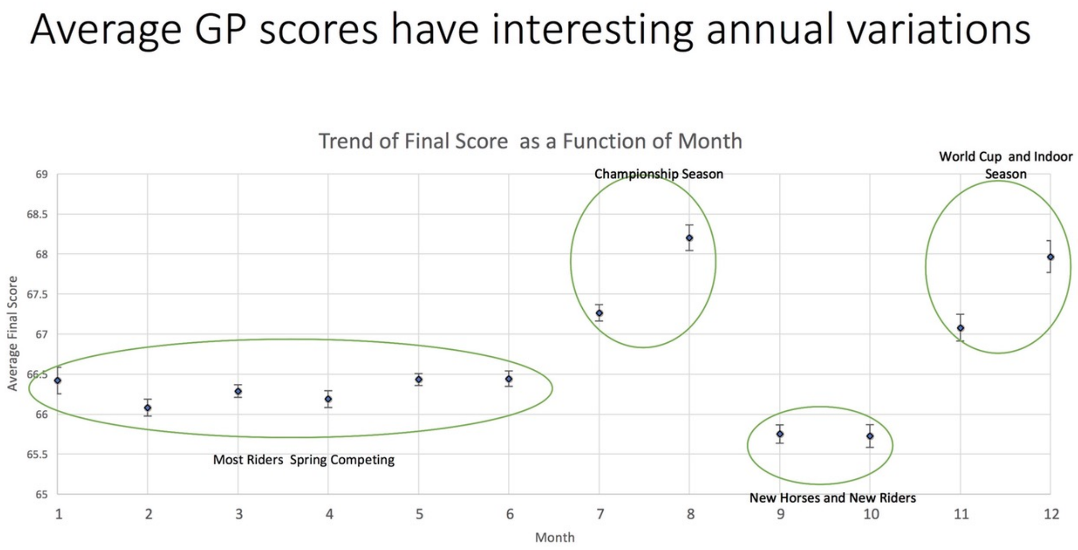
GDA Dressage Facts News

Day 1 April 19

Sorry for the long delay since a posting here. I decided to start a series of postings on Dressage Facts (as opposed to Alternative Dressage Facts) The first one in the series is just a fun one:-) Will post one a day for a few weeks... Please share with your friends if you find them interesting. David
This is just a compilation of International GP scores per month over the past 6 years, I think there is an interesting pattern through the year. the comments are just my guesses as to the possible causes of the patterns. feel free to comment with your interpretations



