Expertise versus Diversity across Epistemic Landscapes

Patrick Grim[1], Daniel J. Singer[2], Aaron Bramson[3], Bennett Holman[4], Sean McGeehan[2] & William J. Berger[5]

[1] Suny Stony Brook
[2] University of Pennsylvania
[3] RIKEN Brain Science Institute
[4] Underwood International College, Yonsei University- Corresponding author
[5] University of Michigan

Abstract

Using computational simulations, Hong and Page have argued that 'diversity trumps ability' (Hong & Page 2004, 2009; Page 2007, 2011), and their result has been used to justify a variety of policies favoring diversity in scientific and other communities. By extending the Hong and Page model to smoother landscapes, we show that the original finding is sensitive to the degree of randomness in the epistemic landscapes that are used to model the questions the community faces. On smoother landscapes the result reverses: it is often 'ability' that trumps 'diversity.' Moreover, it is only by using less random landscapes that the model represents expertise that goes beyond merely performing well on a single question. In this form, the model offers a more nuanced picture of how diversity, ability, and expertise relate. Models of this sort can offer support for various diversity policies, but extra care must be taken in applying the result to specific cases.

Introduction

A number of computational models of group inquiry have garnered wide-spread attention (e.g. Alexander, 2013; Grim 2009; Zollman, 2007). In a formal model of this type, Hong and Page (2004) demonstrate a 'diversity trumps ability' (DTA) result: for a group of agents confronting a task "a randomly selected collection of problem solvers outperforms a collection of the best individual problem solvers" (Page 2007, p. 162). Because the result suggests that an organization is epistemically better off by recruiting a diverse set of candidates instead of just selecting the best individual performers, the Hong-Page result has been taken to have profound implications for policy. The work has been presented to NASA, cited by the USGS, is one of four works cited in support of positive expected institutional effects of UCLA's (2014) proposed diversity requirement, and has recently been appealed in support of promoting diversity in the armed forces in a brief submitted to the Supreme Court (*Fisher v. University of Texas at Austin*, 2016). A number of philosophers have taken the result to apply to the value of epistemic diversity in scientific communities (e.g. Bright, 2016; Martini, 2014; Stegenga, 2016).

Importantly though, there is a difference between the careful way that the Hong-Page result is discussed by those who are sensitive to the details of the model and by others who have interpreted and applied it in a broad range of social contexts. Mayo-Wilson, et. al. (2011) responsibly cite the result to support the "independence thesis"—the claim that the properties of an epistemic community can differ from the properties of their agents. Here the claim is simply that the best epistemic groups are not necessarily composed of the individually highest-performing agents. Such a claim is right in line with Hong and Page's (2004) conclusion that, under the right conditions "a diverse group can often outperform a group of the best" (p. 16386). Other modelers have also been careful to retain the qualified claim that diversity *can* trump ability (Bright, 2016; Zollman, 2011; 2013).

But the result has also been cited in support of a much stronger claim: that diversity is generally (or even necessarily) epistemically beneficial. Nunn (2012) relies heavily on the result to argue that the medical community would be better off if it moved away from evidence-based medicine and incorporated a plurality of "medical models" (e.g. narrative medicine, evolutionary medicine, and complexity medicine) noting: "It is not an a priori assumption or mere hand-waving optimism to say that people working with many models and their associated methods do better than those working with only a few models and methods. Rather, it is a claim that…is grounded in the modeling experiments and theorems of Hong and Page" (p. 976).

The result is also frequently invoked in support of diversity initiatives in science. In support of the claim that consensus conferences are more likely to consider all the relevant evidence if they are socially inclusive, Stegenga (2016) relies on Hong and Page (2004) in claiming that "mathematical models have been employed to show that diverse groups of problem solvers outperform groups of high-ability problem solvers" (p. 45). Claims that the Hong-Page result provides evidence that increasing underrepresented groups will result in disciplinary gains have been made in computer science (Cheryan, Plaut, Handron & Hudson 2013), biomedical engineering (Chesler, Barabino, Bhatia & Richards-Kortum 2010), biomedical science (Gibbs, McGready, Bennett & Griffin 2014; Pickett, Corb, Matthews, Sundquist & Berg 2015) and STEM fields generally (Ferrini-Mundy 2013; Hanley, Brown, Moss-Racusin & Smith 2015). The results have also been cited in support of specific scientific practices, such as developing programs to eradicate tuberculosis (Quissel & Walt 2015).

Here we generalize the Hong-Page model to see how well the claims about diversity and ability hold up. In the first section, we introduce and replicate the Hong-Page result. In section two, we argue against interpreting “best-performing agents” as “experts” in the Hong-Page model. In section three, we modify the Hong-Page model minimally to enable the model to shine light on some forms of expertise. In those modified models, however, it is ability that trumps diversity. In sections four and five, we show that whether diversity or expertise triumphs is affected by other aspects of the model as well, such as what method the group uses to work together and how large a set of problem-solving methods are available. What our results suggest is that even within the highly abstract models considered by Hong and Page, though diversity does trumps ability in some cases, ability trumps diversity in others. Given that both diversity and ability have their place, unqualified assertions of a general triumph of diversity over ability are unwarranted.

I. The Hong-Page Result

Hong and Page offer several variations of a formal model of a group working together to solve a problem (Hong & Page 2004, 2007; Page 2007, 2011). In those models, agents use heuristics to explore an epistemic landscape. The DTA result is that epistemic outcomes for groups of randomly-chosen individuals will consistently exceed the performance of groups composed solely of the best-performing individuals. Although accompanied by a mathematical theorem intended to offer partial understanding, the main result comes from simulations rather than in formal proof. The same will be true of our work here.

In Hong and Page's original model, epistemic exploration proceeds along a looped terrain of 2000 points (so 10 points to the right of 1995 is point 5, for example). For each of the 2000 points of the terrain a height is assigned as a random integer between 1 and 100; higher points are interpreted as better answers to the question. Individual agents are identified by a *heuristic,* modeled as an ordered set of k numbers, each of which is a number between 1 and l. We begin, following the Hong and Page original, with ordered sets of 3 numbers (k =3) between 1 and 12 (l =12). With these parameters, there are 1320 possible agents defined by distinct heuristics (respecting order but avoiding duplication).

Individuals use their heuristics as follows. An agent starts at, say, location 112 of the 2000-point terrain, which carries a value (height) of 80. The agent then applies the first number of its heuristic by asking: Does the point that many steps to the right offer a higher value? If not, it stays put. If so, it moves to that point. In either case it then applies the second number of its heuristic. Does that offer a point with a higher value? If so, it moves to that point. It then uses the third number in the same way. Once the third number has been tried it starts over with the first number. An individual stops only when none of its numbers can reach a higher point; i.e., it has reached its local maximum via applying the cycled heuristic from the initial point of 112. In exploring the terrain in this way, there is a determinant value that the agent reaches starting at each of the 2000 points. The average of those is an individual's score, which we use to rank our 1320 agents. The 9 'best' individuals will be those with the 9 highest scores.

As a model for discussion within a group, Hong and Page employ a sequential ('relay') activation of the agents. Consider a group of 9 participants. Starting from a given point, the first agent uses her heuristic to find the highest point within her reach. Once she has found her maximum reachable height she passes the "baton" to the next agent who begins where she left off. He then searches for a higher maximum by employing his heuristics until his search is exhausted, at which point he passes the baton to the third agent, and so forth until all nine agents have exhausted their searches. At that point the baton is again passed back to the first agent on the list and the agents are activated one by one in the same order. The final decision for the group will be the local maximum from which none of the agents can find a higher point. The discussion can be thought of as a conversational relay that proceeds in orderly fashion around a circular table. The score for the group will be the average height achieved using each of the 2000 locations as starting points.

What Hong and Page compare are the results of a modeled discussion of this form for (a) a group composed of random individuals drawn from the heuristics pool at large and (b) a group composed of the 'best' individuals—those with the highest individual scores. The DTA result is the fact that the random group consistently does better. In our reproduction of the Hong-Page result we compare the scores of (a) 9 random agents and (b) the 9 agents with the highest individual scores. We found the average of the maximal heights reached by the 9 'best' individuals over 1000 different random landscapes to be 92.53 (median 92.67). This is compared to 94.82 (median 94.83) for a group of 9 random individuals, indicating an improvement of roughly 2%. We found a higher score achieved by random agents in 97.6% of the 1000 runs.

In their original presentation, Hong and Page support the claim that diversity trumps ability with both simulations and a mathematical theorem. What the theorem shows is that given strict conditions regarding group and population size and specific definitions of problem difficulty and group diversity, 'diversity trumps ability' with probability 1. Thompson (2014) challenges both the intrinsic value of the mathematical theorem and its relevance to Hong and Page's conclusions. In both their simulations and our replications, the strict conditions required for the theorem are significantly relaxed. Even though the probability falls below 1, the results above show that the central result is still strongly supported. Within the simulation parameters specified the epistemic success of a collection of random heuristics proves consistently superior to that of a collection of those which individually score the best.

