We reply to each of these comments inline with bold text, and appreciate all the advice provided by the AE and the reviewers. We added this acknowledgement to the authors' footnote.

From the AE (we think):

General comments:

With this revision, the authors have significantly improved the paper. However, there are still a few points that are not completely clear. Please see my specific comments for details.

Specific Comments:

1. In the second sentence of Section 3, it may be good to note that part of the alternative hypothesis is that the common judging panel is selected at random. This plays a role in the power calculations described on page 10.

Thank you we made ad addition to clarify this.

2. In the next-to-last sentence of the first paragraph of Section 3, “lower” might be clearer than “more critical”. The last sentence of that same paragraph is awkward and seems to be backwards. If the excluded judges are particularly critical, won’t the difference between the mean for the sub-panel and the mean for the excluded judges tend to be higher rather than lower? It may be a good idea to define the difference metrics more precisely by specifying which mean is subtracted from which mean. Just using the term “difference” leaves ambiguity and may even suggest that an absolute difference is being computed.

We have made the change and tightened the wording to reflect the exact difference used, thank you.

3. Near the bottom of page 8, the authors mention two ways in which ties might be handled. How were ties actually handled in this paper? Since LowExtr and HighExtr will usually be small integer values, there must have been a lot of ties. If the ties were broken at random, how sensitive are the reported p-values to how the ties were broken?

We added a phrase indicating that we used the jittering approach and that this had no impact on the conclusions (see beginning of Section 4). Most of the p-values remained the same, obviously. Although some of the moderate p-values changed in the second decimal place, the results of the study were not substantively changed because none of the small p-values changed much even in the lower decimal places. This is an interesting issue that could be addressed in our paper on the ks-testing, thank you.

4. Just below equation (2), the authors note that they apply “the goodness-of-fit test presented in Gleser (1985)”. It may be good to add a sentence here explaining how the test proceeds.

We added a note that it is a Kolmogorov type of test, which we think is an appropriate level of detail for this paper (e.g. not much detail). We comment on this later, and appreciate all the input from the reviewer on the methodology.

5. Near the middle of page 10, “we reject” might be clearer than “we advocate rejecting”.

Thank you, we made this change.

6. There is some inconsistency in the names of the metrics. The authors use the names “VarDif-
1
fEE ” and “VarDiffPC ” in Section 3, but they use different names in Section 4 and in the tables.

Thank you for catching this.

7. Near the bottom of page 12, the authors suggest that looking for nationalistic bias wouldn’t be useful in this study because such a bias would affect only a few skaters. Isn’t the bigger problem that there is nothing in the data to say which judge is which?

We have clarified the wording on this point, that it isn't useful for answering the question posed in this study. Clearly you are right that a study of bias would be a problem and we couldn't say which judge is which.

8. Near the bottom of page 13, the authors write that “the calculations of exact p-values are only provided for sample sizes of at most 30 because of numerical precision challenges”. However, in the next sentence, they write that for $30 < n < 100$, “exact p-values are used”. Should the second “exact” be “approximate”?

No, the next sentence is for $30 < n < 100$ and no ties in the data; this is really in reference to the default behavior of the R function ks.test() for this particular case. See the next comment, too.

9. I’m surprised to read that the calculation of exact p-values presents challenges for $n > 30$. One good algorithm to use for implementing Gleser’s test is the recursion developed by No´ (1972), who is cited by Gleser (1985). I have used No´’s recursion for sample sizes much larger than 30 without running into any difficulties with precision, and Owen (1995) reports using the recursion for sample sizes as large as 1000. Some of the alternate calculation methods probably do run into trouble, but No´’s recursion is much more stable.

We used one of the methods in Niederhausen (1981), and may have chosen the wrong one! We could not get one of the methods of Steck (1971) working properly (possibly our fault, of course!). We did not try Noe. Our choice of 30 as a cutoff wasn't based purely on numerical challenges, though that was the general ballpark (it depended on the problem). The choice was partly based on a desire not to deviate much from the current behavior of R's ks.test() function, making it more likely that our changes could be added to the R distribution rather than only existing in an extension package. We have a separate paper in submission to the R Journal which focuses entirely on this aspect of the methodology, and the reviewer's many comments have been invaluable for this. Much of this detail isn't necessary in the present TAS paper, of course.

10. On the third line from the bottom of page 13, “sample sample” should be just “sample”.

Thank you!

11. I would find Table 2 more helpful if it provided just a bit more information. Is it possible to add on the sample sizes (28 for the Ladies short program, etc.) and the Kolmogorov-Smirnov distances used in computing each p-value?

We would prefer to keep this as is. We do note in the paper (first paragraph of Section 2) that there were between 20 and 30 skaters per segment. Neither these counts nor the actual values of the statistics add much, we think, and space is at a premium in this table.

References
[1] No´, M. (1972). The calculation of distributions of two-sided Kolmogorov-Smirnov-type statis-
e
tics, Annals of Mathematical Statistics, 43: 58–64.
[2] Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function, Journal of the American Statistical Association, 90: 516–521.

Thank you for the additional references, which we will look at for our paper on the ks.test() work.

From Reviewer 2:

Comments for Authors

You have addressed my previous comments. I have a few, minor revisions to suggest.

1. Thank you for pointing out the combination of tests is across segments (independent), not across metrics (dependent).

2. (p. 2) Not certain the italics are necessary for the events.

We agree and have removed the italics.

3. (p. 10 last paragraph) which −→ that

Changed.

4. (p. 11 first line) the one −→ one

Changed, thank you.

5. (p. 11 line 3) occurrance −→ occurrence

Changed, thank you.

6. (pp. 10-11) The calculation of the six metrics is now clear. Is it worth another sentence or two indicating that
• large values of HighExtr indicate one (or two) of the excluded judges was often the most enthusiastic. I actually think that may be obvious to the reader.
• However, the interpretation of VarDiffPC may not be clear: large

value indicates an excluded judge was highly variable?

We added a final sentence at the very end of section 3 to help clarify this (the reverse of what you stated above, but the point is clear – additional explanation should help).

7. (p. 12 line 12) with −→ within

Changed, thank you.

8. (p. 13 line -7) samples sizes −→ sample sizes

Changed, thank you.