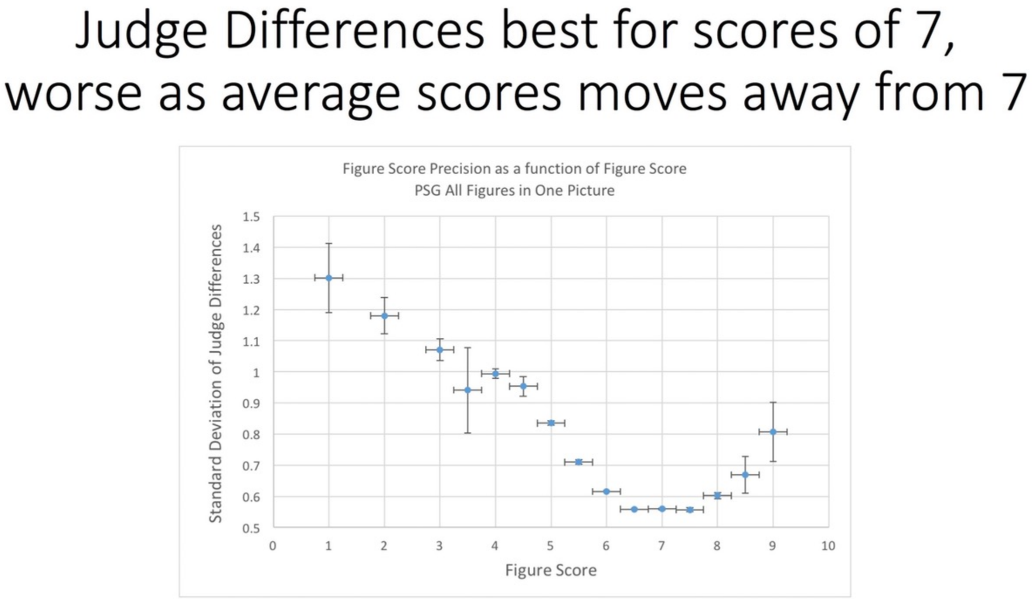
Day 2 April 20

Day 2 Dressage Facts.
In the past week two different Eurodressage articles have used a GDA plot and made what look like two different conclusions from it. (See Wayne Channon and Hans-Christian Mathiessen articles) This plot shows the agreement between judges in Prix St Georges as a function of the final movement score. It tells us that as the score moves away from 7 the agreement becomes progressively worse. Hans-Christian correctly notes that for the majority of scores the agreement is "comfortable", Wayne correctly notes that as the scores move away from 7 the judges start to disagree more. They usually agree they saw an error or an excellence, but they are not so clear as to how much numerically to punish or reward it.

The plot shows the facts but the interpretation can be yours, largely ok or a failing we should try to ameliorate, or both?

http://www.eurodressage.com/.../wayne-channon-who-moved-my-ch...

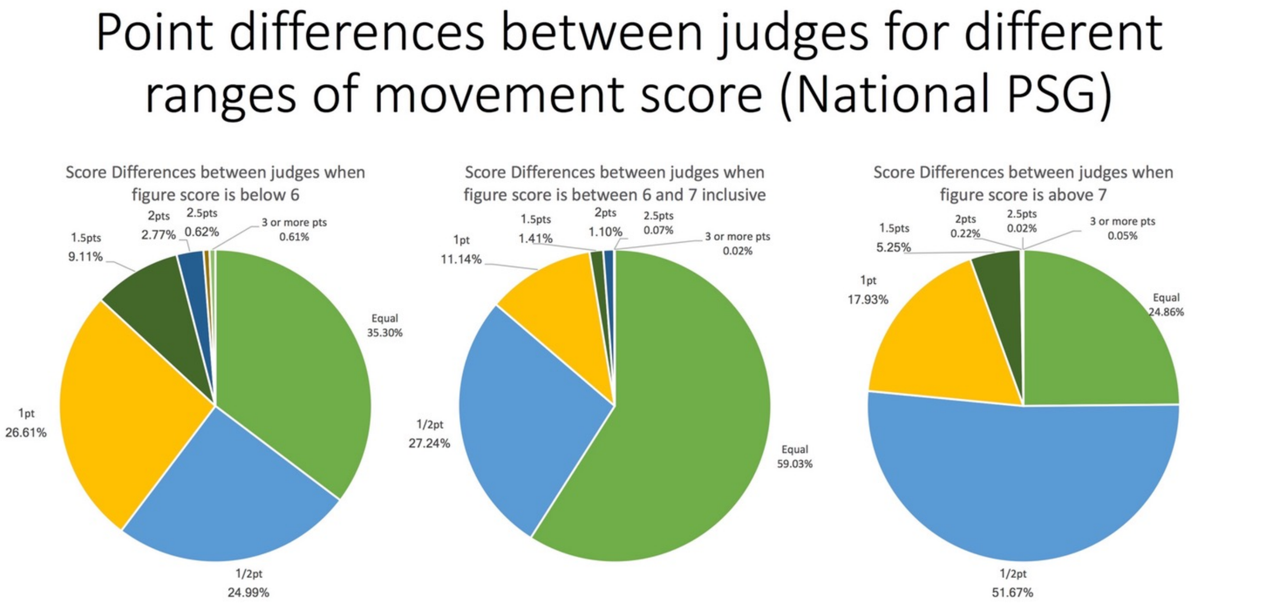
http://www.eurodressage.com/.../hans-christian-matthiesen-per...



