
- The arguments Bruya makes would only go through on the current-affiliation reading.

7. Regression analysis on PGR rank and subject speciality uses strongly correlated variables without comment

Bruya calculates the difference, for each US institution, between (i) the PGR rank and (ii) the "Bruya rank", i.e. the equally-weighted sum of all speciality marks (pp.672-3) He then runs a linear regression analysis to determine the dependence of this difference on the equally-weighted sums of speciality ranking in four areas: History, Value, "M&E" (being Bruya's combination of the PGR categories of M&E and Science) and Other.

The problem is that these are very strongly correlated variables: generally speaking, departments with high PGR ranks in one category have high PGR ranks in others, and running regression analyses on very strongly correlated variables has to be done with great care.

To illustrate: suppose I'm interested in whether certain things about US voters (let's say, their salary) can be predicted by their voting patterns. If one of my regression variables is "usually votes Republican in Senate elections" and another is "usually votes Republican in House elections", then it's very

difficult to determine the relative contributions of those two variables to a voter's salary, since each of them is a very good predictor of the other.

You can work around this with a sufficiently large dataset, and there are statistical tests that can be done to ascertain whether a result is still significant even given a high correlation, but Bruya does not mention any such test. Presenting R-squared values without any methodological discussion and without any quantitative analysis of significance is highly misleading.

(Methodologically, what Bruya should have done is run a regression analysis using the *difference* between average speciality ranking in a given area and average speciality rankings overall, since those variables are much less strongly correlated. Had he done so I suspect he'd actually still have found the effect he discusses, since looking at Kieran Healey's analysis of the PGR certainly suggests better returns to specialisation in M&E (though less obviously in Science) and relatively poor returns to specialisation in History. But I can't know this (not without getting Bruya's dataset, or doing a long data-entry exercise of my own) and it's certainly unsafe to conclude it from Bruya's analysis.

Might Bruya actually have done this? Well, (i) that's not what he said he did; (ii) if he did, he'd have to have left out one of the larger categories, since they'd sum very close to zero for any institution, and run the analysis on the others. Otherwise the correlation between the variables would be even tighter – close to linear dependence. But he reports R-squared values for all the categories.)

8. Conclusion

A large fraction of the "data-driven" part of Bruya's paper is open to severe criticism on methodological grounds, quite apart from one's assessment of the more qualitative issues. The severity of the criticisms is such that it's very hard to see the paper passing peer review in any journal of the quantitative sciences. Indeed, the criticisms in sections 2 and 6 of this note – the silence, in the main part of the paper, about the reclassification of the PGR "Science" category as M&E; the silence, everywhere but in a graph index, about the meaning of an evaluator being "from" an institution, and the use of inconsistent interpretations of that meaning in data and in discussion – would in other contexts be troublingly close to academic malpractice.

I don't use that term lightly and I don't intend any accusation of malice to Professor Bruya. But as a discipline Philosophy needs to be extremely concerned if it allows publication of material that uses the methods of other academic disciplines but which fails to pass the basic methodological standards of those disciplines.

Changelog

V2: added section 6; added page numbers; removed the word "conceal" on page 2, which unintentionally implied deceit.

V3: added section on correlation coefficients; rewritten conclusion to note that section 6 objection is comparable in severity to section 2 objection.