Revisiting the effect of red on competition in humans (supplementary information)

Laura Fortunato[1,2] and Aaron Clauset[3,4,2]

[1] *Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford OX2 6PN, UK*
[2] *Santa Fe Institute, Santa Fe, NM 87501, USA*
[3] *Department of Computer Science, University of Colorado, Boulder, CO 80309, USA*
[4] *BioFrontiers Institute, University of Colorado, Boulder, CO 80303, USA*

Contents

S1 Outline

We begin with detailed descriptions of relevant aspects of the competitions, for the four sports analyzed by Hill & Barton [1], at the 2004 Athens and 2008 Beijing Olympics (Section S2). Our aim in this section is three-fold. First, we aim to provide the reader with a mechanistic understanding of the competition structure for the different sports, focusing on specific features leading to biases towards wins by one color in the outcomes of the competitions. Second, we highlight the equivalence between the 2008 data presented here and the 2004 data analysed by Hill & Barton [1]. Third, we outline changes in the competition structure introduced at the 2012 London Olympics in two of the four sports, which prevent extension of the analysis to this competition.

In Section S3 we provide information on data acquisition and on processing of the data for analysis, together with descriptives for each sport in the 2004 and 2008 competitions.

We then turn to analysis of the data (Section S4). We begin by replicating Hill & Barton's [1] exact analytical approach, in the process uncovering several shortcomings. We re-derive the results underpinning Hill & Barton's [1] main claim of a red effect in the 2004 data and show that they are not robust. We then extend the exact analytical approach to the 2008 data and show that the results do not hold in this case. Next, we propose an alternative analytical approach, which addresses key shortcomings with Hill & Barton's [1] analysis and allows us to quantify the magnitude of any effect that may exist in the data.

Finally, in Section S5 we use Monte Carlo simulation to demonstrate the existence of a structural bias towards wins by red in the outcomes of the 2004 competition, which can explain away the pattern reported by Hill & Barton [1]. This is confirmed by evidence of a structural bias towards wins by blue in the outcomes of the 2008 competition, consistent with the pattern observed in the data. The simulation results indicate that incompleteness in the tournament structures, coupled with variance in skill among the contestants, can induce a bias that shifts the null distribution towards wins by one color.

We conclude with a summary of the insights produced by our multiple lines of analysis and new data (Section S6). In particular, we discuss their implications for the hypothesis of an effect of red on human competition, and on human behavior more generally, in the context of the substantial body of work that has developed over the past decade, building on Hill & Barton's [1] influential study.

S2 Structure of Olympic sports

In the four sports analysed by Hill & Barton [1] (male divisions only), the competition for a given weight class is arranged as a single-elimination tournament (also known as knock-out or sudden death). As illustrated in Fig. 1**b** in the main text, contestants compete in pairs. The winner of a contest, or bout, proceeds to the following round in the tournament. A contestant's placement (top vs. bottom) in a given bout determines the color he wears in that bout. His relative position may change between bouts, as he progresses through rounds in the tournament.

Two possible sources of incompleteness in a single-elimination tournament are byes and walkovers, also illustrated in Fig. 1**b** in the main text. Byes are used if there are fewer than the number of contestants required to "fill" the outermost round in a competition "tree"; in this case, one or more contestants are byed to the following round. A walkover involves a contestant winning the bout by default, because his opponent forfeited the contest (e.g., by withdrawing or by failing to show up).

S2.1 Boxing

The male division at the 2004 and 2008 Olympics included 11 weight classes.

S2.1.1 Tournament structure

For each weight class, the competition is arranged as a single-elimination tournament. If the number of contestants in the weight class is not a power of 2, then there will be byes in the first round of bouts (preliminary round). Contestants not byed compete in this round so that the number in the following round is reduced to a power of 2. The number of byes is the difference between the initial number of contestants and the next higher power of 2.

In the 2004 and 2008 competitions, the number of contestants n in the different weight classes ranged from 16 to 29. In each of the weight classes with more than 16 contestants, $32 - n$ received a bye to the round of 16 (eighth-finals). The other contestants competed in $n - 16$ bouts in the preliminary round, with the winners of the bouts proceeding to the round of 16.

In the 2004 and 2008 competitions, byes were "stacked" at the top of the preliminary round for each weight class. The initial placement of contestants on the tree ("seeding") was drawn by manual lot, and thus at random; byes were determined through this draw, and thus also at random (Official Report of the XXVIII Olympiad 2: the Games, pag. 277; Sébastien Gillot, pers. comm. Nov. 2013; Janusz Majcher, pers. comm. Nov. 2013).

Starting with the 2012 London Olympics, the draw was seeded based on the International Boxing Association (AIBA) ranking and on performance in the World Series of Boxing (WSB) (Sébastien Gillot, pers. comm. Nov. 2013; Janusz Majcher, pers. comm. Nov. 2013), with byes and seeded entries evenly distributed across the tree (see e.g., Appendix E of the AIBA Technical & Competition Rules effective from March 24, 2011). This procedure is used to ensure even strength throughout the competition draw, for example to avoid two top-ranked boxers meeting in an early round, resulting in one of them being eliminated prematurely.

S2.1.2 Color assignment

In each bout, the contestant in the top position of the bracket wears red, the one in the bottom position wears blue. Colors are assigned following this procedure, initially based on the position of the contestants in the outermost round (i.e., the preliminary round); they are re-assigned accordingly as contestants proceed from one round to the next.

In the 2004 and 2008 competitions, contestants wore blue or red uniforms (vest and shorts) with matching equipment (headguard, gloves).

S2.1.3 Placement of winners

The winner of the bout in the final round gets first place (gold), the loser second place (silver). The losers of the two bouts in the semi-finals share third place (bronze). The losers of the four bouts in the quarter-finals share fifth place.

S2.2 Taekwondo

The male division at the 2004 and 2008 Olympics included four weight classes.

S2.2.1 Tournament structure

For each weight class, the competition is arranged as single-elimination tournament with 16 contestants, thus the first round is the round of 16 (eight-finals). There were 16 contestants in the female and male divisions of the 2004 and 2008 competitions, hence no byes. For comparison, there were 15 contestants per weight class in the female division of 2004; the one who picked number 1 in the draw was byed and proceeded to the quarter-finals without a match in the round of 16 (Jeongkang Seo, pers. comm. Oct.–Nov. 2013).

In the 2004 and 2008 competitions, the initial placement of contestants on the tree was drawn by lot, hence the pairing of contestants was by random selection, without consideration of skill. The World Taekwondo Federation (WTF) introduced seeding based on world rankings starting with the 2012 London Olympics (Jeongkang Seo, pers. comm. Oct.–Nov. 2013).

S2.2.2 Color assignment

In each bout, the contestant in the top position of the bracket wears blue, the one in the bottom position wears red. Colors are assigned following this procedure, initially based on the position of the contestants in the outermost round in a competition tree; they are re-assigned accordingly as contestants proceed from one round to the next.

In the 2004 and 2008 competitions, contestants wore a white uniform (called “dobok”, consisting of long-sleeved top, pants), with blue or red equipment (trunk and head protectors).

S2.2.3 Placement of winners

The winner of the bout in the final round gets first place (gold), the loser second place (silver).

Repechage is used to determine the ranking of other contestants, with different types used in 2004 and 2008.

Repechage (2004 Athens Olympics) The upper half of the main competition tree is “Pool A”, the lower half “Pool B”. There is one repechage contest with six contestants, arranged in the following configuration (“b”: blue; “r”: red):

```
1 (b) _____
           |__
2 (b) __   |  |
        |__|  |
3 (r) __|     |
              |__ bronze
4 (b) __      |
        |_    |
5 (r) __| |   |
          |___|
6 (r) ____|
```

1 : loser in semi-final round for Pool A
2 : loser in quarter-final round against finalist for Pool B
3 : loser in eighth-final round against finalist for Pool B

4 : loser in eighth-final round against finalist for Pool A
5 : loser in quarter-final round against finalist for Pool A
6 : loser in semi-final round for Pool B

The winner of the bout in the bronze round is awarded third place (bronze).

Note that the repechage contest is “symmetric” in the allocation of color, in the sense that losers of bouts in the semi-final round wear blue in the top branch (1) and red in the bottom one (6); losers of bouts in the quarter-final round wear blue in the top one (2) and red in the bottom one (5); losers of bouts in the eighth-final round wear red in the top one (3) and blue in the bottom one (4).

Repechage (2008 Beijing Olympics) The upper half of the main competition tree is "Pool A", the lower half "Pool B". There are two separate repechage contests, each including three contestants, arranged in the following configuration ("b": blue; "r": red):

```
1 (b) _____
           |__ bronze
2 (b) __   |
        |__|
3 (r) __|

4 (b) __
        |_
5 (r) __| |
          |__ bronze
6 (r) ____|
```

1 : loser in semi-final round for Pool A
2 : loser in quarter-final round against finalist for Pool B
3 : loser in eighth-final round against finalist for Pool B

4 : loser in eighth-final round against finalist for Pool A
5 : loser in quarter-final round against finalist for Pool A
6 : loser in semi-final round for Pool B

The winners of bouts in the bronze rounds share third place (bronze), the losers get fifth place, and the other two contestants get seventh place.

Note that the two repechage contests are "symmetric" in the allocation of color, in the sense that losers of bouts in the semi-final round wear blue in the top one (1) and red in the bottom one (6); losers of bouts in the quarter-final round wear blue in the top one (2) and red in the bottom one (5); losers of bouts in the eighth-final round wear red in the top one (3) and blue in the bottom one (4).

