FEI Dressage Judging Working Group

Motivation for, and analysis of, the HiLoDrop rule-change proposal.

The DJWG has recommended to the FEI Dressage Committee the adoption of the following rule change to go before the FEI General Assembly:

Article 434 CLASSIFICATION

HILO DROP SCORING PER MOVEMENT

Recommendation: Adopt HiLoDrop scoring per movement for all FEI dressage competitions, with the exception of Young Horse events and for all juries with 3 judges

Executive Summary

This document details the HiLoDrop method and shows how it would have worked if applied to the last 6 months of FEI competitions.
This proposal only affects the way that final scores are calculated for each Rider-Horse combination. No judges score sheets are changed and the individual judge's results are published just as they are today. The only change is that movement by movement, instead of calculating an overall average the average of the middle scores is used where the top and bottom scores are not included. The final score for the combination is then formed by the usual process of coefficient multiplication and summing to form the final percentage result.
For the majority of cases the effect is tiny, much smaller than the actual precision of a Dressage result. But for cases where one judge is "always" high or low for a given rider, or indeed is very different from that of their colleagues, the effect on the result will be significant and it will then represent more the consensus view with a reduced influence from any single judge.

Key features of the proposal

Why is this proposal necessary?

- The task of judging is a complex and challenging one. It involves comparing a visual observation against the agreed standard and then providing a mark out of ten.
- In an ideal world all judges, seated at similar observation points, would interpret each movement against the set standard of criteria, and any variation of marks would be within a close range.
- This result could be assumed to reflect a strong consensus of the judging panel's opinion.

- Currently almost all stakeholders agree that to become practicable and effective on the field of play the format of the current standard (The Judges Handbook) requires comprehensive review and input.
- It will require time to draft a revised practicable standard, which may require testing and updating judges accordingly, before it can be implemented globally for all FEI international competitions.
- Currently the judging system provides scope for individual judges to prioritise certain tendencies and faults, or reward performances, according to their own personal interpretation and opinion. This impacts on the ranking of not only the affected athlete or NF, but also of those ranked nearby.
- While judges may find consensus when discussing priorities in theory, when it comes to the practical delivery on the field of play, there is evidence that a lack of consensus occurs.
- The frequency of lack of consensus is relatively small but it is varied and as the sport develops globally it is an increasing trend and impacts on a growing number of athletes and NFs.
- The FEI is tasked with providing a global competition ranking system that is fair and transparent to all competitors.

What is being proposed?

- In order to provide the greatest consensus of the judging panel's appraisal of each movement it is proposed that, where a panel consists of five or more judges, then the average mark of those judges who evaluate the movement closest to the standard, or are at least closest in agreement, shall be included in the final mark for the movement. This is achieved by excluding the highest and lowest score per movement.

What impact will this have on athletes and NFs?

- As a lack of judging consensus occurs randomly and at different frequencies this will provide a safety net to those athletes and NFs who find themselves affected by it.
- Evidence shows that a lack of consensus can impact on famous riders at the top of the sport, and their NFs, but it more often impacts on developing athletes and their NFs
- Most of the time judging panels mark each movement within a close range, and therefore for most riders the Hi/LO safety net will have little or virtually no impact on their final ranking.

What impact will this have on judges

- FEI Judges are currently tasked with judging each movement against the agreed standard
- Nothing will change as regards to that task and they will continue to judge in the way they have been trained
- Currently each judge's marks have some mathematical adjustments made to them after they have been awarded. For example, sometimes some marks are multiplied, averaged, or (in the case of Championships) adjusted by an independent panel.
- Judges do not, and should not, currently think of what impact a mathematical or system adjustment will have on their result while they are evaluating each mark on a movement by movement basis.
- Judges do not, and should not, consider what the other members of the judging panel are awarding for each movement, but instead should keep in mind the standard according to the Judges Handbook and other directives in-line with their current training and testing.
- If each judge is in consensus with the other members of the judging panel (which occurs most of the time) then this proposal will have little impact on each judges' final ranking. It only comes into action during periods of lack of consensus.

- Judges who unintentionally penalise (or reward) higher or lower marks per movement or for sections of the test, will be relieved of the resultant negative attention from affected riders, NFs or the media.

Details of the HiLoDrop Proposal

What exactly is being proposed?