II. Interpreting Hong-Page: Best Performance, Ability, and Expertise

The Hong-Page result is extremely suggestive, and it has been offered as support for a number of strong conclusions already mentioned. In both their original work and in later applications, Hong and Page allude to diversity as a value in affirmative action (Hong & Page 2004, Page 2007). They also draw conclusions regarding business and research teams: "When selecting a problem-solving team from a diverse population of intelligent agents, a team of randomly selected agents outperforms a team comprised of the best-performing agents" (Hong & Page 2004, p. 16385). It is to their credit, we think, that Page and Hong tend not to use the term 'experts.' In reviews and applications of their work, however, it is probably natural that their results are taken as part of a larger case against expertise (Landemore 2013, Gunn 2014, Weymark 2015). For example, the Princeton University Press's blurb on the back of the book characterizes Page's *The Difference* as revealing "how groups that display a range of perspectives outperform groups of like-minded experts." Elizabeth Anderson also characterizes

Hong-Page as showing “that diverse collections of nonexperts do a better job than experts in solving many problems,” supporting the claim that “democracy, which allows everyone to have a hand in collective problem solving is epistemically superior to technocracy, or rule by experts” (Anderson 2006, 12).

Following Page and Hong (as well as Page, 2007 and 2011), we think of a landscape as representing a specific problem or question; for example, which of this year’s cars has the best gas mileage (Page 2007) or what is the best combination of drugs to treat a particular illness. Importantly, Page and Hong model these as completely random landscapes, where there is no correlation between the heights of any positions on the landscapes and any others. Because of that, different landscapes produce ‘best-performing individuals’ with very different heuristics, and an individual that is ‘best-performing’ on one random landscape is likely to do extremely poorly on another. No matter how linked two problems or the techniques for solving them may be—e.g., calculating gas mileage not merely for 2016 models but for 2017 and 2018 as well—by modeling those problems as distinct random landscapes, best-performing heuristics cannot be expected to carry over from one problem to its close relatives. This spells trouble for interpreting the DTA result as applying to experts, we’ll claim.

Table 1 shows the top 9 heuristic sets in 10 different model runs on different random landscapes. As you’ll notice, there is a large amount of redundancy of heuristic numbers among the ‘best-performing’ agents on each landscape. On the first landscape, for example, the numbers 4 and 12 appear in every one of the ‘best-performing’ heuristic sets. The redundancy of the ‘best-performing’ set is a major part of Hong and Page’s own analysis of both formal results and social implications: why hire 5 individuals with the same background if you will just hear the same message five times? But there is clearly something arbitrary about the numbers that show up as part of a successful heuristic. While the numbers 12 and 4 appear in all of the best-performing heuristics for the first landscape in Table 1, neither number appears in any of the best-performing heuristics for the second or third landscapes.

#	Heuristic Sets for the 'Best-Performing' Agents
1	(12 4 5) (12 2 4) (12 5 4) (12 4 2) (5 12 4) (4 12 2) (6 12 4) (4 5 12) (12 4 6)
2	(5 7 6) (10 8 7) (8 7 10) (7 10 8) (7 5 6) (7 8 10) (11 10 8) (5 6 7) (10 11 8)
3	(1 10 3) (1 6 2) (1 3 10) (3 1 10) (6 2 1) (10 3 1) (10 1 3) (1 10 6) (7 5 3)
4	(11 4 1) (12 2 8) (11 2 12) (4 11 1) (11 1 4) (4 1 11) (12 11 2) (5 8 2) (8 12 2)
5	(6 1 2) (3 6 1) (6 1 3) (1 2 7) (3 6 2) (1 3 6) (2 6 7) (7 1 2) (1 2 6)
6	(4 8 7) (3 4 8) (4 8 3) (7 4 8) (4 3 8) (1 8 7) (3 8 4) (3 8 7) (8 7 2)
7	(3 12 1) (1 3 12) (12 1 3) (3 1 12) (8 3 12) (11 12 8) (1 8 12) (12 1 8) (12 3 1)
8	(2 6 11) (11 2 6) (6 11 2) (11 6 2) (6 2 11) (9 6 11) (2 11 6) (11 9 6) (11 6 9)
9	(8 7 2) (8 2 7) (2 7 8) (8 6 7) (6 8 7) (7 6 4) (6 7 8) (7 8 6) (2 8 7)
10	(2 8 3) (8 3 2) (12 11 3) (3 12 11) (12 3 11) (11 3 12) (2 3 8) (11 12 10) (12 11 10)

Table 1. Heuristic sets for the 'best-performing' agents on 10 different fully random landscapes

Table 2 shows the percentage of cases in which each of the 12 heuristic numbers appears among the 3 heuristic numbers of the 9 'best performing' agents on 100 random landscapes. Importantly, each heuristic number shows roughly equal representation across the random landscapes as a whole. What this shows is that, though certain heuristics perform best on individual random landscapes, that fact that a heuristic performs highly on one random landscape tells us very little about how it will perform on another random landscape.

Percentage of 'best-performing' in which each heuristic value appears

1	2	3	4	5	6	7	8	9	10	11	12
22.7	21.1	19.2	22.3	22.2	24.7	28.3	23.4	31.6	24.3	31.4	28.3

Table 2. Percentage of cases in which each value appears among the 3 heuristic numbers of the 9 'best performers' on 100 random landscapes.

In the original Hong-Page model, the 'best-performing' on a specific landscape might therefore be better thought of as the 'luckiest' on the landscape: those that happen to have heuristic sets attuned to that specific case.

Agents who get things right only by luck hardly qualify as experts. A minimum requirement of genuine expertise, which merely high-performing agents don't meet, is that experts can be expected to perform well on many different questions in their field of expertise. Experts at judging car fuel efficiency who do well on cars produced in 2016 can also be expected to do well on cars produced in 2015 and 2017, since the same methods used to get a good estimate in one year should work to find a good estimate in others years. Gaifman (1988) and Elga (2007) assume this standard in treating someone as an expert in a field when the probability is high that their opinion is right across questions in that field.

The same point applies to a natural conception of 'ability.' While there may be a weak notion of ability whereby someone has ability when they can succeed at just a single instance of a task, a more natural conception of ability treats the needed success as counterfactually robust or transportable: one has an ability to Φ only if one is likely to succeed at that task under a range of conditions (though, of course, the breadth of that range is context dependent). Someone who has the ability to judge the quality of livestock should be able to give us reliable results across multiple herds, for example. An ability to predict the weather requires being disposed to do that well in more than one case.

We'll explore this point further below. We'll argue that Hong and Page's best-performing agents on random landscapes don't meet the minimal requirement of expertise and ought not be considered to have an ability (in the more natural sense). So, while our replication of Hong and Page's simulations shows that the formal result is secure on random landscapes, that result might not have the implications that some have taken it to have. To show that, we explore variations on the model that allow the model to be interpreted in terms of transportable ability and expertise.

III. Ability over Diversity on Smoother Epistemic Landscapes

On the purely random landscape used in the original Hong-Page model, nearby points on the landscape are uncorrelated. Although locations closer to each other on a landscape are more likely to be within the reach of one of the heuristic numbers of an agent, the assigned values or 'answers' at proximate points may or may not have similar heights. If we introduce correlation between the heights of nearby points, we create 'smoother' landscapes. In this section, we explore the robustness of the DTA result on those smoother landscapes.

The interpretation of these smoother landscapes is fairly natural if we conceive of the heuristics as investigatory strategies. Some landscapes allow for strategies that can 'hill climb,' using heuristics that improve incrementally from one answer to a superior neighbor. For these problems a proper methodology allows a sense of incremental progress. Other problems offer essentially no advantage to hill-climbing: a move to nearby solution is just as likely to yield progress as moving to a completely remote part of the landscape. It is this second kind of question, which we might call "strategy-resistant," that is best represented by the random landscapes of Hong and Page's original model. Here we introduce a parameter for the correlation of location and height (smoothness) and vary this parameter in order to explore the relative success of groups of diverse versus best-performing agents on the wider class of problems represented by smoother landscapes.[1]

One way of smoothing a random landscape is to interpolate values between a number of randomly-set points. For example, instead of assigning a random value to each of 2000 points, we could set the height of just the even-numbered locations randomly, filling in the value of the odd points as the means of their immediate neighbors. The result would be a landscape that is less random and less rugged than the original. For a still smoother landscape, we might assign random values to every third point, or every fifth point, and fill in the gaps by drawing descending or ascending lines between the assigned points.[2]

We construct smoother landscapes using a slightly more sophisticated version of this idea. We assign a random height value to point 1. For a *smoothing factor* of x, we pick a random integer between 1 and $2x$ and assign a random height value to that point. The locations of our assigned points therefore have an average distance of x, but without the artificiality of a fixed-length interval. Points between the assigned ones are positioned on a line of ascending or descending values between them (rounding the heights to integer values). Example epistemic landscapes with smoothing factors of 0, 5, 10 and 20 are shown in Figure 1.