Judge Differences best for scores of 7, worse as average scores moves away from 7

Day 3 April 21

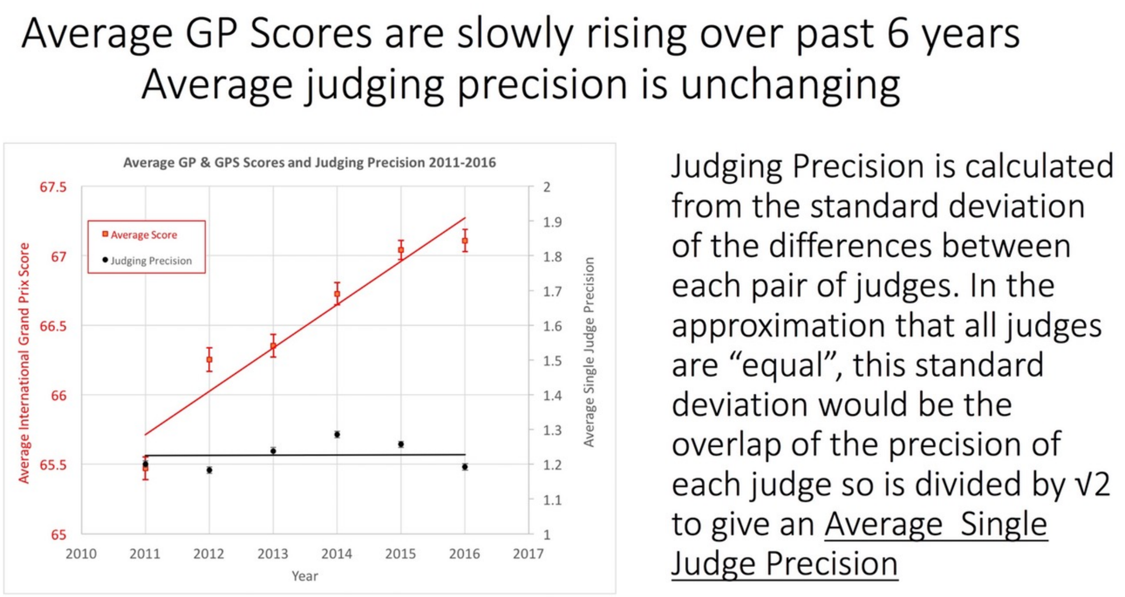
Day 3 Dressage Facts. A continuation of yesterdays theme. Here I show for 3 different score ranges: below 6, 6 to 7 and above 7, how often the judges have the same note, or 1/2 point difference, or 1 point etc. When the score is between 6 and 7 then 60% of the time the judges give the same score, but outside that range its more like 30% of the time. When the score is between 6 and 7 the judges differ by more than 1 point 2.6% of the time, but for lower scores they differ by more than 1 point 13% of the time, that is 5 times more frequently than for the 6-7 range. Above scores of 7 the judges disagree by 1/2 point more than 50% of the time so the most likely scenario is a disagreement, albeit small.
1.5 points or more of difference is not insignificant when we have a 0.5 point scale, so I conclude it would be advantageous to find ways for judges to be more precise in these "abnormal" situations. Of course a legitimate debate is whether this is an education and training issue or a systemic issue with the current judging system.



Point differences between judges for different ranges of movement score (National PSG)

Day 4 April 22

Day 4 of Dressage Facts. This plot was made by Clarissa Stickland and was published in Horse and Hound magazine on February 23rd. In it Clarissa shows how there has been a steady increase in average International GP scores over the last 6 years. She also shows that judging precision (Here defined by using the differences between judges final scores) has remained constant over the last 6 years on average.
An interpretation of this could be that judging precision is at some natural limit of the system we use at present; we could accept that or we could decide to work on the system.