S2.3 Wrestling

The male division at the 2004 and 2008 Olympics included seven weight classes each for Greco-Roman wrestling and free-style wrestling. The tournament structure changed substantially between 2004 and 2008.

S2.3.1 Tournament structure (2004 Athens Olympics)

For each weight class, contestants initially compete in a series of randomly determined elimination pools, with each contestant competing against all others in the pool. The contestant with the greatest number of technical points in each pool proceeds to the qualification round; classification points are used to break ties (Tony Black, pers. comm. July 2013).

Elimination pools include three or four contestants, with pools of four placed at the bottom of the list. Contestants in a pool of three compete in two rounds, those in a pool of four compete in three rounds. Because one contestant per pool proceeds to the qualification round, if the number of pools is not a power of 2, then there will be byes in the qualification round. Contestants not byed compete in this round so that the number in the following round (semi-finals) can be reduced to four. The number of byes is the difference between the number of elimination pools and the next higher power of 2.

In the 2004 competition, the number of contestants in the elimination pools ranged across weight classes from 19 to 22, and they were arranged in one of the following configurations:

19 : 6 elimination pools (5 pools of 3, 1 pool of 4)
20 : 6 elimination pools (4 pools of 3, 2 pools of 4)
21 : 7 elimination pools (7 pools of 3)
22 : 7 elimination pools (6 pools of 3, 1 pool of 4)

In the qualification round six contestants were arranged in two bouts with two byes, seven contestants in three bouts with one bye, irrespective of the specific configuration of the elimination pools. Byes were "stacked" at the bottom of the qualification round for each weight class. Because the initial placement of contestants into pools was based on a random draw (Tony Black, pers. comm. July 2013), byes were effectively determined through this draw, and thus also at random. Because pools of four were placed at the bottom of the list, and byes were drawn from the bottom, contestants from these pools were always byed to the semi-finals; contestants from pools of three may also be byed, depending on the specific configuration.

The winners of bouts in the qualification round and byed contestants proceed to the semi-finals, the losers to the 5–6 final (or the two losers with the most points, in configurations with three bouts in the qualification round). The winners of the two bouts in the semi-finals proceed to the 1–2 final, the losers to the 3–4 final; see below for placement of winners.

S2.3.2 Tournament structure (2008 Beijing Olympics)

For each weight class, the competition is arranged as a single-elimination tournament. If the number of contestants in the weight class is not a power of 2, then there will be byes in the first round of bouts (qualification round). Contestants not byed compete in this round so that the number in the following round is reduced to a power of 2. The number of byes is the difference between the initial number of contestants and the next higher power of 2.

In the 2008 competition, the number of contestants n in the different weight classes ranged from 19 to 21. In each weight class, $32 - n$ received a bye to the round of 16 (eighth-finals). The other contestants competed in $n - 16$ bouts in the qualification round, with the winners of the bouts proceeding to the round of 16.

Byes were "stacked" at the top of the qualification round for each weight class. The initial placement of contestants on the tree, and thus their pairing, was drawn at random; byes were determined through this draw, and thus also at random [see e.g., Articles 8, 12, 14 of the FILA International Wrestling Rules, release Dec. 2006; confirmed in two later versions (updated Feb. 2010 and July 2014), which implies that this approach was used in 2008].

S2.3.3 Color assignment

In each bout, the contestant at the top of the bracket wears red, the one at the bottom wears blue. Colors are assigned following this procedure, initially based on the position of the contestants in the outermost round in a competition tree; they are re-assigned accordingly as contestants proceed from one round to the next.

In the 2004 competition, the same assignment procedure applied to bouts in the elimination pools, but the ordering of contestants in a bout based on draw number "switched" between the three rounds of each pool. Specifically, the contestant with the lower draw number in the pair competed in the top position in Round 1, in the bottom position in Round 2, and in the top position in Round 3.

In the 2004 and 2008 competitions, contestants wore a blue or red one-piece singlet.

S2.3.4 Placement of winners

The winner of the bout in the final round gets first place (gold), the loser second place (silver).

In the 2004 competition, the losers of the two bouts in the semi-finals compete for third place (bronze) in the 3–4 final, with the loser of this bout placed fourth. Two losers of the qualification round compete for fifth and sixth place in the 5–6 final, with any other losers from the qualification round placed seventh, and further placements based on the number of classification points scored in the elimination pools. Note that 5–6 finalists often chose not to compete as they could not medal; in these cases, the bout was won by walkover (Tony Black, pers. comm. July 2013).

In the 2008 competition, repechage is used to determine the ranking of other contestants.

Repechage (2008 Beijing Olympics) The upper half of the tree is the upper branch, the lower half is the lower branch. There are two separate repechage contests, each contest including up to four contestants (depending on the number of contestants in the qualification round and their placement in this round), arranged in the following configuration ("b": blue; "r": red):

```
1 (r) __
        |__
2 (b) __| |
          |__
  3 (b) __| |
            |__ bronze
    4 (b) __|

5 (r) __
        |__
6 (b) __| |
          |__
  7 (b) __| |
            |__ bronze
    8 (b) __|
```

1 : loser in qualification round against finalist for upper branch
2 : loser in eighth-final round against finalist for upper branch
3 : loser in quarter-final round against finalist for upper branch
4 : loser in semi-final round against finalist for upper branch

5 : loser in qualification round against finalist for lower branch
6 : loser in eighth-final round against finalist for lower branch
7 : loser in quarter-final round against finalist for lower branch
8 : loser in semi-final round against finalist for lower branch

The winners of bouts in the bronze rounds share third place (bronze), the losers get fifth place. Further placements (seventh onwards) are based on the number of classification points scored throughout the tournament (Tony Black, pers. comm. July 2013).

In the 2008 competition, no contestants in the upper branch competed in qualification rounds. Consequently, there was no contestant 1, and contestant 2 was byed to the second round of the repechage contest. If the finalists for the lower branch did not compete in the qualification round, then there was no contestant 5, and contestant 6 was also byed to the second round of the repechage contest.

Note that in each bout of the repechage contests, the contestant wearing red had been eliminated "earlier" in the competition than the contestant wearing blue. If there is a link between the round in which a contestant was eliminated and his skill, then this may potentially introduce a bias towards wins by blue. For example, in the outermost round of the repechage contest for the upper branch, the red-wearing contestant (1) had been eliminated in the qualification round, the blue-wearing contestant (2) in the eighth-final round. The winner of this bout wears red in the next bout; the blue-wearing contestant in this bout had reached the quarter-final round (3). The winner of this bout wears red in the next bout; the blue-wearing contestant in this bout had reached the semi-final round (4). We note that there is no evidence of such a bias in the data (Greco-Roman wrestling: Section S3.2.3; free-style wrestling: Section S3.2.4), possibly due to the small number of rounds in each tournament that are potentially affected.

S2.4 Summary

Competitions in the four sports analysed by Hill & Barton [1] are arranged as a single-elimination tournament for each weight class. Generally, contestants compete in pairs, with the winner of a bout proceeding to the next round in the tournament. Details of the tournament structure and related aspects vary, however — both across sports and, with the exception of boxing, within sports between the 2004 and 2008 competitions. Here we outline the implications of this variation for extension of Hill & Barton's [1] approach from the 2004 to the 2008 data.

The structural changes within taekwondo affect only the repechage rounds. As discussed in Section S2.2, the structure of the repechage contests is "symmetric" in the allocation of color for both 2004 and 2008. Consequently, these changes do not invalidate extension of Hill & Barton's [1] approach to the 2008 data.

The structural changes in wrestling are more substantial, affecting the overall tournament structure (Section S2.3). As we discuss below, this provides a candidate mechanism explaining the different patterns observed in 2004 vs. 2008 (namely, a shift in position of the byes from the bottom of the relevant round in 2004 to the top in 2008; Section S5). Even these changes do not, in themselves, invalidate extension of Hill & Barton's [1] approach to the 2008 data. In fact, because Hill & Barton [1] excluded bouts from the elimination pool rounds from analysis of the 2004 data (Section S2.3.1), and elimination pool rounds do not feature in the 2008 competition, the 2008 dataset more than trebles the number of bouts available for analysis in both Greco-Roman wrestling and free-style wrestling (Sections S3.1 and S3.2).

In light of these changes, we determined which bouts to exclude as walkovers from the wrestling 2008 data by comparing possible and realized bout outcomes for the two competitions, as follows. This ensures consistency with the exclusion criteria implemented by Hill & Barton [1] for the 2004 wrestling data.

For the 2004 competition, possible bout outcomes are:

- EF Victory by forfeit, the loser is not classified
- EV Disqualification from all competition for violation of the rules
- EX 3 cautions or violation of the rules
- E2 Both wrestlers are disqualified for violation of the rules
- PA Injury default
- PO Victory by points, the loser without technical points
- PP Victory by points, the loser with technical points
- SP Technical superiority, 10 points difference, the loser with points
- ST Technical superiority, 10 points difference, the loser without points
- TO Victory by fall

All types except EF, EX occur in the Greco-Roman wrestling data analysed by Hill & Barton [1] (i.e., excluding elimination pool rounds). Similarly, all types except EF, EX, E2, SP occur in the free-style wrestling data analysed (i.e., also excluding elimination pool rounds). In both cases, the bouts coded as won by walkover by Hill & Barton [1] are of type EV or PA.