Currently the final score of a dressage test is calculated by averaging the scores assigned by all the judges present for each movement, multiplying them by the movement coefficient if any and summing these movement scores. The final score is given as a percentage of the maximum possible for the test. This method is identical to calculating the final score for each judge and averaging these judge by judge final scores.

In the HiLoDrop system being proposed the method of calculating each judge's final score is completely unchanged and each judge final score will be part of the official record along with all their movement by movement scores. But the final score awarded to the competitor will be calculated using HiLoDrop. In this the highest and the lowest awarded marks for a movement are removed and the other scores are averaged to form the movement score. These movement scores are then summed as before and expressed as a percentage of the maximum possible score. All scores awarded by each judge form part of this process and no judge's score is ever changed.
The following table illustrates how it would work for a randomly chosen block of scores from a recent event;

E	H	C	M	B	Average	HiLo	Difference
7.5	7.5	7	7.5	7.5	7.40	7.50	0.10
6.5	6	5.5	6	5.5	5.90	5.83	-0.07
6.5	7	6.5	7	7	6.80	6.83	0.03
6.5	6.5	6	6.5	5	6.10	6.33	0.23
7	7	7	7	7	7.00	7.00	0.00
7	7.5	7.5	7	7	7.20	7.17	-0.03
6.5	7	7	7	6.5	6.80	6.83	0.03
6.5	7	7.5	7	6	6.80	6.83	0.03
6	6.5	7	7.5	6.5	6.70	6.67	-0.03
7	7	7	7	7	7.00	7.00	0.00

Green indicates "Hi" scores
Pink indicates "Lo" scores
No particular score is "removed", just one of the green and one of the pink is not included in the HiLo result. Where all scores are the same, there is of course no change in the final score

Figure 1 Illustration chosen at random from a recent event

A real-life recent example of such a new score sheet applying the HiLoDrop is shown below. (The current movement score is also shown in this example, for comparison purposes)

Movement #	Description	Coef	E	H	C	M	B		Current Movement Score	HiLoDrop Movement Score
1	Collected walk	1	8	7	8	8	7		7.6	7.7
2	Extended walk	1	7	7	8	7.5	7		7.3	7.2
3	Half-pass right (collected trot)	1	7.5	8	8.5	8.5	8		8.1	8.2
4	Half-pass left (collected trot)	1	7.5	8.5	9	8	8		8.2	8.2
5	Extended trot	1	8	7.5	7.5	7.5	7.5		7.6	7.5
6	Half-pass right (collected canter)	1	7.5	8	8	8	7.5		7.8	7.8
7	Half-pass left (collected canter)	1	7	7.5	8	8	7		7.5	7.5
8	Extended canter	1	8.5	8.5	8	8	8.5		8.3	8.3
9	Flying changes every 2nd stride	1	8	8.5	8.5	8	8.5		8.3	8.3
10	Flying changes every stride	1	8	10	8.5	8.5	9		8.8	8.7
11	Canter pirouette right	2	9	9	8.5	8.5	9		8.8	8.8
12	Canter pirouette left	2	9	8.5	9	8	8.5		8.6	8.7
13	Passage	2	9	9	9	9	9		9.0	9.0
14	Piaffe	2	10	10	9	9.5	9.5		9.6	9.7
15	Transitions passage/piaffe/passage	1	9	9	10	9.5	10		9.5	9.5
16	The entrence and halts	1	8.5	10	8	9	7		8.5	8.5
17	Rhythm, energy and elasticity	4	9	9.5	8	9	9.5		9.0	9.2
18	Harmony between rider and horse	4	9	10	10	9.5	9.5		9.6	9.7
19	Choreography. Use of arena.	4	9	9.5	9.5	9.5	9		9.3	9.3
20	Degree of difficulty	4	9	10	10	9.5	9		9.5	9.5
21	Music and interpretation	4	10	10	10	9.5	10		9.9	10.0
	TECHNICAL		**84.250**	**86.250**	**85.500**	**84.250**	**83.500**		**84.750**	**84.833**
	ARTISTIC		**92.000**	**98.000**	**95.000**	**94.000**	**94.000**		**94.600**	**95.333**
	FINAL		**88.125**	**92.125**	**90.250**	**89.125**	**88.750**		**89.675**	**90.083**

Figure 2 Example of the HiLoDrop system applied to Isabel Werth/Weihegold Old, Aachen CDIO5 2017*

The changes to the movement score are typically small and in this case the final awarded score is slightly higher. For example, a 10 and the 9.5 are removed from the Music and Interpretation movement score resulting in an awarded score of 10.0 instead of the previous 9.9.