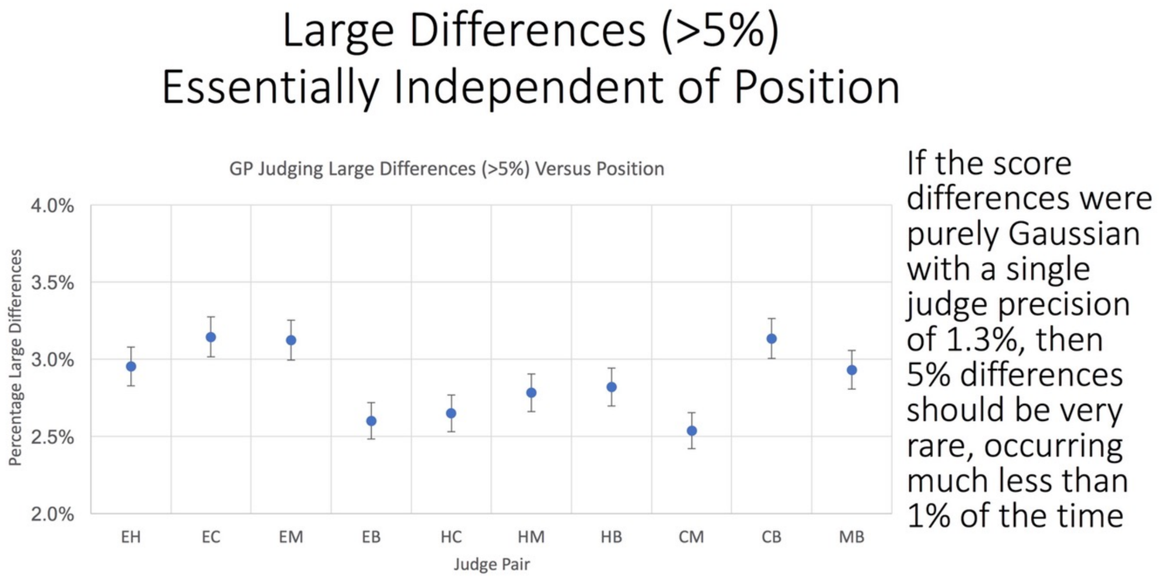


Average GP Scores are slowly rising over past 6 years
Average judging precision is unchanging

Judging Precision is calculated from the standard deviation of the differences between each pair of judges. In the approximation that all judges are "equal", this standard deviation would be the overlap of the precision of each judge so is divided by √2 to give an Average Single Judge Precision

Day 5 April 23

Day 5 of Dressage facts. It is certainly true that some mistakes can be seen very differently from different judging positions, it makes sense to have judges around the arena and it is to be expected that they will differ with each other for certain movements. The final GP centreline for example is clearly different as seen from the judge at C compared to those at E or B. In this picture we show the percentage of what I call Large Deviations (Where judges disagree on a final note by more than 5%) for various pairs of judging position in CDI Grand Prix with 5 judges. It is true that EB judges agree with each other a little more frequently than EC or CB. But the real problem in this plot is that the percentage of large deviations ought to be considerably less than 1% based on the Gaussian Single Judge Precision of 1.3% (Dressage Facts Day 4), wears actually it happens about 3% of the time. Since there are 10 judging pairs in a 5 judge jury this reminds us that in about 30% of all rides two judges will differ by 5% or more. This is telling us that there is a non-Gaussian component to judging, and it is not due to the judges position.



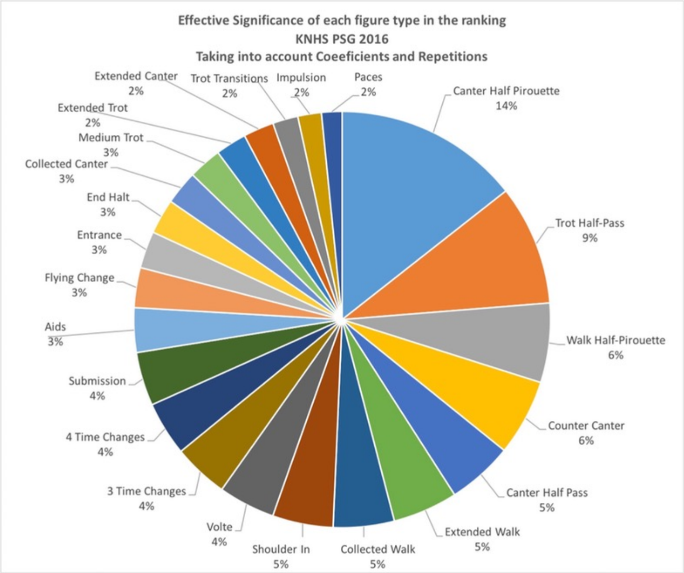
Large Differences (>5%)
Essentially Independent of Position