---

[1] Smooth landscapes are by no means the only variation worthy of study. Many problems, for example, including many problems in science, might be better modeled using NK-landscapes (Alexander, et al. 2015; Fontanari, & Rodrigues, 2016).

[2] Hong and Page emphasize a number of conditions on their result, one of which is that the problem to be solved is ‘difficult.’ Their specification for a ‘difficult’ problem is that there be no individual problem solver who always finds the global maximum (Page 2007, p. 159). All the landscapes employed in our models count as difficult in this sense.

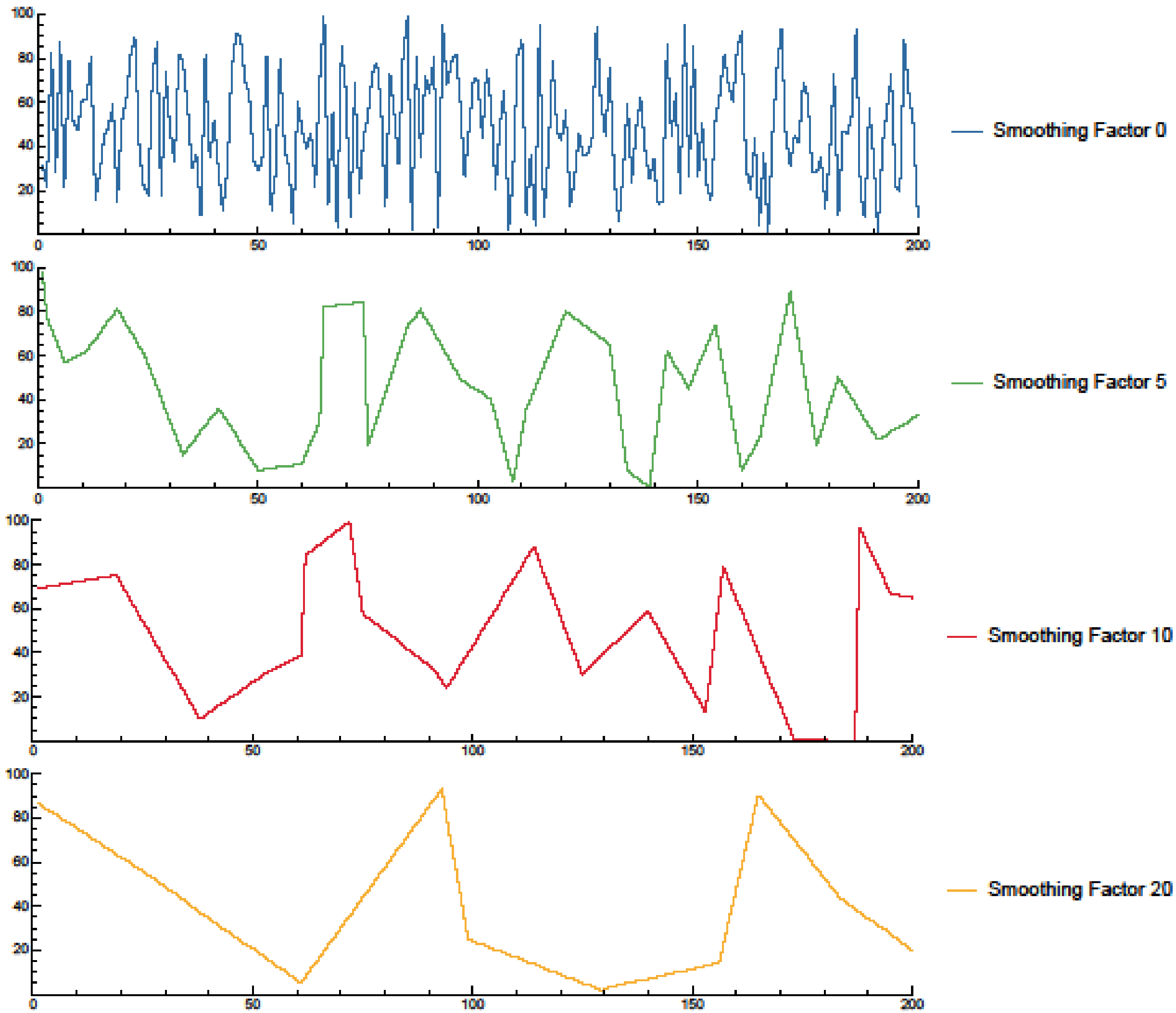

Figure 1. Sample landscapes (up to 200 positions) created with smoothing factors of 0, 5, 10 and 20.

We now ask how robust the DTA result is with increasing landscape smoothness. As before, our agent heuristics consist of ordered sets of 3 numbers between 1 and 12, resulting in 1320 possibilities. Over 100 distinct landscapes, we average the values of the final heights reached when starting from each of 2000 points for (a) a relay group of 9 random individuals and (b) a relay group of the 9 individuals that perform best individually. For landscape smoothness factors from 0 to 20, Figure 2 plots the difference in performance (random group scores minus best-performing group scores). The cross-over point at a smoothing factor of 4 indicates the point at which the DTA result no longer holds. Below this value, random groups outperform groups of the highest-performing individuals. In Hong and Page's terms, 'diversity trumps ability.' Above that value, it is groups of the highest-performing that do better. Here 'ability' trumps diversity.

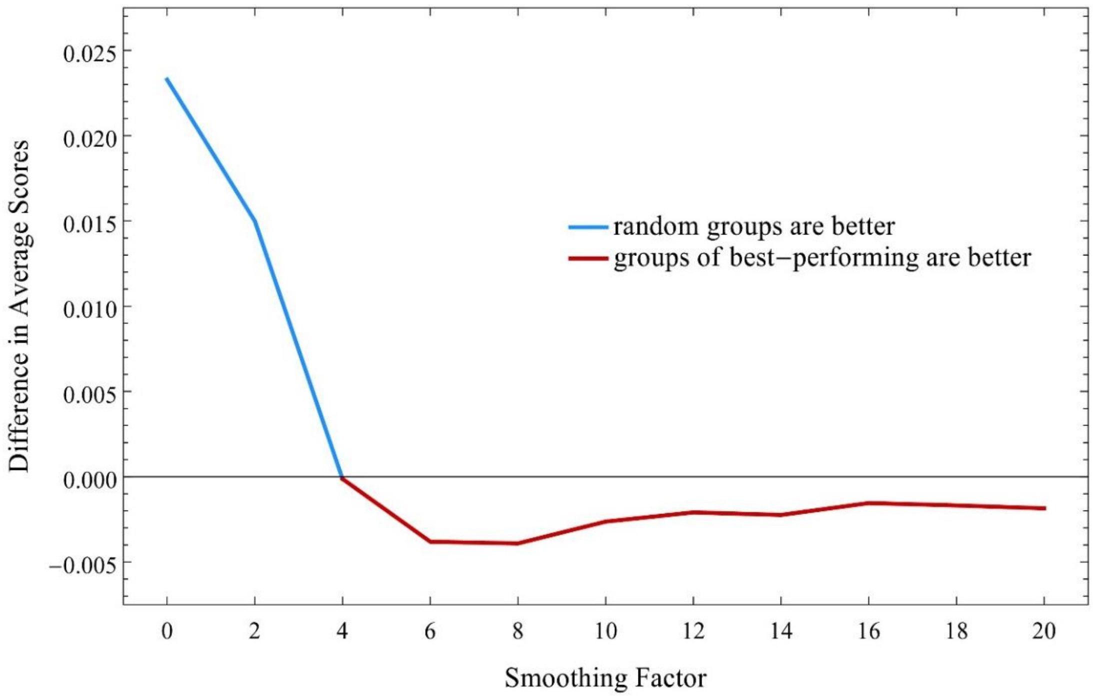

Figure 2. Differences in average performance over 100 different landscapes for groups of 9 individuals using 3 heuristics from a pool of 12.

When groups of the highest-performing individuals do better, the advantage is small: at a smoothness factor of 6, for example, the average performance over 100 landscapes is 0.756 and 0.760 for random and 'best' groups, respectively. That small advantage of the best over the random is, however, clear and robust beyond the cross-over value of 4.