For the 2008 competition, possible bout outcomes are:

- E2 Both wrestlers have been disqualified due to infringement of the rules
- EX 3 cautions '0' due to error against the rules
- PP Decision by points, the loser with technical points
- ST Great superiority, a difference of 6 points, the loser without points
- VB Victory by injury
- VT Victory by fall
- EV Disqualification from the whole competition due to infringement of the rules
- PO Decision by points, the loser without technical point
- SP Victory by technical superiority with the loser scoring technical points
- VA Victory by withdrawal
- VF Victory by forfeit

The only types that occur in the data are VT, ST, PP, PO for Greco-Roman wrestling, and VT, VA, ST, SP, PP, PO for free-style wrestling. Of these, VT, ST, SP, PP, PO can be directly matched to types TO, ST, SP, PP, PO for the 2004 competition. Given that all of these occurred in the 2004 data analysed by Hill & Barton [1] and were retained for analysis, we retained them in the 2008 data. Type VA cannot be directly matched to any type in the 2004 data. This corresponds to a walkover and was therefore excluded from the 2008 data.

Overall, through careful matching of the 2004 and 2008 data, the two datasets are fully equivalent for the purpose of testing Hill & Barton's [1] hypothesis. Crucially, in all sports in both competitions, initial seeding of the competition tree was based on a random draw. Seeding based on skill — a factor proposed as a potential source of bias towards wins by one color [2, 3] — does not apply to these data, and it can therefore be ruled out as explaining for any observed bias. Furthermore, the introduction of seeding based on skill in boxing (Section S2.1.1) and in taekwondo (Section S2.2.1) at the 2012 London Olympics prevents extension of Hill & Barton's [1] approach to this competition.

S3 Olympic data

S3.1 Data for 2004 Athens Olympics

The data were obtained from the supplementary information of Hill & Barton [1] (file `435293a-s1.xls`) and exported to `csv` format.

S3.1.1 Data for boxing

The data include all rounds. We excluded entries marked `Walk Over` in column `Method of Win` (5 total; 3 resulting in a win by red, 2 resulting in a win by blue).

The total number of entries for analysis is $n = 267$, with $n_{\text{red}} = 147$ resulting in a win by red, $n_{\text{blue}} = 120$ resulting in a win by blue.

S3.1.2 Data for taekwondo

The data include all rounds. We excluded entries marked `Withdrawn` in column `Method of Win` (5 total; 2 resulting in a win by red, 3 resulting in a win by blue).

The total number of entries for analysis is $n = 75$, with $n_{\text{red}} = 43$ resulting in a win by red, $n_{\text{blue}} = 32$ resulting in a win by blue.

S3.1.3 Data for Greco-Roman wrestling

The data include only the qualification round, semi-finals, 1–2 final, 3–4 final, and 5–6 final (i.e., bouts from the elimination pools were not included; Section S2.3.1). We excluded entries marked `Yes` in column `Won by Walkover` (3 total; 0 resulting in a win by red, 3 resulting in a win by blue).

The total number of entries for analysis is $n = 48$, with $n_{\text{red}} = 25$ resulting in a win by red, $n_{\text{blue}} = 23$ resulting in a win by blue.

S3.1.4 Data for free-style wrestling

The data include only the qualification round, semi-finals, 1–2 final, 3–4 final, and 5–6 final (i.e., bouts from the elimination pools were not included; Section S2.3.1). We excluded entries marked `Yes` in column `Won by Walkover` (3 total; 1 resulting in a win by red, 2 resulting in a win by blue).

The total number of entries for analysis is $n = 51$, with $n_{\text{red}} = 27$ resulting in a win by red, $n_{\text{blue}} = 24$ resulting in a win by blue.

S3.2 Data for 2008 Beijing Olympics

Results books were obtained from the archive at `http://library.la84.org/6oic/OfficialReports/2008/` (boxing: file `2008Results_Book1.pdf`; taekwondo, wrestling: file `2008Results_Book2.pdf`).

The relevant data were entered manually and exported to `csv` format.

S3.2.1 Data for boxing

The data include all rounds. We excluded entries marked `WO` (walkover) in column `method` (2 total; 0 resulting in a win by red, 2 resulting in a win by blue).

The total number of entries for analysis is $n = 270$, with $n_{\text{red}} = 133$ resulting in a win by red, $n_{\text{blue}} = 137$ resulting in a win by blue.

S3.2.2 Data for taekwondo

The data include all rounds. We excluded entries marked `WDR` (withdrawn) in column `method` (1 total; 0 resulting in a win by red, 1 resulting in a win by blue).

The total number of entries for analysis is $n = 75$, with $n_{\text{red}} = 38$ resulting in a win by red, $n_{\text{blue}} = 37$ resulting in a win by blue.

S3.2.3 Data for Greco-Roman wrestling

The data include all rounds. No entries were excluded, because none corresponding to walkovers are represented in the data (Section S2.4).

The total number of entries for analysis is $n = 164$, with $n_{\text{red}} = 80$ resulting in a win by red, $n_{\text{blue}} = 84$ resulting in a win by blue.

Of the 32 entries corresponding to repechage and bronze rounds, 17 ended in a win by red, 15 in a win by blue (one-sided binomial test, $H_0 : f_{\text{blue}} \leq 0.5; H_{\text{A}} : f_{\text{blue}} > 0.5$, $p = 0.702$). There is thus no evidence of a bias towards wins by blue in the repechage contests (Section S2.3.4).

S3.2.4 Data for free-style wrestling

The data include all rounds. We excluded entries marked `VA` (victory by withdrawal; Section S2.4) in column `method` (1 total; 1 resulting in a win by red, 0 resulting in a win by blue).

The total number of entries for analysis is $n = 164$, with $n_{\text{red}} = 67$ resulting in a win by red, $n_{\text{blue}} = 97$ resulting in a win by blue.

Of the 33 entries corresponding to repechage and bronze rounds, 13 ended in a win by red, 20 in a win by blue (one-sided binomial test, $H_0 : f_{\text{blue}} \leq 0.5; H_{\text{A}} : f_{\text{blue}} > 0.5$, $p = 0.148$). There is thus no evidence of a bias towards wins by blue in the repechage contests (Section S2.3.4).

S4 Data analysis

S4.1 Hill & Barton's [1] analytical approach

The main claim of a red effect in Hill & Barton's [1] analysis is in two parts. First, in each of the four sports, over 50% of the bouts resulted in a red win, with a fraction that was statistically significantly different from 0.5 in the data pooled over the four sports. Second, further pooling of the data across sports by round or by weight class revealed a consistent pattern: in each case, the fraction including over 50% of red wins was statistically significantly different from 0.5. Specifically, the four sports combined comprised 21 rounds: of these, 16 presented a majority of red wins, and only four a majority of blue wins. Similarly, of the 29 weight classes across the four sports, 19 presented a majority of red wins, and only six a majority of blue wins.

As a corollary of the main claim, a final analysis divided the pooled data into four groups, based on the competitive ability of the contestants (as judged by the number of points scored). In each of the three most symmetric groups, over 50% of the bouts were won by red, with a fraction statistically significantly different from 0.5 only in the most symmetric group; in the least symmetric group, over 50% of bouts were won by blue.

We begin by replicating the analysis underpinning Hill & Barton's [1] main claim, applied to both the 2004 and the 2008 data, in the process uncovering several shortcomings (Section S4.1.1). Next, we outline a series of shortcomings that apply to the skill-based analysis (Section S4.1.2).

S4.1.1 Main claim

Table S1 summarizes χ^2 test results for the 2004 data, following the analytical approach used by Hill & Barton [1]. In all cases, the fraction of red wins $f_{\text{red}} > 0.5$, but the only significant result at the $\alpha = 0.05$ level is for the data aggregated over the four sports. Even this result is not robust, however: just one additional blue win would tip the p-value over the significance threshold. In any case, as each hypothesis test is carried out in parallel with additional tests on (sub-sets of) the same data, the analysis falls squarely within the setting of multiple hypothesis testing. Therefore, a threshold of $\alpha = 0.05$ overstates the likely true statistical significance (discussed in Section S4.2).

Furthermore, we note that the χ^2 test rests on an asymptotic approximation that only holds for large sample sizes. The binomial test is exact and correct for all sample sizes, and in its one-sided form it provides a more accurate representation of Hill & Barton's [1] hypothesis (i.e., $H_{\text{A}} : f_{\text{red}} > 0.5$ vs. $H_{\text{A}} : f_{\text{red}} \neq 0.5$). A final issue is that the data present structural dependencies, in violation of the independence assumption of standard hypothesis testing (discussed in Section S4.2).

Table S2 summarizes the sign test results for the 2004 data, following the analytical approach used by Hill & Barton [1]. The test is equivalent to a two-sided binomial test ($H_{\text{A}} : f_{\text{red}} \neq 0.5$), where the number of successes is n_{red} and the number of trials is

Table S1: Number (n_{red}) and fraction (f_{red}) of bouts won by red, total number (n), and χ^2 test results in data for boxing (BOX), taekwondo (TKD), Greco-Roman wrestling (GRW), free-style wrestling (FSW), and aggregated (ALL), at the 2004 Athens Olympics.

Sport(s)	n_{red}	n	f_{red}	χ^2	d.f.	p-value
BOX	147	267	0.551	2.73	1	0.098
TKD	43	75	0.573	1.61	1	0.204
GRW	25	48	0.521	0.08	1	0.773
FSW	27	51	0.529	0.18	1	0.674
ALL	242	441	0.549	4.19	1	0.041

$n_{\text{red}} + n_{\text{blue}} < n$, i.e., rounds or weight classes with an equal number of red and blue wins are excluded. For both rounds and weight classes, $f_{\text{red}} = n_{\text{red}}/(n_{\text{red}} + n_{\text{blue}})$ is significantly different from 0.5 at the $\alpha = 0.05$ level. In line with the hypothesis, more present a majority of red wins than a majority of blue wins.