If the score differences were purely Gaussian with a single judge precision of 1.3%, then 5% differences should be very rare, occurring much less than 1% of the time

Day 6 April 24

Day 6 Dressage Facts. Today information for riders. In the picture you see how each movement in the Prix St Georges contributes to the final ranking. What does that mean? If everyone gets the same score for a movement then it contributes nothing to the ranking. So in this chart each slice of the pie has a size proportional to the spread in scores (actually the standard deviation of scores for each movement) taking into account the coefficients where relevent. Not surprisingly the canter half-pirouettes are the most decisive, perhaps surprisingly the next most important are the trot half pass, significantly more important than the change lines.
Tomorrow we will investigate how the judging system enters into this equation.

Importance of Each Figure in the Ranking

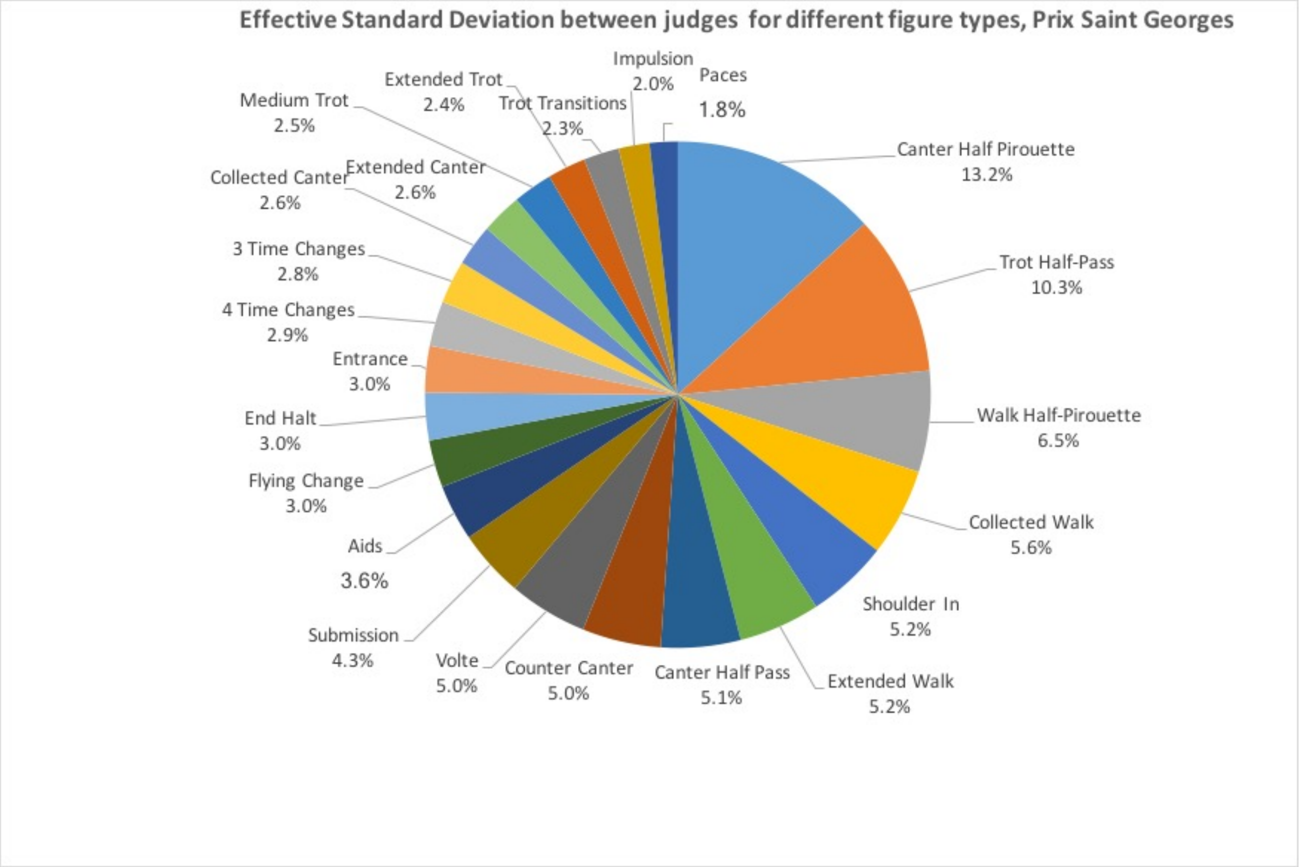


Day 7 April 25

Day 7 Dressage Facts. Yesterday I showed which figures have the most variation between riders, and hence have the most effect on the rider ranking. Today I show the same Prix St Georges movements but this time each slice of the pie is proportional to the variation between the judges - the standard deviation of judge differences. The main features of the two charts are the same. Now this does take into account repetitions and coefficients, so it is going to highlight the figures that are shown twice and have double coefficients, so this may be an important contributor to both the rider variation and the judge variation.

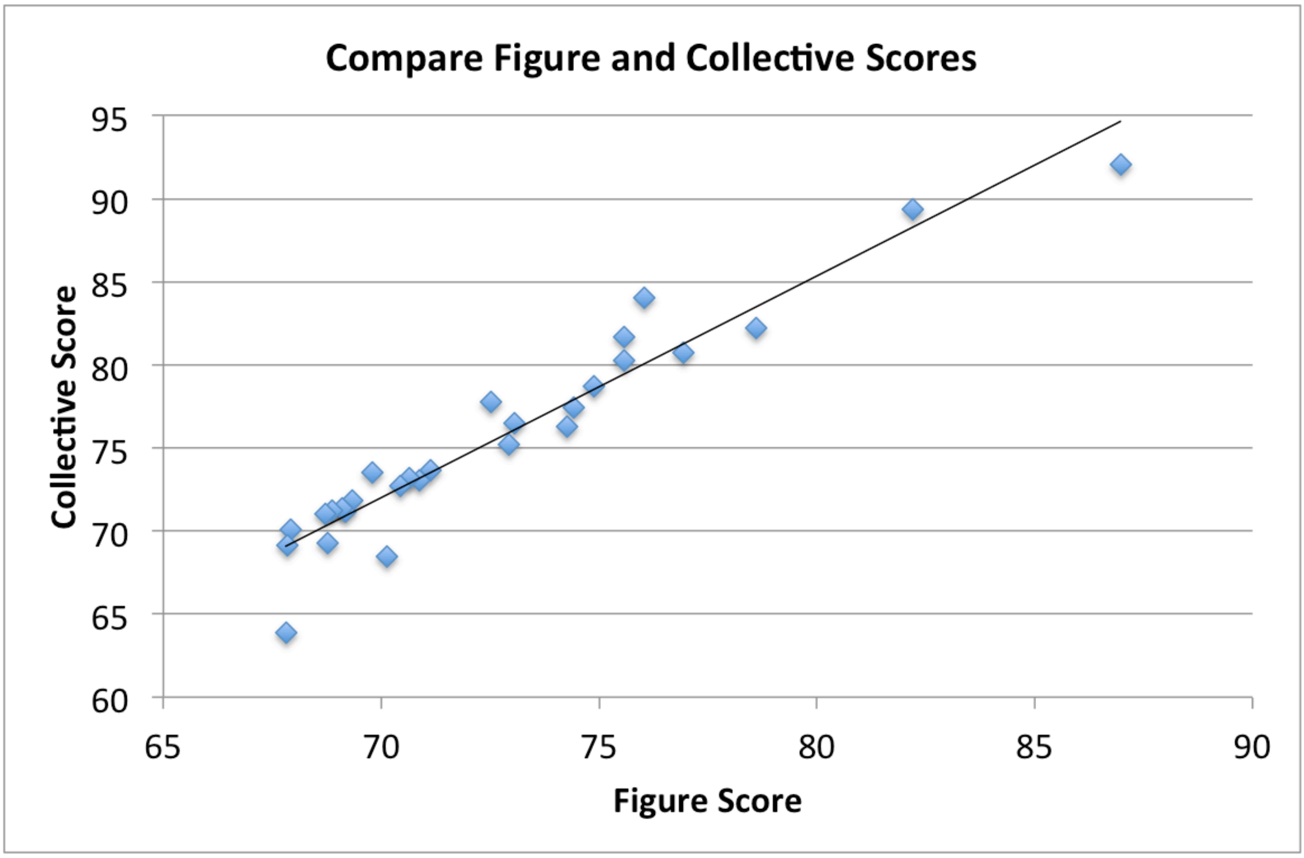
Nevertheless, the figures that most separate the riders are also the ones that the judges have the most difficulty to agree on.

