Response to EJSS reviewers' comments on manuscript #2019-0005

We are grateful for being granted the opportunity to revise and resubmit our manuscript. We have considered all the comments by both the reviewers and the Editor. They have helped us craft a stronger manuscript than we otherwise would be able to do. We believe our responses to the comments and criticisms made have made the work a stronger contribution to the literature.

In what follows, we respond to each of the comments and explain how we accordingly have adapted our work. In the cases where we have chosen not to follow the advice, we motivate our choices for doing so.

Thank you for reading and commenting on our work!

Responses to comments made by the Editor

From the document attached with the Editor's Decision letter, we identified twelve comments and criticisms. We list the items as we have read them below, in order of appearance in the document, along with a brief statement on how we have dealt with the issue identified by the Editor.

1. *Overemphasis on the prestige of expert judgement and being a sports judge.*

We have reworded the period on prestige of sport judges.

2. *The size of the scholarly literature*

We have rephrased the first period of the literature review.

3. *Typo: Em-dash before "effectively handing them the gold medal"*

The period is rewritten in the new version.

4. Typo: "Our second hypotheses*" should be "Our second hypothesis"*

This is corrected in the new version.

5. The age of the data used in the analysis.

In the new version, we also use data from the 2015-2016 season's World Cup competitions. We describe how we obtained these data and how they were treated in detail under our responses to reviewer #2's comments.

6. The argument that results from ski jumping competitions may have relevance beyond the sport, needs a more solid foundation and a more precise characterization of ski jumping.

We have toned down this argument somewhat in the new version.

7. The Editor appreciates the thorough explanation of statistical models used.

Thank you! In this version, we have only changed small bits of material to make the models even easier to understand.

8. Sources of national differences in tendencies of national bias (french vs. others) – what may they be? The Editor suggests that the potential growth in the sport nationally may be a factor.

We now build upon sociological theory on evaluation to propose hypotheses on variation in biases in the field of ski jumping.

9. Our estimates of bias are less than Zitzewitz', which may be the result of a general professionalization of the sport. New data may cast further light on this issue.

We now, with our new data, explicitly discuss and test a hypothesis of change over time.

10. Editor suggest that safeguards against biases are already in place (5 judges from different nations; lowest and highest scores excluded).

This is of course correct. In the new version, we make more room for discussing these safgeguards and their effects.

11. Editor suggests reference to sports-related political turmoil is too far fetched. Instead, we should provide more social context of the ski jumping sport.

We have followed this advice, and removed the political turmoil-related reference.

12. Reference list: There were missing data for two references (Sampaio 2012 and Yang 2006).

The appropriate information has been added for these references.

We are grateful to the Editor for her/his comments on our manuscript. They helped us writerethink and rewrite the work in a way that we believe is a stronger contribution to the field. Thank you!

Responses to comments made by reviewer #2

Let us first express our thanks for the Reviewer's thoughtful comments on our manusript. In the folling paragraphs we respond to each of the comments and criticisms. We quote the Reviewer's comments (with minor edits) first before offering our response.

Distortion effects on the evaluation of performance in ski jumping are to be investigated. One is the national origin of the judges and the other is the home advantage of the athletes. The research situation on the problem of the influence of nationality seems to be quite well worked out. The following hypotheses are formulated (p. 6):

1) There will be a "positive distortion" of the results when the judges evaluate performances of athletes who come from the same nation as themselves.

2) Another hypothesis claims that it is precisely the assumed effect that diminishes in strength when other variables are controlled (here I wonder about the sense). It is normal and plausible that the more variables are put into a model, the lower the relative explanatory share of a single variable. It would be interesting if the explanatory value of a variable remained high despite the control of disturbing variables (secondary variance).

Yes, it is typical that when a new variable (Z) is introduced the regression coefficient of a first variable (X) may change both magnitude and sign if the new variable is correlated with X. In our case, it may that competitors from certain nations are better on average than competitors from other nations, and it may be that the above-average nations also are more often represented on judge panels. This will induce a correlation between the same-nation indicator (X) and qualities of the performance (Z). Controlling for Z, should then reduce the magnitude of the regression coeffient for X.