As above, we note that a one-sided test provides a more accurate representation of Hill & Barton's [1] hypothesis, and that we are in a setting of multiple hypothesis testing, with structural dependencies in the data (discussed in Section S4.2). Furthermore, we contend that the number of trials to be used in the test is the total number of rounds or weight classes n, rather than $n_{\text{red}}+n_{\text{blue}}$. That is, rounds and weight classes with an equal number of red and blue wins provide useful information for evaluating the hypothesis, and they should not be excluded from analysis. A final issue is that rounds vary greatly in number of bouts, both within and between sports. Consequently, the probability that different rounds will end with a majority of red wins is not uniform. Similarly, weight classes vary in number of bouts, thus the probability that different weight classes will end with a majority of red wins is also not uniform.

Table S2: Number of rounds and weight classes with $> 50\%$ of red (n_{red}) or blue (n_{blue}) wins, total number (n), and sign test results in data aggregated for boxing, taekwondo, Greco-Roman wrestling, and free-style wrestling, at the 2004 Athens Olympics.

Test	n_{red}	n_{blue}	n	p-value
Rounds	16	4	21	0.012
Weight classes	19	6	29	0.015

Tables S3 and S4 summarize the χ^2 and sign test results for the 2008 data, reported here only for consistency with Hill & Barton's [1] analysis. The fraction of bouts won by red $f_{\text{red}} < 0.5$ in all cases, except taekwondo (with an "excess" of only one red win; Section S3.2.2), and more rounds and weight classes present a majority of blue wins than a majority of red wins. Across the two tests, the only significant result at the $\alpha = 0.05$ level is for bouts in free-style wrestling, but in the opposite direction than predicted by the hypothesis (Table S3).

Table S3: Number (n_{red}) and fraction (f_{red}) of bouts won by red, total number (n), and χ^2 test results in data for boxing (BOX), taekwondo (TKD), Greco-Roman wrestling (GRW), free-style wrestling (FSW), and aggregated (ALL), at the 2008 Beijing Olympics.

Sport(s)	n_{red}	n	f_{red}	χ^2	d.f.	p-value
BOX	133	270	0.493	0.06	1	0.808
TKD	38	75	0.507	0.01	1	0.908
GRW	80	164	0.488	0.10	1	0.755
FSW	67	164	0.409	5.49	1	0.019
ALL	318	673	0.473	2.03	1	0.154

Table S4: Number of rounds and weight classes with > 50% of red (n_{red}) or blue (n_{blue}) wins, total number (n), and sign test results in data aggregated for boxing, taekwondo, Greco-Roman wrestling, and free-style wrestling, at the 2008 Beijing Olympics.

Test	n_{red}	n_{blue}	n	p-value
Rounds	8	13	25	0.383
Weight classes	11	17	29	0.345

Overall, these results call into question the main claim of a red effect in Hill & Barton's [1] analysis. First, the tests used are mis-specified in several respects, outlined above. Second, even ignoring all issues with test mis-specification, the pattern reported for the 2004 data is not robust, with statistical significance hinging on only one or two observations (in the case of bouts and rounds/weight classes, respectively). In any case, the pattern does not hold in the equivalent 2008 data, even replicating the exact analytical approach used by Hill & Barton [1].

S4.1.2 Corollary

Analogous considerations apply to Hill & Barton's [1] skill-based analysis (which we have been able to replicate only approximately, due to ambiguities in processing of the data). A first issue relates to sample size: the analysis involves dividing the $n = 441$ bouts in the aggregated 2004 data into four sub-samples of approximately 110 bouts each, likely reducing the statistical power of the hypothesis tests. A second issue is that the sub-samples retain the structural dependencies of the full sample. A third issue is that, as above, we are in a setting of multiple hypothesis testing (i.e., four separate tests, one for each sub-sample). These considerations suggest that the pattern reported by Hill & Barton [1] is confounded by both Type I and Type II errors, at rates that are unknown or difficult to estimate.

A fourth and final issue applies specifically to the skill-based analysis. Using points as a proxy for skill, and aggregating the information over the four sports, is an uncontrolled procedure with unknown properties. The results yielded by such a procedure are, at best, of difficult interpretation; at worst, they are meaningless. This is because point-scoring systems vary greatly across sports, and the number of points scored is not "linear" with respect to skill. For example, boxing presents the simplest scoring system, which assigns one point for a punch meeting specific requirements that lands to the head or torso. However, even in this case judges rely on additional considerations (e.g., better style, better defense) to break ties. Thus, the assumption of a "linear" relationship between the number of points scored by a contestant and his skill is questionable.

The assumption is even less tenable for the scoring systems of taekwondo and wrestling, where different "actions" are assigned different number of points, with an elaborate set of additional considerations used to break ties (see e.g., the diverse set of potential bout outcomes in wrestling; Section S2.4). Additional complications include the deduction of points in taekwondo (which may result in a contestant ending a bout with an overall negative score), and the distinction between technical vs. classification points in wrestling (which Hill & Barton [1] sidestep by adding the two). In fact, in wrestling a bout is won by the contestant who prevailed in two of the three periods constituting the bout; consequently, the winner may actually have an overall lower score than his opponent. These and related factors are likely to crucially compromise any attempt to extract information about variation in contestant skill from the points scored. We contend that the results of Hill & Barton's [1] skill-based analysis must consequently be discounted.

S4.2 Alternative analytical approach

Here we implement an alternative approach to the analysis underpinning Hill & Barton's [1] main claim (Section S4.1.1). Our approach is in two parts. First, we test for an effect of red in the 2004 and the 2008 data using a series of one-sided binomial tests (Section S4.2.1). Second, we study the rates of Type I and Type II error associated with the analysis, linked to the multiple hypothesis testing and to the variable sample sizes (Section S4.2.2).

While this approach addresses several of the issues with test mis-specification in Hill & Barton's [1] analysis, we note that it does not resolve one fundamental issue, related to structural dependencies in the data. It is a feature of single-elimination tournament competitions that winning contestants compete in multiple rounds (e.g., as they progress to the final; Fig. 1**b** in the main text). Consequently, individual bouts cannot be considered independent observations for the purpose of hypothesis testing. Specific features of the tournament structure may lead to associations between a contestant's skill (and hence his probability of winning) and color, creating a bias towards wins by one color when the data are aggregated over multiple rounds, competitions, and so on [2]. However, any such bias is not easily distentangled from a "real" effect of red, as hypothesized by Hill & Barton [1], in observational data. For example, excluding specific bouts from analysis to remove potential biases [e.g., 2, 4, 5] may lead to a reduction in statistical power. Therefore, we turn to simulation in Section S5 to investigate the bias arising from non-independence in the data-generating process.

S4.2.1 Binomial tests

Following Hill & Barton's [1] approach (Section S4.1.1), we test for an effect in the fraction of bouts won by red (separately for individual sports, and aggregated over the four sports by year), and in the fraction of rounds and weight classes with a majority of red wins (aggregated over the four sports by year). Additionally, we test for an effect in the fraction of bouts won by red, aggregated over the four sports and over the two years; this test maximises the sample size. Finally, we test for an effect in the fraction of bouts won by red excluding data for wrestling, aggregated over the two remaining sports (boxing, taekwondo) by year and over the two years. Our simulation results show that exclusion of the wrestling data minimizes the effects of bias in the data-generating process (Section S5).

This gives a total of 18 one-sided binomial tests ($H_0 : f_{\text{red}} \leq 0.5; H_{\text{A}} : f_{\text{red}} > 0.5$), summarized in Table S5. Only 3 results are significant at the $\alpha = 0.05$ level, relating to (i) the fraction of bouts won by red, aggregated over the four sports, in 2004, (ii) the fraction of bouts won by red, aggregated over boxing and taekwondo, in 2004, and (iii) the fraction of rounds with a majority of red wins, aggregated over the four sports, also in 2004. The latter result is of difficult interpretation due to variation in the number of bouts for different rounds, both within and across sports (Section S4.1.1). Furthermore, we emphasize that we are in a setting of multiple hypothesis testing, hence a threshold of $\alpha = 0.05$ overstates the likely true statistical significance, i.e., we may be making a Type I error of incorrectly rejecting the null hypothesis in these 3 cases. At the same time, we may be making a Type II error of not rejecting the null hypothesis in the other cases. We investigate these issues in detail below.

S4.2.2 Type I and Type II error rates

In statistical hypothesis tests, the critical value α is used as a threshold to decide whether a given pattern is statistically unlikely under a particular null hypothesis H_0. It is thus the probability of a Type I error, i.e., of incorrectly rejecting the null hypothesis when in fact the null is true. If the p-value exceeds α, then the pattern is plausibly within the range of natural variation under the null hypothesis. In our case, natural variation around $f_{\text{red}} = 0.5$, and α is the probability that we find an effect of red when no such effect exists in the underlying data-generating process.

The structure of the analysis underpinning Hill & Barton's [1] main claim is an example

Table S5: Results of one-sided binomial tests of a red effect in data for boxing (BOX), taekwondo (TKD), Greco-Roman wrestling (GRW), free-style wrestling (FSW), and aggregated over the four sports (ALL), at the 2004 Athens and 2008 Beijing Olympics. Tests denoted "bouts" compare the number of bouts won by red, n_{red}, to the n total bouts. Tests denoted "rounds" compare the number of rounds with a majority of red wins, n_{red}, to the n total rounds. Tests denoted "weight classes" compare the number of weight classes with a majority of red wins, n_{red}, to the n total weight classes. In all cases, $f_{\text{red}} = n_{\text{red}}/n$.