In the vast majority of cases this is typically what will happen, there will be a small change of awarded score. From a study of 1320 GP level tests with 5 or more judges in 2017 the average score change per rider was +0.09%. 224 scores dropped a little and 1096 scores went up a little. In the Appendix, more details are shown.

As expected and desired, most of the time HiLoDrop does nothing significant to a dressage result. Two-thirds of all result changes are less than 0.2%, and since the intrinsic accuracy of a dressage scores is considerably larger than this, most of the time HiLoDrop satisfies a form of Hippocratic Oath and "does no harm".

HiLoDrop is however extremely powerful at correcting important judging differences that occasionally make ranking changes in key events. In particular the following types of anomaly are completely corrected through HiLoDrop:

1. A Single judge is high or low for the majority of a test for whatever reason.
    a. If the judging is "nationalistic" in nature, pushing a rider up or down, the entirety of these scores will effectively be discounted.
    b. If it is a legitimate opinion that is significantly higher or lower throughout the test than that of the other judges it will effectively be discounted. The awarded score will reflect the consensus of the judges rather than the simple average.
2. "Poor judging", where a judge arrives at the same final score as their colleagues but follows a different route movement by movement. The likelihood is that many of those scores will be discounted by HiLoDrop
3. One judge missing a mistake in a figure (For example miscounted changes), this will typically be corrected by HiLoDrop. (The opposite effect where only one judge

sees the mistake is much less likely, but if it did occur it is true that the "correct" judge's note would be discounted.)

4. The HiLoDrop would have the same effect as the "6%" rule currently in use at JSP invigilated events. But it does not have a sharp cut-off at any fixed score such as 6%. This is much less likely to artificially disrupt a key result.

The use of the average of 5 or 7 judges is just one way to calculate the score to award. From a purely statistical point of view the average is known to be a biased estimate of the "correct" value when there are only a few measurements (as is the case with 5-7 judges for each movement). Better measures are, for example, the median (or central) score. However, the median method fully removes all but the central value and surely some information is lost.

Our goal in proposing this rule change is to establish for dressage a system that is in common use in other sports, including other equestrian disciplines. It is easy for the public and rider to understand and reduces the impact of judges with a divergent mark on the score of each movement. We fully recognize that sometimes this will discount a score that may after full analysis be considered to have been the most correct. That score will remain on the score sheet but the awarded final score will reflect the general consensus at the time of the test.

Only by education, training and experience, or a change in judging system, can that consensus be more often the perfect correct result we would wish for, but HiLoDrop will ensure that the general consensus is more often the one that is used to award the final score and to determine the final ranking.

HiLoDrop does not tell judges what to award. As today, a judge who sees an exceptional execution or a mistake is encouraged to award the score that they believe is the most correct. Their score will stand and can be discussed as today with other judges after the event and can be understood by the rider in the same way as today. No judge's notes are ever actually discounted from the final score in totality unless they are high or low for every figure. There is no reason for judges to be more conservative in their judging. If two judges give a 10 or two judges give a 4 then the 10 or 4 will strongly influence the final result. Just as today, judges must award the score they believe in. Most riders will actually receive a boost in their final score.

HiLoDrop favours the consensus opinion of the judges whilst reducing the individual influence of a single judge on the final score. The changes introduced due to differences on a single figure are typically immeasurably small, even a 4-point difference for a single figure, today changes the final score by just .001%. It is only in the case of a judge being very different for the majority of the test that any significant change in the awarded final score will be visible. The goal is not to change the important role that multiple judges and judges position brings to the sport, but just to arrive at a more representative final result for prizes, medals and public understanding.

Summary

- For the majority of cases, scores will change very little, although on average there will be a small increase in final scores
- A judge unsure if an error has occurred can confidently give the score corresponding to 'no-error', knowing that if their colleagues have seen the mistake then it will be correctly penalized in the final result
- No single judge can ever determine the final ranking, the consensus result will be the determining factor
- Nationalistic or other biases, deliberate or not, will always be removed from the final score, it is impossible for a single judge to push a rider up or down compared to their colleague's appreciation

- The system will act in a similar way to the JSP now, but for all FEI events, not just a select few. The 6% rule would become redundant.
- Judging difference such as at Herning in the CH-EU that very nearly changed a medal, will no longer pose any issues in the final result.