This is probably closely related to the observation on Day 2 that when movement scores move away from the 6-7 range then judges have more difficulty to agree, the fact that these figures have a wider score range will inevitably also mean that they will populate the left and right hand tails of the Day2 plot. So the Day 2 message is perhaps still the key to success: a more clearly defined and or a more finely executed award of penalties and reward for scores outside the 6-7 comfort zone? As always your comments and different perspectives are very welcome



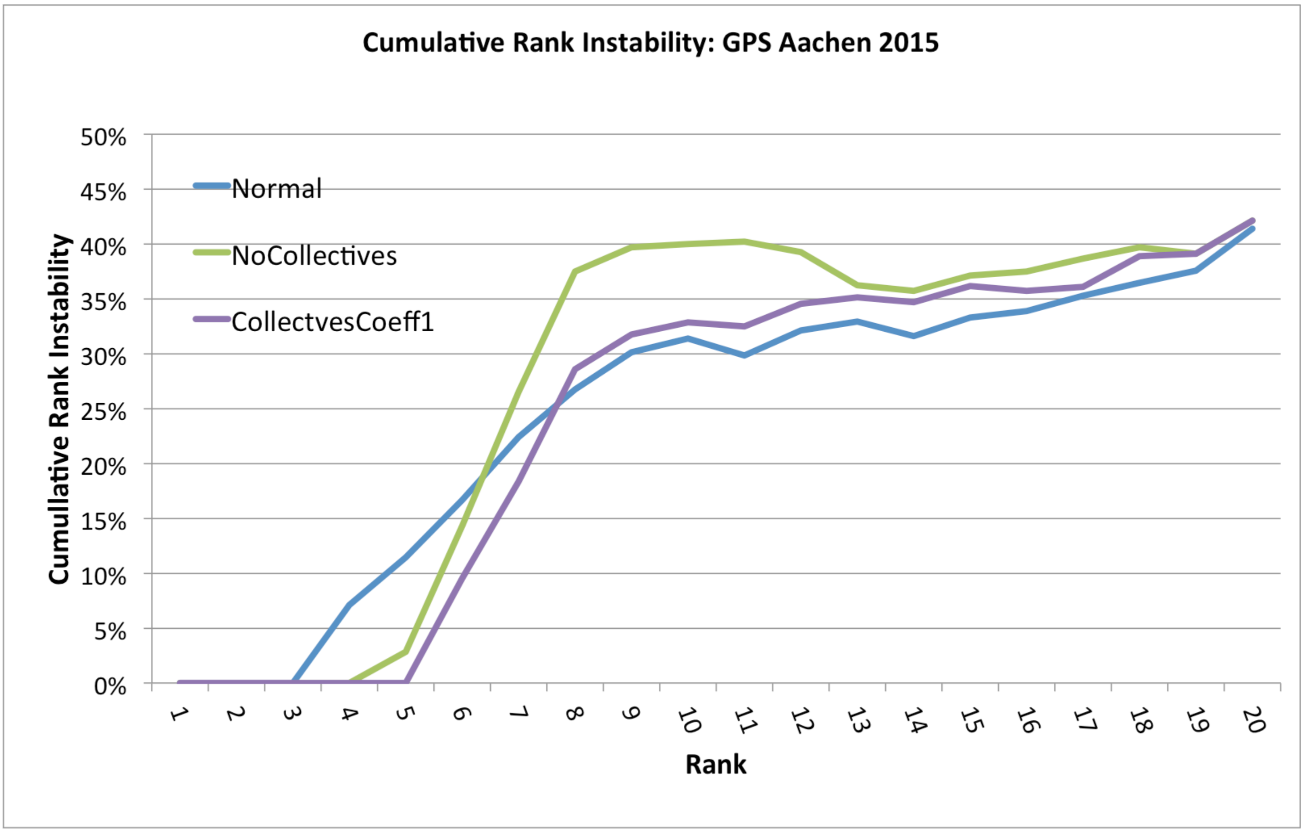
Day 8 April 26

A comparison from a major Grand Prix event plotting the technical mark versus the collective mark. Not surprisingly you see they are highly correlated. So, does that mean they are useless, as they add no new real information to what has already been judged, or does it mean they are not an important issue, as they don't really change anything? I'm pretty sure it does mean that either way we could live without them. In tomorrows Dressage Fact we will look at their effect on the ranking



Day 9 April 27

Day 9 Dressage facts
In todays post we introduce a new measure called Rank Instability. Basically it comes from calculating how many changes of rank would occur if one removed one judge at time from the jury. Put another way, it is the effect of each single judge on the ranking. The higher the instability the more subject to a single judge in the jury the final ranking is. I think its a useful way to measure the effect of new scoring algorithms. In this case we compare the rank instability for a normal, in this case 7-judge, jury with what would happen if there were no collective marks and what would happen if all the collective marks were coefficient 1. In the middle of the field then rank stability is never very small, rides are close in score and so its very easy for one judge to tip the balance up or down. In the top three ranks all algorithms are insensitive to dropping a judge. But we see here that with either no collectives or coefficient 1 collectives the 4-7 positions have less rank instability than in the normal system. So one answer to yesterdays question could be that the collective marks are largely inoffensive but they do increase the sensitivity to a single judge changing the ranking. So perhaps they should be discontinued, or maybe just have their overall weight reduced (coef 1).