Here we return to the issue of ability and expertise mentioned above. If we think of a landscape as representing a particular question within a particular discipline or subject matter, for the best performing agents to be considered to be experts or to have an ability, we'd expect to see their skill as transportable: we would expect them to perform roughly as well on other related questions.[3] Smoother landscapes do exhibit this form of transportability of best-performing heuristics. Interpreting 'best-performing' as reflecting ability or expertise thereby becomes more plausible on smoother landscapes than on the random landscapes of Hong and Page's original model. Here's how we can tell:

We generated pairs of landscapes of equal smoothness and found the Pearson correlation of the performance of each of the 1320 agents on those landscapes. This process was repeated 100 times to obtain the average value for a given smoothness factor and the entire process was repeated for each smoothness factor from 0-10. The square of the Pearson correlation tells us what percent of the variance in performance is explained by the individuals' heuristics and thus to what extent performance on one landscape predicts performance on the next (Figure 3).

[3] Thinking of alternative landscapes as different specific questions within the same general problem space follows Page (2007, 2011). An interpretation that demands a content-distinction between sets of questions in different fields would require significantly more complex modeling assumptions.

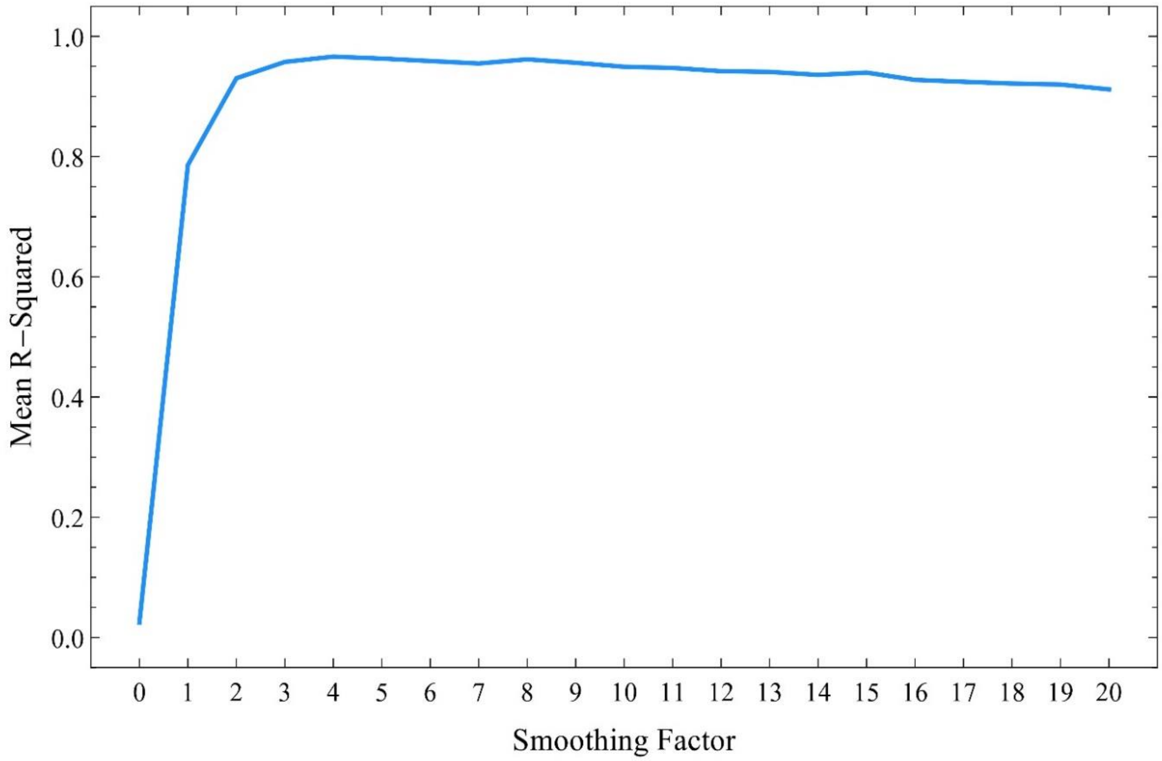

Figure 3. Average percentage of variance of performance of an agent on landscapes of the same smoothness that is explained by individual heuristics

Figure 3 shows a clear and sudden initial jump as smoothness increases in the correlation of individual performance on one landscape to another of the same. It is only after the jump, we propose, that interpreting results in terms of 'ability' or 'expertise' becomes plausible, since it is only then that one's performance on one landscape is correlated with one's performance on another. The highest level of transferability occurs approximately when it is no longer true that diverse groups are best-performing (smoothness 4). So it is roughly where an interpretation of high performance in terms of expertise becomes more plausible that groups of the best performing start to outperform groups of random heuristics.[4]

[4] Here we mean "expertise" in the sense we discussed above, as being transportable between landscapes. That said, one might think that expertise need only be transportable to different places in the landscape, either because one rejects our interpretation of landscapes or our

By looking at who the experts are in these smoothed landscapes, we can also say something more about what's involved in the expertise modeled. Recall Table 2 showing the percentage of cases in which each of our 12 numbers appears among the 3 heuristics of the 9 'best performing' agents on 100 landscapes. In that case there appears to be no clear signature of the transportability required for general 'ability' to exist: each heuristic number occurs roughly equally across the random landscapes. Table 3 expands Table 2 to show the percentage of cases in which each of our 12 numbers appears among the 3 heuristic numbers of the 9 'best performing' agents on 100 landscapes with increasing smoothing factors.

Smoothing Factor \ Heuristic Number	1	2	3	4	5	6	7	8	9	10	11	12
0	22.7	21.1	19.2	22.3	22.2	24.7	28.3	23.4	31.6	24.3	31.4	28.3
1	100	0	3.8	23.4	20.4	17.0	21.4	19.9	19.3	22.4	20.6	31.4
2	100	0	0	0.2	10.3	32.7	35.7	19.6	3.7	11.1	24.4	62.3
3	100	0	0	0	1.7	21.0	46.4	26.8	4.1	0.4	13.8	85.8
4	99.4	0.5	0	0.5	7.8	23.4	33.8	27.7	6.6	0.1	3.9	96.1
5	98.7	1.4	1.7	5.8	14.2	21.6	26.8	21.0	7.7	1.2	0.3	99.7

Table 3. Percentage of cases in which each number appears among the 3 heuristic numbers of the 9 'best-performers' on 100 landscapes for smoothing factors 0 through 5.

While a purely random landscape (smoothness 0) shows no consistent bias toward any specific heuristic numbers, a pattern immediately emerges at smoothing factor 1; specifically, the

---

understanding of expertise. That said, the same story plays out if we compare intra-landscape performance. We tested the average percentage of variance of performance of an agent on the first half of a landscape and the second half of the same landscape that can explained by individual heuristics, and the graph was virtually identical to the one described above. Just as with cross-landscape comparisons, it showed a clear and sudden initial jump in the correlation of individual performance on both halves of the landscape as smoothness increases. So again here, our result about expertise only playing a role in higher smoothness landscapes applies even if you reject our interpretation of expertise needing to be transportable *between* landscapes.

heurtistic 1 appears among all of the best 9 heuristics in all 100 cases. The large numbers, and 12 in particular, become increasingly prevalent as the smoothing factor increases, as do middle numbers around 7. On the other hand, the number 2 disappears entirely at smoothing factor 1, joined by the number 3 and then 4 with increasing smoothness. The numbers 2, 3, 4, 9, 10, and 11 all become rare quite quickly. In all these cases, unlike the random landscape, it is clear that there are certain patterns of heuristics—the 'expert sets'—that tend to do best quite generally across most landscapes of a particular smoothness.

There are a number of possible explanations for why we see this particular pattern of which heuristic values do best on these smoothing factors. The heuristic 1 is valuable because it is the ultimate hill-climber: should other numbers in rotation not interfere, repeated access to '1' alone would allow a heuristic to climb to the highest point on any incline to reach a local maximum. With 1 present, 2 is at best redundant on landscapes with smoothing factor 1, and potentially disruptive—pushing one over the top of a local maximum to a decline on the other side—hence its total disappearance at smoothing factor 1. This phenomenon also accounts for the disappearance of 2 and 3 given a smoothing factor of 2, and of 2, 3, and 4 given a smoothing factor of 3. The value of 12—the highest number available—is that it offers the best hope of leaping over declines to an incline on the other side of a valley. Because the width of valleys widen as smoothing factor increases, this capability becomes increasingly important. Values in the middle are likely useful on occasion for jumping over more narrow valleys, but are less precisely selected for partly because the widths of these valleys vary. They are also weakly selected for plausibly because having 1 and 12 suffices for high performance, and the third number which will put an agent in the top 9 depends on the specific shape of the landscape. The data we present here doesn't completely determine exactly why these heuristics are best though.