Year	Test	Sport(s)	n_{red}	n	f_{red}	p-value
2004	Bouts	BOX	147	267	0.551	0.056
2004	Bouts	TKD	43	75	0.573	0.124
2004	Bouts	GRW	25	48	0.521	0.443
2004	Bouts	FSW	27	51	0.529	0.390
2004	Bouts	ALL	242	441	0.549	0.023
2004	Bouts	BOX, TKD	190	342	0.556	0.023
2004	Rounds	ALL	16	21	0.762	0.013
2004	Weight classes	ALL	19	29	0.655	0.068
2008	Bouts	BOX	133	270	0.493	0.620
2008	Bouts	TKD	38	75	0.507	0.500
2008	Bouts	GRW	80	164	0.488	0.652
2008	Bouts	FSW	67	164	0.409	0.992
2008	Bouts	ALL	318	673	0.473	0.929
2008	Bouts	BOX, TKD	171	345	0.496	0.585
2008	Rounds	ALL	8	25	0.320	0.978
2008	Weight classes	ALL	11	29	0.379	0.932
Both	Bouts	ALL	560	1114	0.503	0.440
Both	Bouts	BOX, TKD	361	687	0.525	0.097

of multiple hypothesis testing, in which more than one test is applied to different aspects of the same data. For example, in our analytical approach, which follows closely Hill & Barton's [1], there are 18 binomial tests, with $\alpha = 0.05$ (Section S4.2.1). If α is the desired rate of false positives, then in a setting of multiple hypothesis testing the significance threshold to be applied to each test is $\alpha_c < \alpha$, i.e., a lower value that depends on the number of tests being conducted. Thus, using a threshold of $\alpha = 0.05$ to evaluate the tests in Table S5 overstates the likely true statistical significance, i.e., we may be making a Type I error of incorrectly rejecting the null hypothesis in cases where the p-value does not exceed the threshold.

There is a rich literature on corrections for multiple hypothesis testing [6]. A standard approach is the well-known Bonferroni correction [7], which gives $\alpha_c = \alpha/m$ for m tests. This is not the most conservative correction, but it is widely used and its theoretical deficiencies are well understood [6]. With $m = 18$, the Bonferroni-adjusted significance threshold for the tests in Table S5 is $\alpha_c = 0.003$. None of the results are significant under this threshold. We emphasize that analogous reasoning applies to the 11 hypothesis tests (seven underpinning the main claim, four underpinning the corollary; Section S4.1) reported by Hill & Barton [1].

In failing to reject the null hypothesis, could we be making a false negative error instead? β represents the probability of a Type II error, i.e., of incorrectly failing to reject the null hypothesis H_0 when in fact it is false and the alternative hypothesis H_A is the correct data-generating process. The power of a test is defined as $1 - \beta$, and a value of 0.8 is a conventional threshold for a test with sufficient power to distinguish between the null and the alternative hypotheses.

For each of the statistical tests in Table S5, we conducted two power analyses, separately for $\alpha = 0.05$ and for $\alpha_c = 0.003$ (Tables S6 and S7, respectively). First, we calculated the parameter $f_{\text{red,alt}}$ of the most likely alternative model, if in fact we have committed a Type II error in failing to reject the null hypothesis (no effect of red). This parameter thus represents the smallest effect of red that is outside the range of natural variation under the null, given the size of the sample. If the null is correct, then as the sample size increases, $f_{\text{red,alt}}$ necessarily converges on 0.5. For large values of n, $f_{\text{red,alt}}$ may be very close to 0.5 and would represent a statistically significant effect, even if it is a very small one.

Second, we calculated the smallest parameter $f_{\text{red,pt}}$ for which we have sufficient power to rule out as an alternative, in the case where we committed a Type II error. As discussed above, under α_c we fail to reject the null hypothesis in all of the tests we conducted (Table S5). Thus, $f_{\text{red,pt}}$ can be interpreted as an upper bound on the possible effect size, and $f_{\text{red,alt}}$ as a lower bound. Their difference Δf can then be interpreted as the largest possible effect of red given the observed data, above and beyond the range of natural variation under the null.

That is, suppose the true data-generating process includes an effect of red, and suppose that by chance, our observed data f_{red}, which are drawn from this process, are sufficiently close to 0.5 that we fail to reject the null hypothesis because the observed fraction is not statistically unusual relative to the null (although it may be unusual relative to the true parameter of the data-generating process). In this case, we have made a Type II error. However, given f_{red}, we can calculate how much statistical power it has against different choices of red effect $f_{\text{red,pt}} > 0.5$. The bigger the choice of $f_{\text{red,pt}}$, the more statistical power our observed f_{red} provides against it. In this way, for each candidate choice of $f_{\text{red,pt}}$, we can calculate the probability of observing f_{red}. The less likely our observed f_{red} is under a particular $f_{\text{red,pt}}$, the more statistical power we have against it. The more statistical power, the less likely it is that that parameter could have generated our observed data.

Table S6: Results of power analyses of one-sided binomial tests of a red effect in data for boxing (BOX), taekwondo (TKD), Greco-Roman wrestling (GRW), free-style wrestling (FSW), and aggregated over the four sports (ALL), at the 2004 Athens and 2008 Beijing Olympics. Tests denoted "bouts" compare the number of bouts won by red, n_{red}, to the n total bouts. Tests denoted "rounds" compare the number of rounds with a majority of red wins, n_{red}, to the n total rounds. Tests denoted "weight classes" compare the number of weight classes with a majority of red wins, n_{red}, to the n total weight classes. In all cases, $f_{\mathrm{red}} = n_{\mathrm{red}}/n$. k is the critical value for a one-sided binomial test ($H_0 : f_{\mathrm{red}} \leq 0.5; H_{\mathrm{A}} : f_{\mathrm{red}} > 0.5$), i.e., the test will be significant at α for $n_{\mathrm{red}} \geq k+1$; $f_{\mathrm{red,alt}} = (k+1)/n$. Power is the power of the test. n_{pt} is the smallest value of n_{red} consistent with a power threshold of 0.80; $f_{\mathrm{red,pt}} = n_{\mathrm{pt}}/n$. $\Delta f = f_{\mathrm{red,pt}} - f_{\mathrm{red,alt}}$.

Year	Test	Sport(s)	n_{red}	n	f_{red}	k	$f_{\mathrm{red,alt}}$	Power	n_{pt}	$f_{\mathrm{red,pt}}$	Δf
2004	Bouts	BOX	147	267	0.551	147	0.554	0.476	155	0.581	0.026
2004	Bouts	TKD	43	75	0.573	45	0.613	0.281	49	0.653	0.040
2004	Bouts	GRW	25	48	0.521	30	0.646	0.055	34	0.708	0.062
2004	Bouts	FSW	27	51	0.529	31	0.627	0.103	35	0.686	0.059
2004	Bouts	ALL	242	441	0.549	238	0.542	0.632	248	0.562	0.020
2004	Bouts	BOX, TKD	190	342	0.556	186	0.547	0.649	195	0.570	0.023
2004	Rounds	ALL	16	21	0.762	14	0.714	0.784	17	0.810	0.095
2004	Weight classes	ALL	19	29	0.655	19	0.690	0.431	22	0.759	0.069
2008	Bouts	BOX	133	270	0.493	149	0.556	0.022	157	0.581	0.026
2008	Bouts	TKD	38	75	0.507	45	0.613	0.041	49	0.653	0.040
2008	Bouts	GRW	80	164	0.488	93	0.573	0.017	99	0.604	0.030
2008	Bouts	FSW	67	164	0.409	93	0.573	0.000	99	0.604	0.030
2008	Bouts	ALL	318	673	0.473	358	0.533	0.001	370	0.550	0.016
2008	Bouts	BOX, TKD	171	345	0.496	188	0.548	0.030	197	0.571	0.023
2008	Rounds	ALL	8	25	0.320	17	0.720	0.000	20	0.800	0.080
2008	Weight classes	ALL	11	29	0.379	19	0.690	0.001	22	0.759	0.069
Both	Bouts	ALL	560	1114	0.503	584	0.525	0.071	599	0.538	0.013
Both	Bouts	BOX, TKD	361	687	0.525	365	0.533	0.366	377	0.549	0.016

Table S7: Results of power analyses of one-sided binomial tests of a red effect in data for boxing (BOX), taekwondo (TKD), Greco-Roman wrestling (GRW), free-style wrestling (FSW), and aggregated over the four sports (ALL), at the 2004 Athens and 2008 Beijing Olympics. Tests denoted "bouts" compare the number of bouts won by red, $n_{\rm red}$, to the n total bouts. Tests denoted "rounds" compare the number of rounds with a majority of red wins, $n_{\rm red}$, to the n total rounds. Tests denoted "weight classes" compare the number of weight classes with a majority of red wins, $n_{\rm red}$, to the n total weight classes. In all cases, $f_{\rm red} = n_{\rm red}/n$. k is the critical value for a one-sided binomial test ($H_0 : f_{\rm red} \leq 0.5; H_{\rm A} : f_{\rm red} > 0.5$), i.e., the test will be significant at $\alpha_{\rm c}$ for $n_{\rm red} \geq k+1$; $f_{\rm red,alt} = (k+1)/n$. Power is the power of the test. $n_{\rm pt}$ is the smallest value of $n_{\rm red}$ consistent with a power threshold of 0.80; $f_{\rm red,pt} = n_{\rm pt}/n$. $\Delta f = f_{\rm red,pt} - f_{\rm red,alt}$.

