Dear Score editors:

We thank the reviewers for their thorough and helpful reviews for our MMA inter-rater reliability module. We have addressed all comments (bold notes below provide more detail) and believe that the module is much improved. In particular, we have better organized/labeled sections and subsections and better delineated the exercises. We added a conclusion and expanded the introduction to include more motivation and learning objectives.

In addressing the comments, we modified the format(s) of the modules:

* **RMarkdown version – exercises “static”, fill in the blank (with an instructor copy which has exercise solutions)**
* **“learnr” package version; R package to produce interactive exercises.**
* ***NOTE: we have the original submission in ISLE; we did not update our changes in ISLE instead opting for the learnr approach, but the initial submission could be used as a start point if an ISLE version is created.***

*Associate Editor Items:*

1.      Add specific learning goals to the module content.  (There were part of the module submission notes but please add them to the module materials itself.)

* **Added learning objectives in a subsection of the introduction**

2.      Consider adding something to the overview/motivation describing why an organization such as UFC or the Nevada State Athletic Commission might be interested in the consistency of their judges.

* **Expanded the introduction to address the module motivation and objectives**

3.      Provide a conclusion/summary section at the end that ties back to the original motivation.

* **Added a section**

4.      Look over the additional comments and notes that the reviewers have provided including the pdf with notes that reviewer #1 has shared which is attached.

* **Addressed all but a very few of the reviewer comments (notes below for those we did not)**

*REVIEWER #1 Comments (in pdf)*

* **Made all suggested changes (wording etc) and modifications. Some specific notes or exceptions:**
* **Most notably, we better labeled sections to improve flow and consistency. (ie Changed “Overview” to “Introduction” and then made separate sections for “Inter-rater Reliability Overview” and “MMA Overview” to better distinguish (first comment on page 2).**
* **We also clarified which sections were exercises, and numbered questions and removed solutions.**

REVIEWER #2 Comments

*This is just a half-baked thought: fighter 1 and fighter 2 have been randomized in the data, and I wonder if this impacts some of the results, like the simulations. Assuming this data was collected from the official UFC statistics website, the red and blue corners should be known. If so, randomizing fighter 1 and fighter 2 does throw out some information that the judges do have, even if the judges are not supposed to use that information in close rounds. Randomization has yielded an outcome that is balanced – fighter 1 and fighter 2 each win on the judges’ scorecards approximately half of the time; however, we know that the red corner wins much more often that 50% of the time (and I assume this rate holds for decisions but haven’t checked it).*

*At minimum, I believe retaining the red and blue corner as fighter 1 and fighter 2, respectively, would make interpreting the corresponding Fleiss’ Kappa values meaningful and potentially interesting.*

* **RESPSONSE:** ***the measures described are not impacted by choice of which fighter to designate as fighter 1; they are measures of how much the judges agree on a decision and that is not dependent on which fighter is which. We agree that information such as which corner (red/blue) wins more is interesting and, in fact, may be one explanation for judges that are not “reliable” or consistently disagree with others – perhaps some bias exists that they are, for example, more likely to score the red corner higher. However, that is a different question and not in the scope of this module. Note that this is a lengthy module as is!***

*Even though the review questions I was given reference a motivation section, there is no specific motivation section here – though I am not sure if there should be since the Overview section covers it. There also is not a summary-like section at the end that brings the analysis together, makes some sort of sport-specific argument or conclusion, etc. Again, I am not sure if there should be such a section. For what it’s worth, I think some expanded commentary on drawing sport-specific conclusions from the results (i.e. something more than just interpreting a hypothesis test) would help make the module more relevant to the sport.*

* **RESPONSE: Added a conclusion and put more motivation for the module in the introduction section**

"Since judges in MMA score each round, this module is actually exploring the upper bound of inter-rater reliability by using bout-level scores. For example, two judges can disagree on the scores of two of the three rounds in a given bout and still agree on the final score of the bout (e.g. judge 1’s scorecard is 10-9, 10-9, 9-10 and judge 2’s scorecard is 9-10, 10-9, 10-9), and these instances are missed in this module. Consider at least noting this. To make this module more pertinent to the sport, this entire analysis would ideally be performed using round-level scores.

* **RESPONSE: An excellent point – we added a note at the end of the section discussing the data to this effect, and also included this as an important point in the conclusion.**

*By looking across this many years and promotions, this data likely covers multiple rulesets, which may be problematic if certain rulesets are more prone to scorecard disagreements."*

* **RESPONSE: we added a note in introducing the data to address this point.**

*"Overall, the module seemed well written to me. Though I did not specifically check things like grammar, I did find two typos in the MMA Overview section. I don't believe referring to ""The MMA"" is correct.*

* **RESPONSE: Corrected**

*The interpretation of percent agreement and Cohen’s Kappa sound identical to me. The former is: “One simple way is to calculate the percentage of fights in which they agree on the outcome. This is appropriately termed percent agreement.” The latter is: “The kappa value represents the percentage of the two judges’ results that agree with one another.” I suspect the latter may need to be reworded since it is not literally the percentage of scores that agree."*

* **RESPONSE: Corrected as part of reviewer 1 comments; added “over and above what we would expect from chance” to the kappa statement.**