Those details aside, the broader lesson of these results is a warning against accepting the DTA result without a qualification regarding the character of the epistemic landscape at issue. Keeping other values in the Page and Hong simulation constant, groups of random agents do better than groups of high-performing individuals only for a very narrow range of highly random landscapes. For smoother landscapes, those on which successful individuals are more plausibly viewed as experts, it is the groups of high-performing individuals that do better. Given the other assumptions in play, it is ability that trumps diversity on smoother landscapes.

IV. Diversity over Ability with Larger Heuristic Pools

In the previous section, we showed that with a heuristic pool limited to numbers between 1 and 12, there is a cross-over in favor of experts once the smoothness factor exceeds 4. Beyond that point DTA, no longer holds. What happens when the heuristics can be any triplet of numbers from 1 to 16 or 1 and 20, rather than being confined to numbers from 1 to 12 though? What we see is that diversity again shows its strength. We'll explore that variation on the model here.

Recall that at a smoothness 8, the best-performing do better than a random group when the three numbers of heuristics are chosen from a set of 12 numbers (see Figure 2). When heuristic numbers are chosen from the set 1 to 24 or more, however, the group of random heuristics again does better. Figure 4 shows the difference in average score for groups of random heuristics minus the best-performing as we increase the size of the heuristic pool for a smoothing factor of 8.

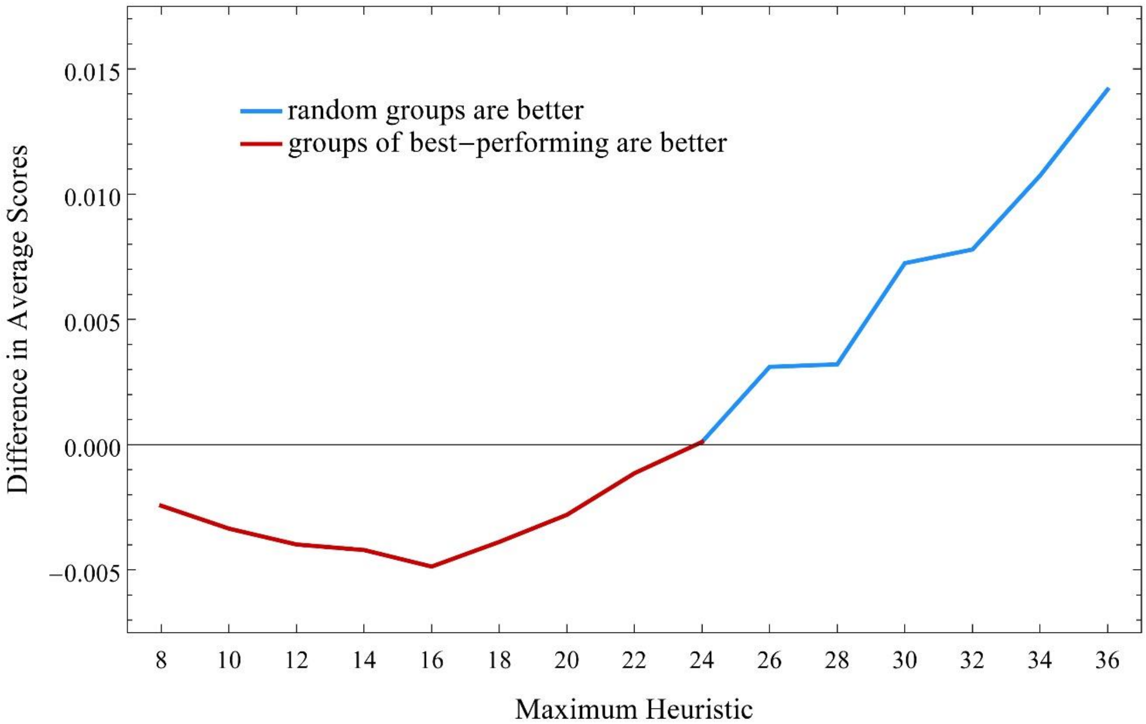

Figure 4. Using a smoothness of 8 as an example, a crossover in favor of random groups occurs when the maximum heuristic number is set to 24. Differences in averages over 100 landscapes shown.

A similar crossover occurs for other smoothing factors as well, and as we vary the smoothness we have found a very rough ‘rule of three’ for the value where this crossover occurs. For heuristic pools that are *less than three times the smoothing factor* of the landscape, the best-performing outperform random groups (as outlined in the previous section). For heuristic pools roughly three times the smoothing factor or greater, we once again see a DTA effect. Although increases in landscape smoothness favor groups of the best-performing, such an advantage is always relative to the maximum value in the heuristic pool from which strategies are drawn.

The virtues of diversity and ‘ability’ are therefore relative to the interaction of at least *two* important factors: landscape smoothness and heuristic pool. Figures 5 through 7 show a parameter sweep across both variables, indicating distinct areas of relative strength for diverse groups as opposed to groups of the individually best-performing.

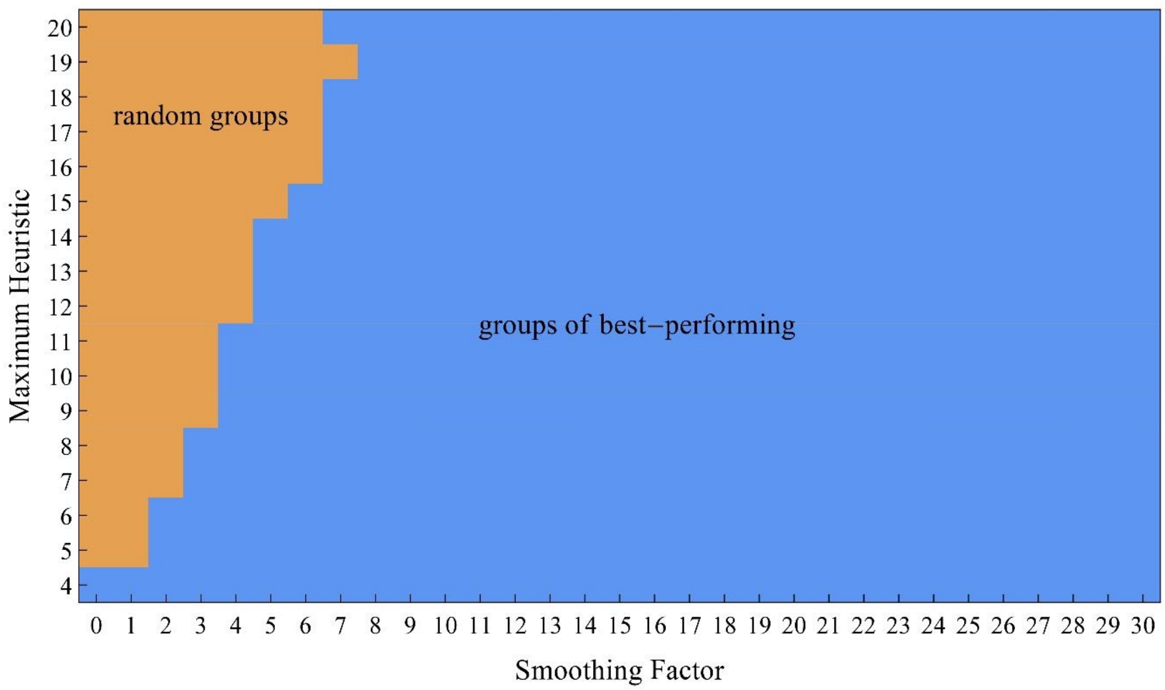

Figure 5. Parameter combinations for which groups of random heuristics do better (orange) and areas in which groups of the best-performing perform better (blue).