Year	Test	Sport(s)	$n_{\rm red}$	n	$f_{\rm red}$	k	$f_{\rm red,alt}$	Power	$n_{\rm pt}$	$f_{\rm red,pt}$	Δf
2004	Bouts	BOX	147	267	0.551	156	0.588	0.121	164	0.614	0.026
2004	Bouts	TKD	43	75	0.573	49	0.667	0.063	53	0.707	0.040
2004	Bouts	GRW	25	48	0.521	33	0.708	0.006	36	0.750	0.042
2004	Bouts	FSW	27	51	0.529	35	0.706	0.008	39	0.765	0.059
2004	Bouts	ALL	242	441	0.549	250	0.569	0.208	260	0.590	0.020
2004	Bouts	BOX, TKD	190	342	0.556	197	0.579	0.207	206	0.602	0.023
2004	Rounds	ALL	16	21	0.762	17	0.857	0.227	19	0.905	0.048
2004	Weight classes	ALL	19	29	0.655	22	0.793	0.082	25	0.862	0.069
2008	Bouts	BOX	133	270	0.493	158	0.589	0.001	166	0.615	0.026
2008	Bouts	TKD	38	75	0.507	49	0.667	0.004	53	0.707	0.040
2008	Bouts	GRW	80	164	0.488	100	0.616	0.001	106	0.646	0.030
2008	Bouts	FSW	67	164	0.409	100	0.616	0.000	106	0.646	0.030
2008	Bouts	ALL	318	673	0.473	372	0.554	0.000	384	0.571	0.016
2008	Bouts	BOX, TKD	171	345	0.496	198	0.577	0.002	207	0.600	0.023
2008	Rounds	ALL	8	25	0.320	19	0.800	0.000	22	0.880	0.080
2008	Weight classes	ALL	11	29	0.379	22	0.793	0.000	25	0.862	0.069
Both	Bouts	ALL	560	1114	0.503	603	0.542	0.005	618	0.555	0.013
Both	Bouts	BOX, TKD	361	687	0.525	380	0.555	0.068	392	0.571	0.016

Following standard conventions, any choice of $f_{\mathrm{red,pt}}$ against which we have power of at least 0.8 can be ruled out by the value of f_{red} we did observe. That is, the observed value of f_{red} allows us to rule out extreme scenarios (strong red effect) because they would be unlikely to produce an observed f_{red} so far below $f_{\mathrm{red,pt}}$. The smallest value of $f_{\mathrm{red,pt}}$ that we cannot rule out in this way is the largest value (i.e., the biggest effect) that is consistent with our observed value f_{red} under these rules (which assume that we have committed a Type II error).

Now we know how big a red effect cannot be ruled out by the observed f_{red}, but we would like to estimate the impact of this effect on our observed data. The null hypothesis of $f_{\mathrm{red}} = 0.5$ has a range of natural variation, and this is naturally defined as all values of f_{red} below the critical value. Any observed outcomes that fall within this region can be attributed to the null. Variation above and beyond this region can be attributed to an effect of red. The limit of this variation is precisely what we have calculated above, and the difference between the upper limit of natural variation, given by the null and the sample size, and the upper limit of a possible effect of red, permitted by the observed data f_{red}, is the maximum impact that red could be having on the observed data.

Across all of our tests, the values for Δf are all small, indicating that if an effect of red does exist in these data, it is a modest one, accounting for altering the outcomes of only a handful of bouts. In particular, in the data aggregated over the four sports and over the two years, $\Delta f = 0.013$, indicating that at most wearing red could be altering the outcomes of about 1.3% of bouts. We find similar results when we focus only on the more unbiased tournaments, namely boxing and taekwondo (Tables S6 and S7). We emphasize that these values do not take into account the potential biases arising from asymmetries in tournament structure, thus they may even overestimate the true impact.

S5 Monte Carlo simulation of competition

Hypothesis tests like the χ^2, sign, and binomial tests make the standard assumption that observations are independent and identically distributed (i.i.d.). As discussed in Section S4.2, individual bouts in a single-elimination tournament cannot be considered independent observations for the purpose of hypothesis testing. Here we use simulation to investigate the bias arising from non-independence in the data-generating process. Specifically, we focus on two sources of incompleteness in the tournament structures for the four sports in the 2004 Athens and 2008 Beijing Olympics: byes and walkovers (Sections S2 and S3). We show that incompleteness in the tournament structure, coupled with variance in skill among the contestants, can induce a bias that shifts the null distribution (no effect of red) away from $f_{\text{red}} = 0.5$.

The first source of incompleteness comes from the number of contestants in the outermost round in a competition tree being different from a power of 2 (Fig. 1**b** in main text). In this case, a subset of contestants are awarded byes, which effectively allows them to skip to the next round. If n is the number of contestants in a given round, the number of byes awarded is $\delta = 2^{\lceil \log_2 n \rceil} - n$, and the number of contestants competing in that round is $n - \delta$. For example, the average number of byes per weight class in the 2004 data is $\langle \delta \rangle = 3.4$ for boxing, 1.6 for Greco-Roman wrestling, and 1.3 for free-style wrestling (taekwondo has no byes because the number of contestants per weight class is 16, and thus a power of 2; Section S2.2). Byes are "stacked" either at the top of the tree's upper branch (2004 and 2008 boxing, 2008 wrestling; Sections S2.1 and S2.3) or at the bottom of the lower branch (2004 wrestling; Section S2.3).

The second source of incompleteness comes from bouts won by walkover, in which a contestant fails to show up for a bout or withdraws, leading his opponent to win by default; for example, this was common in the 5–6 final round of 2004 wrestling (Section S2.3). These bouts are reasonably excluded from analysis as their outcome is unlikely to be influenced by the colors worn by the contestants [1]. Unlike in the case of byes, there is no systematic pattern governing the position of bouts won by walkover. Across all sports, there were 16 walkovers in 2004 and four in 2008 (Section S3).

S5.1 Simulation specification

To control for structural sources of non-i.i.d. behavior in the data-generating process for the correct null distribution, we implemented an exact Monte Carlo simulation of competition on the observed tournament structures for each weight class in each sport, including the specific asymmetries generated by byes and walkovers observed in each weight class. This simulation allows us to numerically estimate the correct distribution for the null hypothesis (no effect of red) within an individual bout.

The simulation was parameterized in a way that allows us to systematically vary the skill levels of the competitors. Specifically, for each simulation, each competitor is assigned

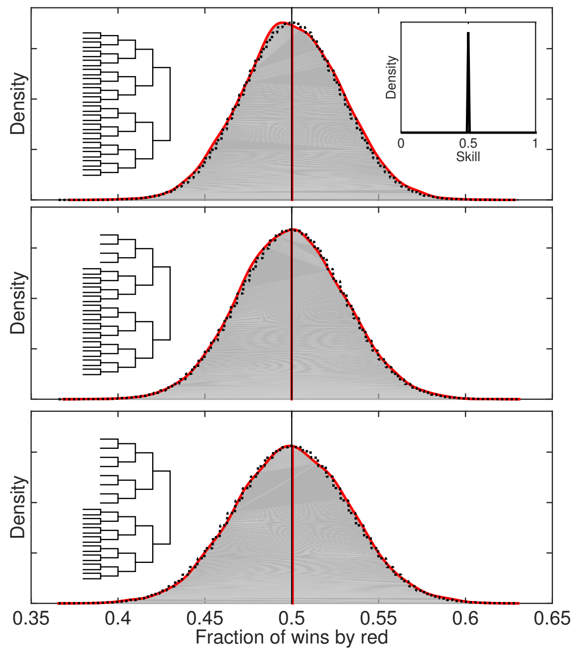

Figure S1: Fraction of red wins under Bradley–Terry competition on tournaments with different degrees of incompleteness due to byes (left insets), when all competitors are equally skilled (right inset, top panel), for 10^4 repetitions. Color assignment is red/blue to the top/bottom positions of the bracket in each bout. Regardless of the degree of incompleteness, the null distribution takes the shape of a binomial distribution (dotted line) centered at $f_{\text{red}} = 0.5$.

a latent skill value x drawn from a symmetric Beta distribution $x \sim \text{Beta}(\beta, \beta)$ over the unit interval, but independently of the color initially assigned to the competitor. A competitor's skill value is fixed over all bouts in which he participates. In the limit of $\beta \to \infty$, this distribution converges on a delta function at $x = 0.5$, meaning that all competitors have equal skill. For finite values of β, the distribution has non-zero variance but is symmetric about $x = 0.5$. When $\beta = 1$, $x \sim \text{Uniform}(0, 1)$, and for $\beta < 1$, the distribution exhibits a symmetric "U" shape, with the modal skill values being close to 0 or 1.