Our estimate from model 1, the "naïve" model, may be overestimating the strength of bias for such reasons, and in model 2 we control for *all* unobserved characteristics of the jump that affects the judge scores.

3) Finally, in a third hypothesis, cultural influences on stylistic elements are assumed (and I wonder whether this would not be an explanation for the variance of the variable "national influence"). So why is this separated? The authors differentiate between nations with a strong tradition of ski jumping and those with a rather weak tradition. In a first step there are some points critical to mention:

1) The authors use a fixed effects-term for each jump in their regression-model to remove confounding factors and thereby estimate national bias more precisely. This is done under the assumption that the other - in this case four - judges give an unbiased score. Although this often so called „true score" is the common modus operandi in terms of the questions to be answered, the authors are to be pointed out on the fact that this assumption only holds true if no set of five judges contains two or more judges of identical nationality.

Specific rules from the regarded years concerning the composition of the judging panel in ski jumping could be presented to support the assumption in that sense.

Yes, we use a fixed-effects term for each jump to remove confounding factors from characteristics of the jump. However, we believe this model does *not* rely on an assumption of unbiasedness of the other four scores nor an assumption that there is only one judge from the competitor's nation in the judge panel. If there are multiple judges from the competitor's country, there will simply be multiple 1's among the five observations that make up the observations of judge scores for each jump. What identifies the B (in the revised ms denoted β) parameter in the models, is the within-jump *variation* in judge-competitor correspondence. The identification criterion is that the data must include some performances (jumps) where there is one (or more) judges from the same nation as the competitor, and thus scores 1 on the dummy variable for national bias. As we have quite a few of such performances, we are able to identify and estimate an effect of nationalistic bias on judge scores.

2) Additionally the first part oft the third hypothesis proposed by the authors seems not to be testable the way it is formulated, it is rather formulated as an additional assumption. („A third hypothesis is that this bias may be partly due to cultural influences on stylistic elements,[…]"). For the second part of the third hypothesis however, there was no test of the hypothesis performed, nor was the concept of differing countries according to tradition for ski jumping further elucidated.

We agree with you that the third hypothesis was mangled and untestable the way it was formulated. In the revised version, we have rewritten the hypothesis section and motivated the hypotheses better, in accordance with the new theoretical background.

3) Furthermore on page 12 it is stated that there would be a fluctuation in the number of observations due to missing values in the independent variables. Table 2 yet shows 18.860 observations for all three models presented, which indicates that all the available observations (3772 jumps x 5 judges) have been used to calculate the first three estimations.

Thank you for pointing this out. This was an error on our part. It is corrected in the current version of the manuscript.

> *Home advantage is divided into a physical effect and a social effect. These constructs should be explained more clearly (see e.g. p. 10). I suspect that - according to the state of research (referenced on p. 5) - actual factors influencing performance exist (the sporting performance in the home country is actually on average higher than in a host country), but also influenced due to social pressure from the audience on the adjudicators, which can be assessed as higher in the home countries of the athletes than in the host countries (this is also explained on p. 10, this would then again be a distortion of judgement).*

We have expanded our discussion of home field advantages somewhat in this version, and make amongst other things the discintion you point to in your comment. TODO: OR SHOULD WE JUST DROP HOME FIELD?

> *It is expectable that the more variables are put into a model, the lower the relative explanatory share of a single variable. It would be interesting if the explanatory value of a variable remained high despite the control of disturbing variables (secondary variance).*
>
> *Some remarks on the regression model: The following regression models are proposed.*
>
> *National affiliation of the judges:* * sijp = B-(Lj- ϕ(I = J)) + qp + ⊠ijp s = score B = coefficient Φ = Dummy variable (nationality judge and athlete equal or unequal) L = leniency ("mildness" of the judges) q = jump fixed effect (as far as I understood it is the "ski jump tradition" of a country, is also explained * as dummy variable, probably nations with ski jump tradition and those without such a tradition??, that * should be explained better) ⊠ = disturbance term (Unclear: where is the actual power to be measured)?*

As the Editor seemed to be satisfied with the explanation of the statistical models, we have only made smaller refinements to the text on this area. We are willing, of course, to provide even deeper explanations than we already give, if need be.