Appendix : Analysis of the effect of HiLoDrop on all competitions in the first 5 months of 2017.

As noted above the average change in awarded score is +0.09%. The standard deviation of this difference is 0.21%. 17% of results are adjusted down by a small amount, 83% are adjusted up. Four out of Five riders receive a small boost from HiLoDrop. The actual distribution of score changes is shown below

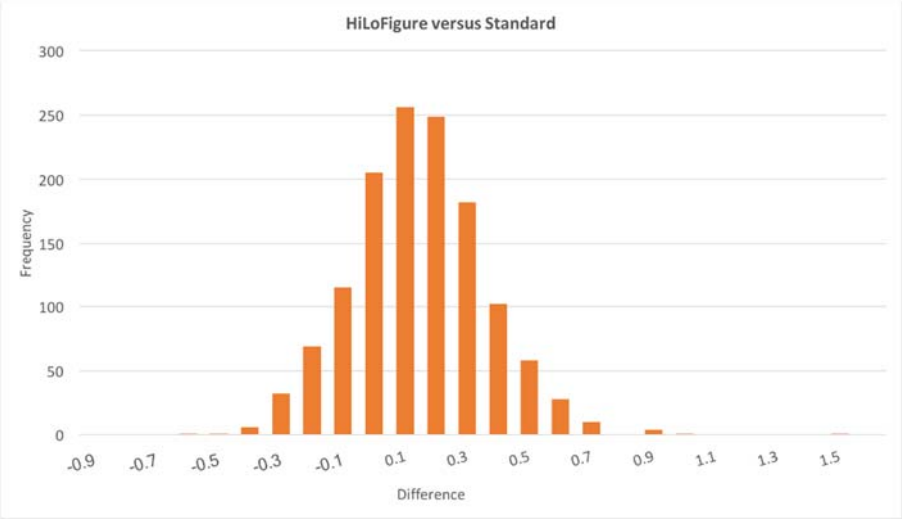

Figure 3 HiLoDrop final score minus the Standard Final Score (GP level events)

The largest difference in this sample was +1.47%, where the judge's final scores were: 57.4, 64.3, 68.0, 66.0, 68.3. The awarded score in the current system was 64.800 that would have been 66.267 with the HiLoDrop system

Comparing the standard and HiLoDrop score shows that they are perfectly correlated with each other

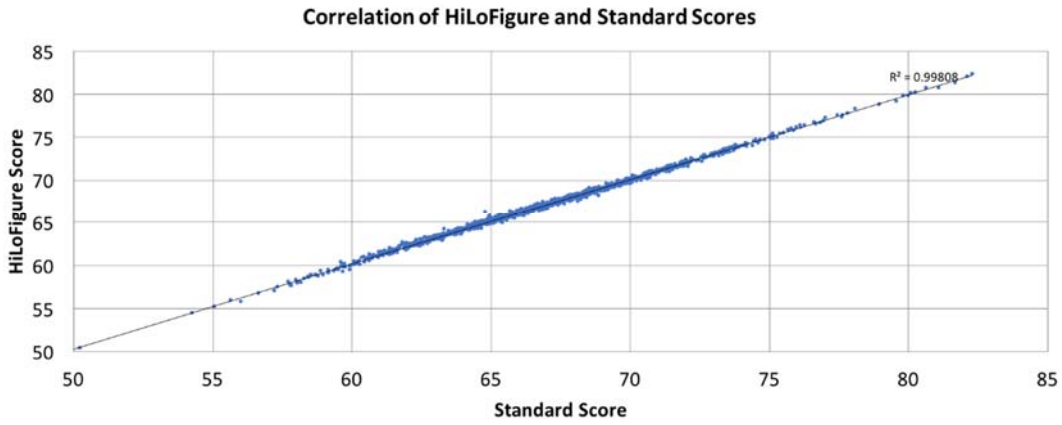

Figure 4 Comparison of Standard and HiLoDrop score (GP level tests)