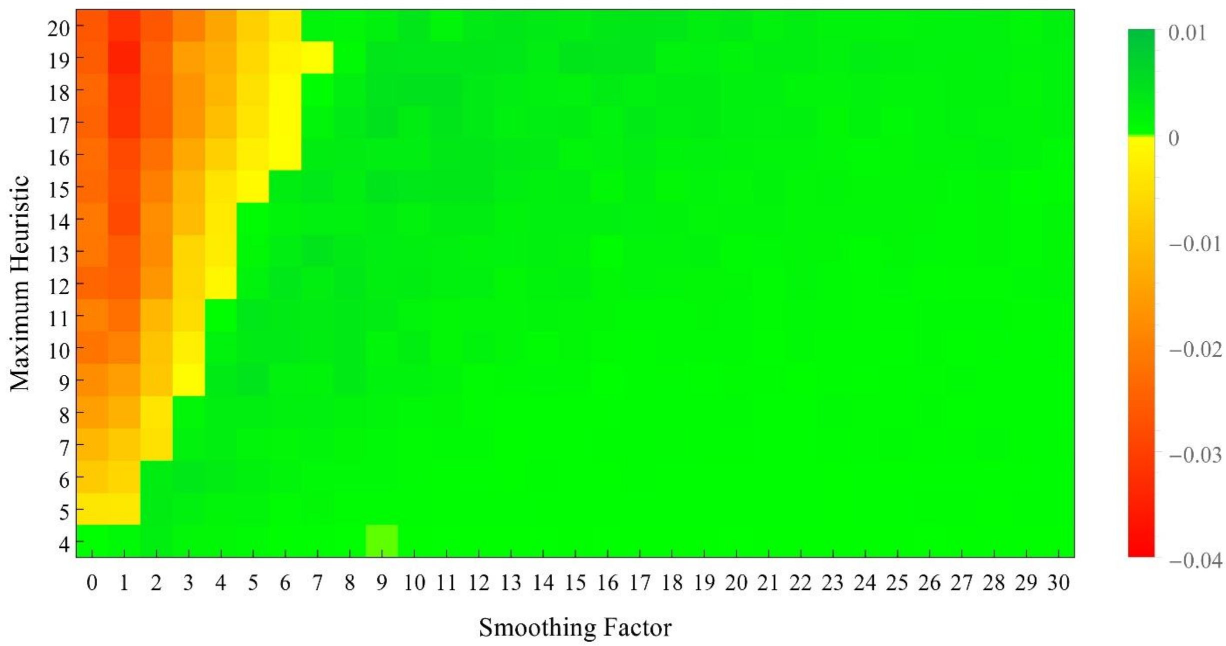

Figure 6. Differences in average scores for groups of random heuristics and groups of the best-performing over 100 landscapes at different parameter combinations of smoothing factor and heuristic pool. Positive values (in green) show higher averages for groups of the best-performing. Negative values (in yellow and red) show higher averages for groups of random heuristics.

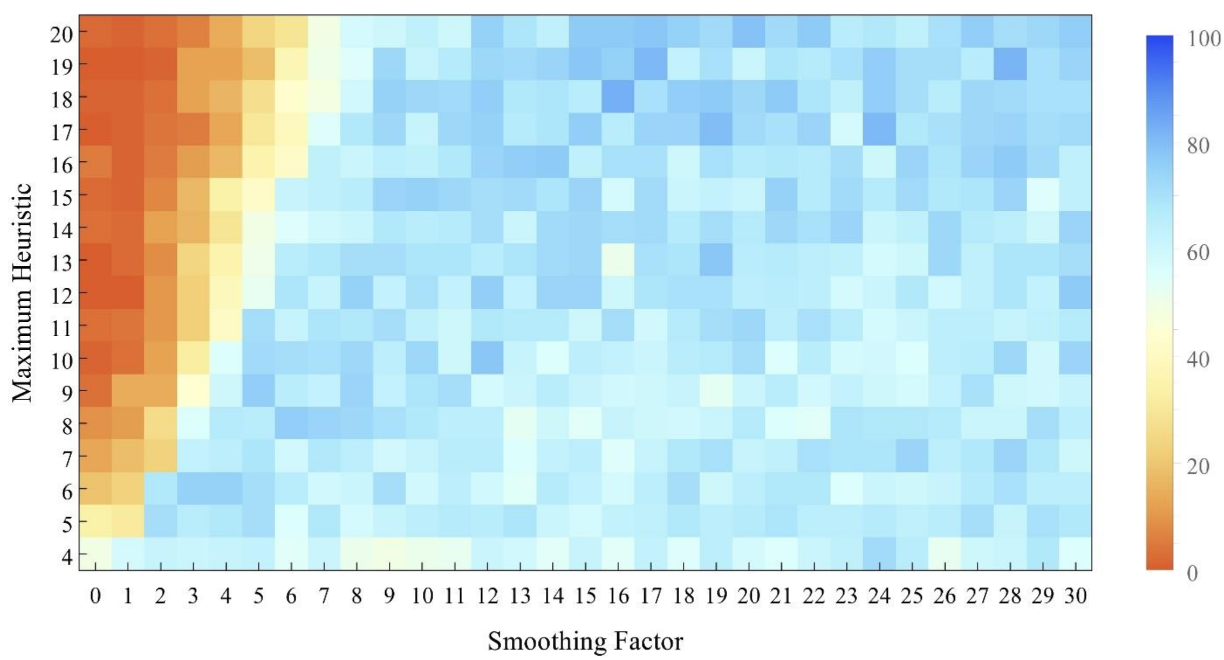

Figure 7. Percentages of runs in which groups of the best-performing do better than groups of random heuristics.

Figure 5 presents the data in the roughest form, showing those areas in which the average score for each is greater over 100 landscapes. Figures 6 and 7 show the more nuanced reality behind this result. Even where an average over 100 runs is higher for diversity as opposed to 'ability,' the difference may be very slight. Figure 6 shows the same data mapped in terms of the difference in the average scores. Figure 7 shows the percentage of 100 runs in which a random group or group of the best-performing does better at each setting of maximum heuristic and landscape smoothness. This demonstrates that even at combinations where best-performers or randoms generally win out, there are still some landscapes for which the other set does better.

If we think of the heuristic pool as representing the conceptual or problem-solving resources available to tackle the problem, what these results suggest is that there are different niches in which groups of the best-performing and diverse groups are each of particular value. Groups of the best-performing are better for a wide range of smoothness, but only where the available conceptual resources are relatively limited. With a wider pool of conceptual resources, a diverse group will do better even on problems of that same character.[5]

One key factor to understanding many of these results, we think, is the extent of heuristic coverage represented in groups. As landscapes increase in smoothness, the best-performing individuals tend to become very much alike, as indicated in table 4. A small number of the available heuristic numbers will be best on landscape of that smoothness, and all of the individually best-performing will share that small set of numbers. A *group* of the best-performing will therefore show high redundancy: their collective numbers will not be dense on

[5] As documented in auxiliary materials online, we have tested the robustness of these results with groups of sizes 3, 6, and 9 (bit.ly/DAESupp). The smaller the group, the greater the advantage for groups of the best-performing. The larger the group, all things considered, the greater the advantage for groups of random heuristics.

the space of heuristic numbers. So, contrary to Thompson (2014), we suggest that it is not the randomness that is an epistemic virtue of random groups, but the extent to which their heuristics collectively cover the available space.[6] Among all the members, a group of random strategies will have more numbers to try, and so have a prospect of reaching higher peaks and avoiding more local maxima.

As Page and Hong hint in their original work that greater coverage explains the success of groups of random heuristics on random landscapes: the random groups have more heuristics to work with in their union, and so have a greater number of options to pursue in finding the highest peaks. Employing coverage, rather than randomness, also helps explain why groups of the best-performing do better with smaller heuristic pools but random groups pull ahead with an expanded heuristic pool: In the larger heuristic pool, the percentage of 'expert' numbers is smaller, so they are relatively more concentrated. Even groups of random heuristics have some redundancy, but in a larger heuristic pool that expected redundancy will be smaller. So, at least one reason why random groups do better with increased heuristic pools than groups of the best-performing seems to be because their coverage of available heuristic numbers increases with a larger pool.

V. Discussion Dynamics: Diversity and Expertise

There is a further factor that surprisingly and dramatically favors diverse groups and is largely ignored in other discussions of the Hong-Page result. The original Hong-Page model uses a 'relay' dynamics, as we have done above. Starting from a given point, the first agent in the

---

[6] We explore this point more in our discussion of mixed groups in the online supplement (bit.ly/DAESupp).

group finds the highest point her heuristic will reach. The second agent then starts from that point in search of a higher one and so forth. Once all members of the group have sequentially sought for the highest point from the last point of their predecessor, the baton is passed again to the first agent of the group.

A clear alternative to ‘relay’ dynamics is a ‘tournament’ in which all agents of a group simultaneously strive to identify the point on the landscape that earns them the highest value. In tournament dynamics, the point with the highest value that any agent could identify in the first round then becomes the starting point for everyone in the next go-round. What is eliminated in tournament dynamics is the around-the-table sequencing of a relay. Hong and Page consider both dynamics, saying that their result “do not seem to depend on which structure was assumed” (2004, 16386). Yet our results do depend on which dynamic is used. Just as a larger heuristic pool favor diverse groups, so does the use of tournament over relay dynamics. In comparison with Figures 5 through 7, Figures 8 through 10 show results for tournament dynamics in place of relay.