When two competitors r and b face off, the outcome is determined by a standard Bradley–Terry model of competition [8], in which the probability that a competitor wearing red wins is $x_r/(x_r + x_b)$. The winner of the bout advances to the next relevant position in the tournament. If a particular bout was won by walkover in the empirical data, then

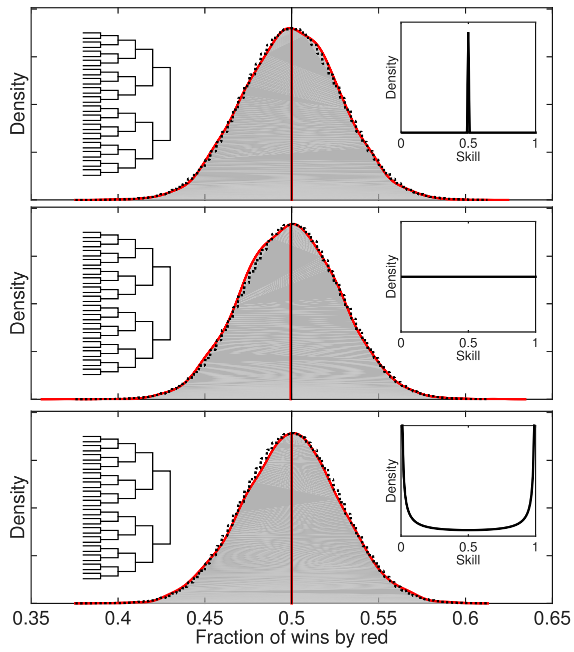

Figure S2: Fraction of red wins under Bradley–Terry competition on three complete tournaments (left insets), when competitors have unequal skills (right insets), for 10^4 repetitions. Color assignment is red/blue to the top/bottom positions of the bracket in each bout. Regardless of the variance in competitor skill, the null distribution takes the shape of a binomial distribution (dotted line) centered at $f_{\text{red}} = 0.5$.

the corresponding winner in the simulation automatically wins the simulated bout and advances. If additional bouts were a part of a particular tournament, these were included, with competitors allocated to these bouts according to the same rules as were applied in the tournament (e.g., repechage rounds in 2004 and 2008 taekwondo and in 2008 wrestling; Sections S2.2 and S2.3).

We use the above Monte Carlo simulation, with at least 10^4 repetitions, to numerically estimate the correct null distribution. This distribution can be used to calculate a standard p-value under a hypothesis of a particular distribution of competitor skills. It can also be used to quantify the impact of different seeding procedures (e.g., seeding by skill) on the null distribution, which can shift the fraction of red wins away from 0.5 [2].

Finally, the simulation can be used to characterize the bias induced in the distribution of the fraction of red wins under the null hypothesis (no effect of red) by asymmetries

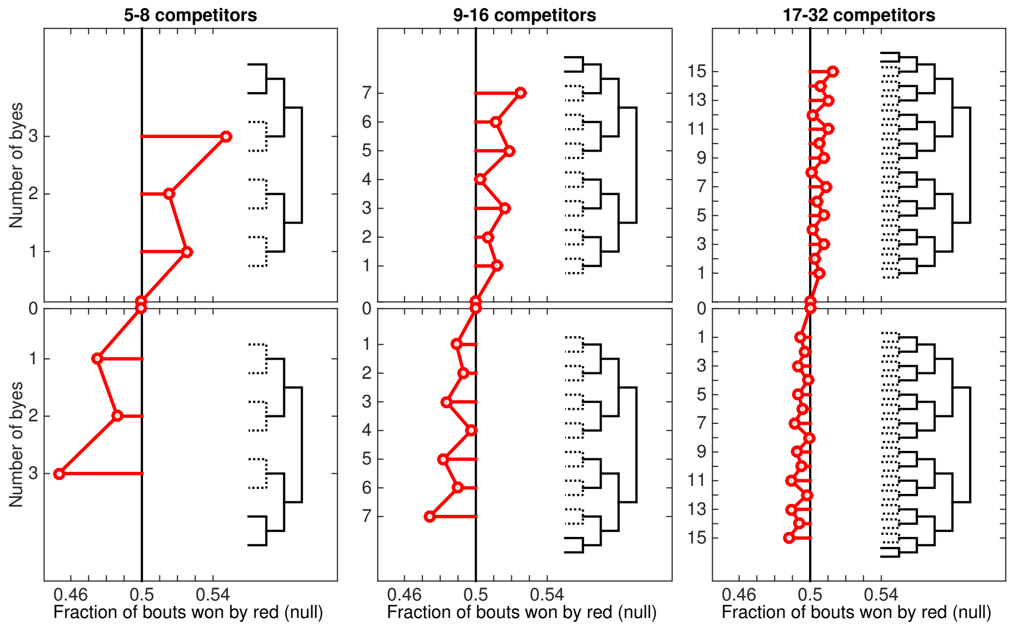

Figure S3: Fraction of red wins under the null hypothesis (no effect of red) on tournaments with 5–8, 9–16, and 17–32 competitors, covering all possible number of byes in the outermost round. Byes are stacked either at the bottom of the round (and hence drawn beginning from the "lowest" bout in the round; upper panels) or at the top (and hence drawn beginning from the "highest" bout in the round; lower panels). Across all simulations, the skill distribution is fixed with $\beta = 0.1$ (the largest level of variance shown in Figs. 1**c,d** in the main text). Color assignment is red/blue to the top/bottom positions of the bracket in each bout. Dots show the mean fraction under 10^4 repetitions.

in the tournament structure. For instance, Figs. S1 and S2 show that when competitors with equal skill compete in an incomplete tournament, or when competitors with unequal skill compete in a complete tournament, the null distribution of the fraction of red wins is given by a binomial distribution centered at $f_{\text{red}} = 0.5$. In contrast, Fig. S3 shows that when competitors have unequal skill and compete in an incomplete tournament, the null distribution is shifted away from $f_{\text{red}} = 0.5$ by an amount that varies non-trivially with (i) the number of competitors and (ii) the number of byes in the outermost round. However, the direction of the bias relative to 0.5 depends only on whether the byes are stacked at the bottom of the round, which leads to $f_{\text{red}} > 0.5$ (i.e., more wins by red), or at the top of the round, which leads to $f_{\text{red}} < 0.5$ (i.e., more wins by blue).

S5.2 Simulation results

To characterize the causal role that tournament asymmetries play in inducing a bias towards one color, we compared the above simulation, on real tournament structures, with an identical simulation in which the empirical tournament asymmetries were removed. That is, in this second simulation, no byes or walkovers were allowed and every weight class tournament was complete and symmetric. Symmetrizing the tournament structures necessarily adds bouts, thereby increasing the simulated sample size. To compensate, we pruned a corresponding number of bouts from locations of symmetry within the tournament, e.g., the final (gold) round in any sport, the 5–6 final round in 2004 wrestling, or an entire repechage tournament in 2004 and 2008 taekwondo and 2008 wrestling (Section S2).

Fig. 1**c,d** in the main text shows how the location of the null distribution varies as a function of competitor skill variance for both simulations, for 10^5 repetitions for the 2004 and the 2008 tournament structures. Clearly, the location of the distribution shifts substantially in the presence of tournament asymmetries, but remains centered at $f_{\text{red}} = 0.5$ when asymmetries are absent. That is, tournament asymmetries induce a bias in the null distribution, generating a tendency for red to win more often in 2004, and for blue to win more often in 2008.

Most of these deviations are driven by the wrestling tournament structures, which exhibit large degrees of asymmetry in both 2004 and 2008. For a tournament with n contestants and δ byes, the outermost round of competition will have $n - \delta$ contestants, and thus a "completeness" fraction linked to byes of $\rho = (n - \delta)/2^{\lceil \log_2 n \rceil}$. The average value of ρ across weight classes in a sport provides a simple measure of how systematically asymmetric its tournaments are. The table below gives these calculated values for the four sports in 2004 and 2008.

Sport	Year	$\langle \rho \rangle$	Year	$\langle \rho \rangle$
BOX	2004	0.79	2008	0.79
TKD	2004	1.00	2008	1.00
GRW	2004	0.61	2008	0.24
FSW	2004	0.68	2008	0.24

Disaggregating the simulation results for each year by individual sport shows that the primary source of the bias is in the tournaments with the greatest degree of incompleteness linked to byes, Greco-Roman wrestling and free-style wrestling (Figs. S4 and S5). No bias appears in taekwondo because in both years there are no byes (Section S3). The boxing tournaments present incompleteness linked to byes in both years (Section S3). However, in these cases the size of the tournaments is large enough that the bias is offset by the relatively large number of symmetric bouts (Figure S3).

Under the null hypothesis of no effect of red, the expected number of red wins within any particular sport will vary stochastically relative to the null. In the cases of boxing and taekwondo individually, the observed number of red wins is within the natural variation

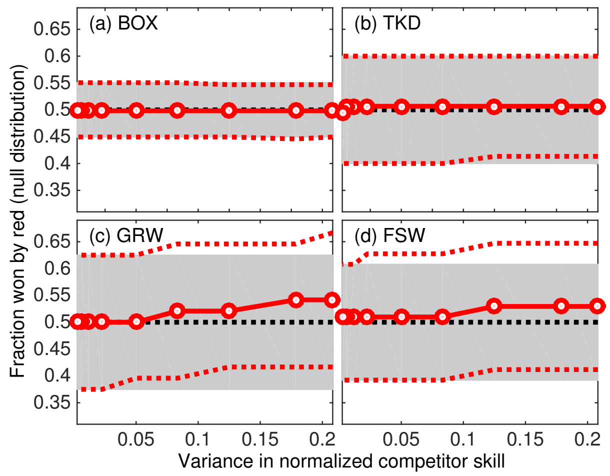

Figure S4: Distributions (5, 50, and 95% quantiles) of the fraction of red wins under the null hypothesis (no effect of red), for the asymmetric 2004 tournaments and equivalent symmetric tournaments, by sport. The distributions were evaluated by Monte Carlo at the locations of the red dots. Combining these distributions yields those shown in Fig. 1**c** in the main text.

we expect, for a null with no structural bias. In the cases of Greco-Roman wrestling and free-style wrestling, the observed number of red wins in 2004, and the observed number of blue wins in 2008, are likely enhanced as a result of the structural bias we identified in these sports.