A brief comment: The "jump fixed effect" represents a dummy variable for each jump (each performance), which means that we "trend" the judge scores against their mean. This implies that the model is the "within-jump" estimator of nationalistic biases.

- Formal notes: The following sources do not appear in the bibliography: Meyer & Booker, 1991; Bursell 2012, both indicated on p. 2, are not listed in the bibliography.*

This has been corrected in the revised version.

- Typos on p. 5 Line 5: pscyhological => psychological*

This is corrected in the revised version.

- Overall I recommend major revisions. In general: The limitations of the study should be explained better.*

TODO: In the revised version, we have tried to make the limitations of our study clearer. To this end, we have added … …

Thank you for taking your time to read and commenting upon on our work. In our view that your comments made us produce a stronger manuscript than the original submission, and we hope you concur.

Response to comments made by reviewer #3

We respond the comments made by reviewer #3 below.

- The manuscript is a quantitative analysis of judges' biases and home field advantages on results in ski jumping, based on data from the FIS World Cup in 2006-2008. In general, the subject is relevant and interesting; and authors apply sound methodology and statistics. All formal academic standards are met, and the paper is well written. However, data are rather old, even though they could be easily complemented by more recent data.*

In this version, we use both our original data from the mid 2000s, but also a new data set. This data set was created by scraping the FIS website using automated procedures. In short, we wrote pieces of software that 1. walked through the FIS website and identified URLs pointing to results in PDF files on FIS' server 2. downloaded all these PDFs automatically (and with time gaps to avoid overloading servers) 3. scraped results data from these files according to format specifications.

This procedure was very successful, and let us quickly assemble a rather large data set with results from the 2015-2016 season. The software will be available on the GitHub repository associated with this paper, once the paper is published.

The procedure also has limitations. We exclude some events due to differences in results formats, and include only the first page of results from each file. These minor issues should not affect the results of our statistical analysis in any way. The competitions we include should be considered a random sample of competitions from the season, as the format differences are in all likelihood completely exogenous to the actual results.

> *But most of all, insights are quite limited because the authors follow a solely empiricist approach. Theoretical considerations are missing from the start, and thus the empirical findings are not very intriguing in the end:*

We concur that the original submission was less rich in theory than it was in analysis. Our aim with the study is not to make a theoretial contribution above and beyond testing implications of existing theory drawn from across the social sciences. We have however, inspired by your comments, added relevant theoretical scaffolding for our analysis. In our opinion, this move has improved the mansucript over the original version.

- The literature review is not very extensive and only comprises empirical studies. Hypotheses refer only broadly to some of the literature or seem to be generated ex-post, instead of being deduced from any theory. Not only sociological literature is missing, but there aren't any systematic theoretical considerations, e.g. on the role of judges, on nationalism in sports, or on the specific context of ski jumping

> in comparison to other sports. In the end then, authors face the typical problem of such an empiricist approach: Their results - significant but weak biases, which seem to be addressed quite well by the sport's regulations - do not "speak for themselves", and thus you wonder which insight you have gained. Usually, a more thorough theoretical consideration will also lead to a more insightful interpretation of the results. Instead of treating ski jumping as just one more arbitrary case and speculate about potential threats of biases (p. 20), authors should reconstruct the social context (or field, figuration, system) of ski jumping as a specific case. Then they would be able compare their findings to other sports and reflect upon the meaning of biases in different constellations.*

Inspired by your comments, we have rethought these issues and added some material on the practice of evaluations. Here, we sketch the ski jumping field as a field in a pseudo-Bourdieusian sense and connect the national variation in judge scores to struggles of definition of what good performances are. In particular, we build on the literature on evaluation that has flowed from the seminal work of Boltanski & Thevenon, Lamont and others. We also draw upon the historical example of the Boklöv style, which at first was not taken seriously but over time has come to dominate the field.

Thank you again for your efforts! Your comments made us rethink and add components to the analysis and manuscript that in turn led to a stronger contribution to the literature.

END.