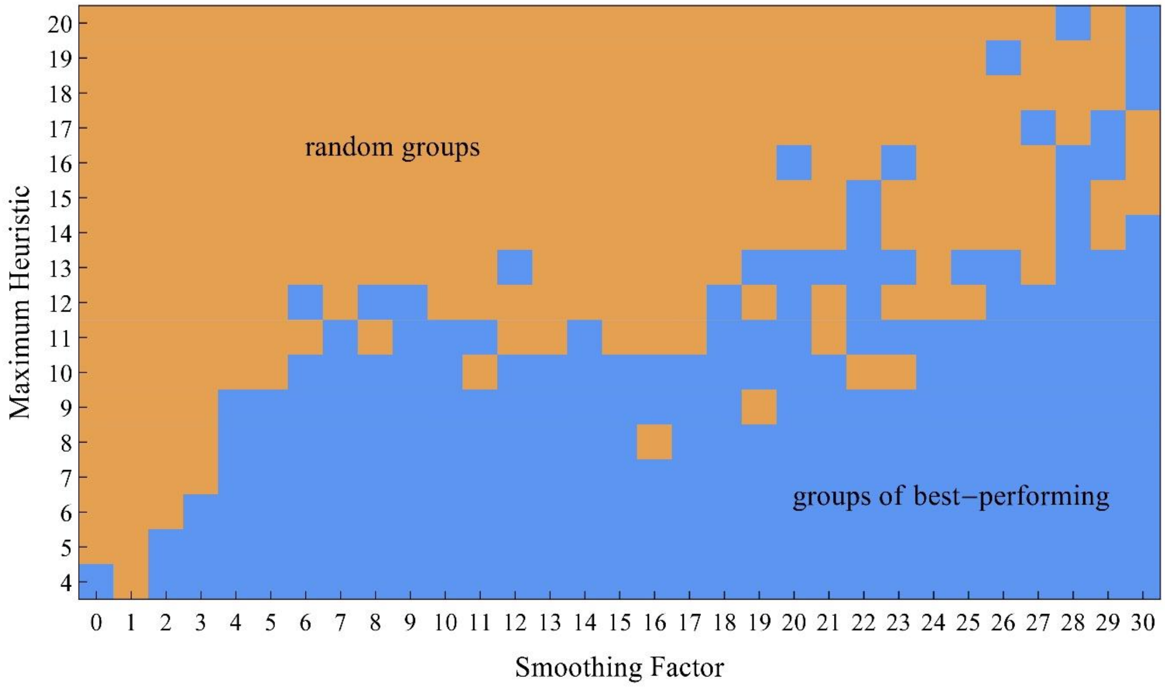

Figure 8. Tournament results corresponding to Figure 5's relay results showing parameter combinations in which groups of random (orange) and groups of best-performing (blue) do best.

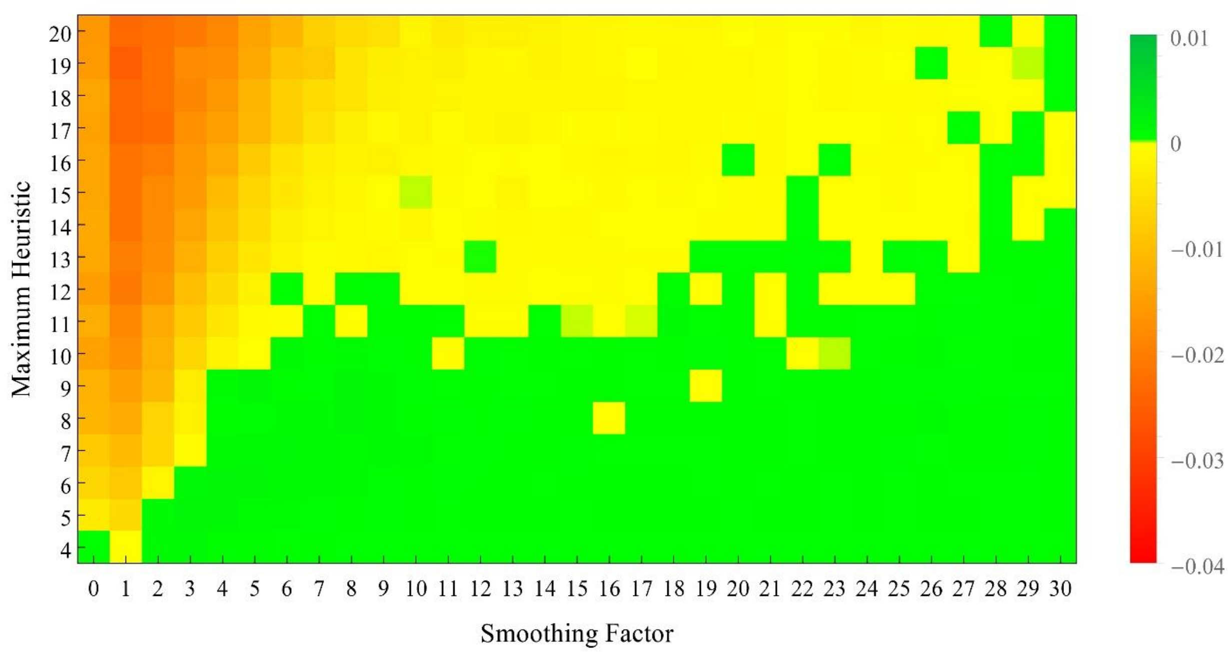

Figure 9. Tournament results corresponding to Figure 6's relay results showing differences in averages over 100 landscapes, with positive values (green) showing advantage to the best-performing and negative values (yellow to red) for random groups.

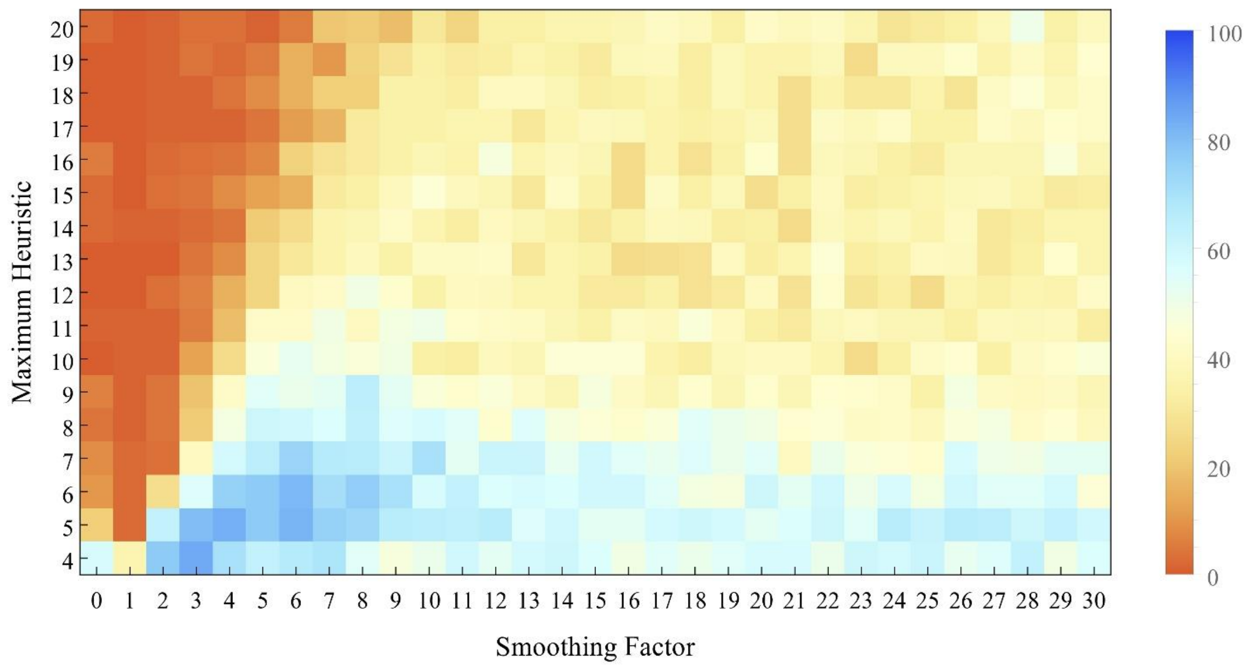

Figure 10. Tournament results corresponding to Figure 7's relay results showing percentages of runs in which each group does better, with blue values reflecting more wins by best-performing and yellow to red for groups using random heuristics.

For a maximum heuristic over 10, 'tournament' rather than 'relay' updating gives a strong advantage to random groups. Indeed, the tournament group dynamics dramatically reduces the advantage possessed by groups of the best-performing on smoother landscapes with the relay dynamic. Again here, these are areas in which one can more reasonably interpret best-performance as ability or expertise.[7] So group dynamics in the form of a tournament rather than a relay makes an important difference in the relative value of diversity and groups of the best-performing.

[7] As documented in auxiliary materials online, we have tested the robustness of these results with groups of sizes 3, 6, and 9 (bit.ly/DAESupp). The smaller the group, for tournament as well as relay dynamics, the greater the advantage for groups of the best-performing. The larger the group, all things considered, the greater the advantage for groups of random heuristics.