S5.3 Summary

Our simulations show that in the case where all contestants have equal skill, and thus the chance that any color wins in a given bout is even, the tournament structure has no impact on the null distribution of the fraction of red wins f_{red} (Figure S1); in this case, a standard hypothesis test would be sufficient to detect the presence of a red effect. However, there is no evidence supporting an assumption that Olympic athletes have equal skill. In the unequal skill case, if the tournament structure is symmetric, with an equal number of bouts occurring in upper and lower branches of the competition tree, then the null distribution of the fraction of wins by red is also symmetric about $f_{\text{red}} = 0.5$ (Figure S2) and a standard hypothesis test would be sufficient. However, when contestants vary in skill and the tournament structure is incomplete, the null distribution is shifted away from $f_{\text{red}} = 0.5$ (Figure S3). Crucially, the direction of the bias depends only on whether the byes are stacked at the bottom of the outermost round, which leads to $f_{\text{red}} > 0.5$ (i.e.,

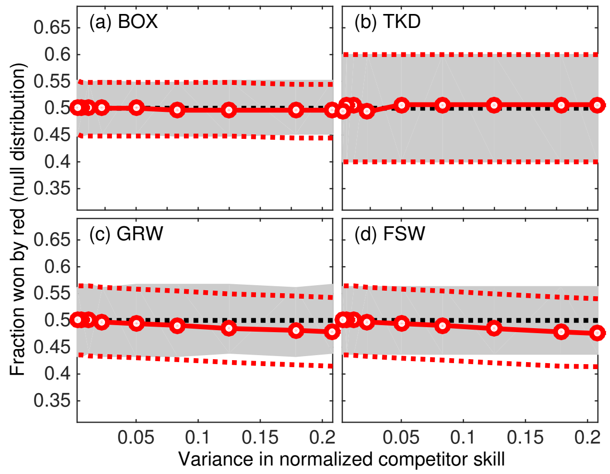

Figure S5: Distributions (5, 50, and 95% quantiles) of the fraction of red wins under the null hypothesis (no effect of red), for the asymmetric 2008 tournaments and equivalent symmetric tournaments, by sport. The distributions were evaluated by Monte Carlo at the locations of the red dots. Combining these distributions yields those shown in Fig. 1**d** in the main text.

more wins by red), or at the top of the round, which leads to $f_{\text{red}} < 0.5$ (i.e., more wins by blue).

While the true variance in skill among contestants in the four sports at the 2004 and 2008 Olympics is not known (Section S4.1.2), simulations on the actual tournament structures for the two competitions show a structural bias towards wins by red and blue, respectively. This is consistent with the empirical pattern observed for each competition (Table S5 and Fig. 1**a** in the main text). Furthermore, as discussed in Section S2.4, the only major structural change that occurred between the two competitions is in the wrestling tournaments, including a shift in the position of the byes from the bottom of the qualification (8-contestant) round in 2004 to the top of the qualification (32-contestant) round in 2008. Consistently, our simulations show that byes at the bottom of the round create a bias towards wins by red, whereas byes at the top a create a bias towards wins by blue.

S6 Discussion

We have provided multiple lines of analysis, together with new data, to evaluate Hill & Barton's [1] hypothesis that the effect of red on human competition is a response shaped by sexual selection, analogous to the response observed in other animal species.

First, we have shown that the results reported by Hill & Barton [1] in support of the hypothesis, based on analysis of data for four sports in the 2004 Athens Olympics, present several shortcomings, from issues of test mis-specification to issues with interpretation. Even ignoring these, the results underpinning the main claim of a red effect in Hill & Barton's [1] analysis are not robust. Consistently, we find that the pattern does not hold in equivalent data for the 2008 Beijing Olympics, even replicating the exact analytical approach used by Hill & Barton [1].

Second, we have implemented an alternative analytical approach, which addresses several of the issues with Hill & Barton's [1] analysis. This approach allows us to investigate the rates of Type I and Type II errors associated with our analysis. The results show that there is no evidence of a red effect in either the 2004 or the 2008 data, and that the magnitude of any effect that may exist in these data is necessarily small.

Finally, we have used simulation to investigate systematic biases in the data-generating process, due to asymmetries in the tournament structures. The results provide evidence of a structural bias towards wins by red in the outcomes of the 2004 competition, confirmed by evidence of a structural bias towards wins by blue in the outcomes of the 2008 competition, and consistent with patterns observed in the data for the two years.

These multiple lines of analysis and new data converge to show that the effect of red on human competition reported by Hill & Barton [1] can be fully accounted for by structural features of the tournaments in the four sports analysed. Contrary to previous claims [e.g., 9], the reported pattern is not due to randomness, but to systematic bias towards wins by red in the outcomes of the 2004 competition. Thus, it is not necessary to invoke any of the behavioral, structural, and other confounds that have been proposed over the years as alternative interpretations to Hill & Barton's [1]. These include, for example, differences in visibility between colors, asymmetries in prior experience across contestants (e.g., win–lose effects and/or number of previous bouts fought), and differences in recovery time due to variation in intervals between contests [2, 4, 5]. Of course, based on our results we cannot exclude that these or other factors may operate [see e.g., 10, for suggestive evidence of a bias in competition judges in taekwondo favoring contestants wearing red]. Yet in the absence of strong evidence that such factors were at play in the 2004 Athens Olympic competition, our explanation provides the most comprehensive and parsimonious account to date of the pattern reported by Hill & Barton [1].

What are the implications for the hypothesis of an effect of red on human competition, shaped by evolutionary processes? A large body of work has developed over the past decade, straddling the biological and social sciences, building on Hill & Barton's [1] influential study [reviewed in 11–13]. Because of the difficulties that arise in disentangling a "real" effect of red from potential confounds in observational data, researchers have increasingly turned to experiments to demonstrate the existence of such an effect. Irrespective of the approach used, work in this area routinely points to Hill & Barton's [1] results as *the* key evidence that the effect of red on human competition is a response shaped by sexual selection and, by implication, *the* key evidence of "parallels between the human and nonhuman response to color" [13, p. 115]. By refuting Hill & Barton's [1] results, our work calls for a critical re-evaluation of this body of work, together with re-assessment of the theoretical premise of future studies that depend on it conceptually. In particular, extreme caution is required when invoking an evolutionary basis for any effect of red on human competition, and on human behavior more generally, that is demonstrably robust.

Ultimately, there is no question about whether the human response to color has been shaped by evolutionary processes over time. In what way it has been shaped by them, and how this is reflected in present-day human behavior, remain open questions.

References

1. Hill, R. A. & Barton, R. A. Red enhances human performance in contests. *Nature* **435,** 293–293 (2005).
2. Dijkstra, P. D. & Preenen, P. T. No effect of blue on winning contests in judo. *Proceedings of the Royal Society of London B: Biological Sciences* **275,** 1157–1162 (2008).
3. Seife, C. *Red does not enhance human performance in the Olympics* `http://www.users.cloud9.net/~cgseife/SeifeOlympicsManuscript12February.pdf`. [Online; retrieved on 2014-03-06].
4. Rowe, C., Harris, J. M. & Roberts, S. C. Corrigendum: Sporting contests: Seeing red? Putting sportswear in context. *Nature* **441,** E3 (2005).
5. Rowe, C., Harris, J. M. & Roberts, S. C. Sporting contests: Seeing red? Putting sportswear in context. *Nature* **437,** E10–E10 (2005).
6. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society: Series B* **57,** 289–300 (1995).
7. Dunn, O. J. Estimation of the medians for dependent variables. *Annals of Mathematical Statistics* **30,** 192–197 (1959).
8. Bradley, R. & Terry, M. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39,** 324–345 (1952).
9. Seife, C. *Proofiness: how you're being fooled by the numbers* (Penguin Publishing Group, 2010).
10. Hagemann, N., Strauss, B. & Leißing, J. When the referee sees red... *Psychological Science* **19,** 769–771 (2008).
11. Wiedemann, D., Barton, R. A. & Hill, R. A. in *Applied evolutionary psychology* (ed Roberts, S. C.) 290–307 (Oxford University Press, Oxford, 2012).
12. Maier, M. A., Hill, R. A., Elliot, A. J. & Barton, R. A. in *Handbook of color psychology* (eds Elliot, A. J., Fairchild, M. D. & Franklin, A.) 568–584 (Cambridge University Press, Cambridge, 2015).
13. Elliot, A. J. & Maier, M. A. Color psychology: effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology* **65,** 95–120 (2014).

Acknowledgments

We thank Tony Black of USA Wrestling, Jeongkang Seo of the World Taekwondo Federation (WTF), Sébastien Gillot of the International Boxing Association (AIBA), “Iceman” John Scully, Janusz Majcher, Andrew Baldwin, Justin Rao, for helpful conversations about Olympic sports. Adam Kenny and Sears Merritt provided research assistance. Mirta Galesic, Robert Barton, and Russell Hill provided helpful feedback on the manuscript.

Session information

- R version 3.3.1 (2016-06-21), `x86_64-pc-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: dplyr 0.5.0, knitr 1.14, xtable 1.8-2
- Loaded via a namespace (and not attached): assertthat 0.1, DBI 0.5-1, evaluate 0.9, formatR 1.4, highr 0.6, lazyeval 0.2.0, magrittr 1.5, R6 2.1.3, Rcpp 0.12.7, stringi 1.1.1, stringr 1.1.0, tibble 1.2, tools 3.3.1