In additional work (available as an online supplement: bit.ly/DAESupp), we explore other variations of the Hong-Page model. As opposed to groups composed exclusively of the best-performing agents or random agents, we look at the performance of mixed groups, consisting of both experts and randomly selected agents. What we see there is that, in many respects, mixed groups do better than either kind of pure group, but their performance is importantly affected by the group dynamics. We also explore the performance of groups of different sizes. Smaller groups are more advantaged by having best-performers, we find, but for larger groups, random agents do best. Like above, this is naturally explained by the amount of heuristic coverage offered by the agents in the group, we argue.

VI. Conclusion

Our results indicate that the slogan 'diversity trumps ability' can easily be overstated. By exploring minimal variations of the original Hong-Page simulation, we have shown that the DTA result of random groups outperforming groups of the best-performing holds only within a small window of low landscape smoothness. Within roughly that same window, moreover, the success of the best-performing heuristics on a specific landscape is limited to that specific landscape: success on one random landscape cannot be expected to yield success on another. As we saw above, 'diversity trumps ability' only in those cases where it is unclear that best-performance should really be considered ability or expertise.

On smoother landscapes, there is a connection between performance on one landscape and another. Interpreting a set of landscapes as specific questions within a given domain, successful heuristics have a better claim to be modeling ability or expertise, and it's here that expertise shows its value. For landscape smoothness above 4, using the Hong-Page relay dynamics, groups of the individually best-performing agents outperform groups of random agents. With an increase in landscape smoothness, leaving other parameters in place, it is ability that trumps diversity.

Diversity again shows its strength, however, when other parameters are changed. Widening the pool from which heuristic numbers are drawn increases the advantage for random groups. Given a landscape smoothness factor at which groups of the best-performing do better with a given set of available heuristics, groups of random agents perform better once we increase the conceptual space to a larger heuristic pool.

Contrary to Hong and Page's indication of little difference between the relay dynamics used in their simulation and an alternative 'tournament' dynamics, we find a major difference between the two. In 'tournament' dynamics, agents deliberate and navigate a problem landscape with simultaneous suggestions from the floor rather than in a round-the-table 'relay.' It turns out that a tournament dynamic further favors the value of diversity. Many of the points at which groups of the best-performing show an advantage within a relay dynamics disappear in favor of groups of random heuristics once the dynamic is changed to a simultaneous tournament.

The variety and sensitivity of these results shows that uncritical applications of the Page-Hong result are risky. As mentioned above, the Hong-Page model has been presented to NASA, cited by the USGS, and has played a role in U.S. Supreme Court reasoning. The results have been used as a critique of expertise as part of the epistemic argument for democracy (Anderson 2006; Landemore 2013, Gunn 2014, Weymark 2015). In both theory and application, the slogan 'diversity trumps ability' has been used as a general claim in support of diversity initiatives in science (Nunn 2012; Stegenga 2016; Cheryan, Plaut, Handron & Hudson 2013; Chesler, Barabino, Bhatia & Richards-Kortum 2010; Gibbs, McGready, Bennett & Griffin 2014; Pickett, Corb, Matthews, Sundquist & Berg 2015; Quissel & Walt 2015). What our results indicate is that diversity does *not* always trump ability.

Policy makers across the board must consider the specific character of the problem sets at issue and the decision procedures to be employed. Here as elsewhere, moving from formal results to real world applications is a long, laborious, and, most importantly, empirical process (Alexandrova & Northcott, 2009). Policy makers wishing to assess whether a particular issue might benefit from a more diverse community have a significant amount of additional bridgework to engage in before they can derive support for policies from these modeling results (for details see Grim, et al., 2013). Our results show that diversity is not always epistemically beneficial, even at the model level. Diverse groups and groups of the individually best-performing both have a place. Our expanded model of group inquiry reveals a nuanced interplay between them and points towards a greater understanding of the strengths of each.

References

Alexander, Jason. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, *82*(3), 424-453.

Alexandrova, A. and Northcott R. (2009). "Progress in Economics: Lessons from The Spectrum Auctions" in H. Kincaid and D. Ross (eds.), Oxford Handbook of Philosophy of Economics (Oxford University Press), 306-336.

Anderson, Elizabeth (2006). The Epistemology of Democracy. *Episteme* 3.1, 8-22.

Bright, Liam. K. (2016). Decision Theoretic Model of the Productivity Gap. *Erkenntnis*, 1-22.

Cheryan, S., Plaut, V. C., Handron, C., & Hudson, L. (2013). The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex Roles*, *69*(1-2), 58-71.

Chesler, N. C., Barabino, G., Bhatia, S. N., & Richards-Kortum, R. (2010). The pipeline still leaks and more than you think: a status report on gender diversity in biomedical engineering. *Annals of Biomedical Engineering*, *38*(5), 1928-1935.

Elga, A. (2007). Reflection and disagreement. *Noûs*, *41*(3), 478-502.

Ferrini-Mundy, J. (2013). Driven by diversity. *Science*, *340*(6130), 278-278.

Fisher v. University of Texas, Austin, 2016 (no. 14-981) Brief for Lt. Gen. Julius W. Becton, Jr., Gen. John P. Abiziad, Adm. Dennis C. Blair, Gen. Bryan Doug Brown, Gen. George W. Casey, Lt. Gen Daniel W. Christman, Gen. Wesley K. Clark, Adm. Archie Clemins, Gen. Ann E. Dunwoody, Gen. Ronald R. Fogleman, Adm. Edmund P. Giambastiani, Jr., et al., as Amici Curiae in Support of respondents. .

Fontanari, J. F., & Rodrigues, F. A. (2016). Influence of network topology on cooperative problem-solving systems. *Theory in Biosciences*, *135*(3), 101-110.

Gaifman, Haim (1988). 'A Theory of Higher Order Probabilities', in Brian Skyrms and William L. Harper (eds.), Causation, Chance, and Credence, Kluwer Academic Publishers, pp. 191–219.

Gibbs Jr, K. D., McGready, J., Bennett, J. C., & Griffin, K. (2014). Biomedical science Ph. D. career interest patterns by race/ethnicity and gender. *PloS One*, *9*(12), e114736.

Grim, P. (2009). Threshold Phenomena in Epistemic Networks. In AAAI Fall Symposium: Complex Adaptive Systems and the Threshold Effect (pp. 53-60).

Grim, P., Rosenberger, R., Rosenfeld, A., Anderson, B., & Eason, R. E. (2013). How simulations fail. *Synthese*, 190(12), 2367-2390.

Gunn, Paul (2014) “Democracy and Epistocracy.” *Critical Review* 26:1-2, 59-79.

Hong, Lu & Scott E. Page (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences. 101*, 16385-16389.

Hong, Lu & Scott E. Page (2009). Interpreted and Generated Signals. *Journal of Economic Theory* 144, 2174-2196.

Landemore, Helen (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rules of the Many*. Princeton, NJ: Princeton University Press

Martini, Carlo (2014). Experts in science: A view from the trenches. *Synthese*, *191*(1), 3-15.

Mayo-Wilson, Connor, Zollman, Kevin J., & Danks, David. (2011). The independence thesis: when individual and social epistemology diverge. *Philosophy of Science*, *78*(4), 653-677.

Page, Scott E. (2007) *The Difference*. Princeton University Press.

Page, Scott E. (2011) *Diversity and Complexity*. Princeton University Press.

Pickett, C. L., Corb, B. W., Matthews, C. R., Sundquist, W. I., & Berg, J. M. (2015). Toward a sustainable biomedical research enterprise: Finding consensus and implementing recommendations. *Proceedings of the National Academy of Sciences*, *112*(35), 10832-10836.

Quissell, K., & Walt, G. (2015). The challenge of sustaining effectiveness over time: the case of the global network to stop tuberculosis. *Health Policy and Planning*, czv035.

Stegenga, Jacob (2016). Three Criteria for Consensus Conferences. *Foundations of Science*, *21* (1), 35-49.

Thoma, Johanna (2015). The Epistemic Division of Labor Revisited. *Philosophy of Science*, *82*(3), 454-472.

Thompson, Abigail (2014). Does diversity trump ability?: An example of the misuse of mathematics in the social sciences. *Notices of the American Mathematical Society*, 61, 1024-1030.

UCLA (2014). Proposed Diversity Requirement. Retrieved from https://ccle.ucla.edu/pluginfile.php/743624/mod_resource/content/6/082014%20REVISED%20DiversityReqProposal.pdf

Weymark, J. A. (2015). Cognitive Diversity, Binary Decisions, and Epistemic Democracy. *Episteme*, *12*(04), 497-511.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74 (5), 574-587.

Zollman, K. (2011). Computer simulation and emergent reliability in science. Journal of Artificial Societies and Social Simulation, 14(4), 15.

Zollman, K. J. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass*, *8*(1), 15-27.