It is difficult to come up with objective ways to measure new scoring algorithms, It's easy to see it changes something but hard to know if it's better. I think that rank Instability is a potentially useful measure



Day 10. April 28 A summary and a brief pause

There is a lot of information in Dressage scores, much more than we usually imagine. Observations are what they are, but of course they can be interpreted in different ways when trying to dig down to the underlying reasons for them. Obviously judge education and training is very important and must always be pursued.But I think the evidence points to the conclusion that we fundamentaly have a lack of precision in defining penalties and rewarding exceptional performance, so I welcome ideas that might improve that. It is very easy for even the best judges in the world to head off down a different route to their colleagues and end up 5% or more different, we should try to reduce the likelihood of this or make it more clear why it has happened when it eventually does

We need to decide what we want; perfection in judging will not happen and each method will fail in various ways. Is the average opinion the most right? If one judge is very different they can be the most correct, but that surely isn't always the case. Protect the opinion of the individual or arrive at the most likely consensus of the group accepting that sometimes that will be wrong, but which type of "wrong" happens most frequently?

Ten Day Fact Summary

- Judging differences are smallest for scores in the range 6-7 and larger for scores outside this range
- Judging Precision is more or less constant over the past 6 years, and about 1.3% per judge
- Differences between judges of 5% or more occur about 3% of the time but from the 1.3% single judge precision they would be expected to occur more like 0.5% of the time
- There is a small variation in large differences depending on the judges position, but it is not the dominant source of these differences
- Judge variation (in the PSG at least) is biggest for the very same movements that also influence most significantly the ranking
- Collective marks are highly correlated with technical marks, so they do not typically have a large effect, but doing away with them does reduce rank instability

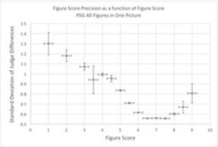
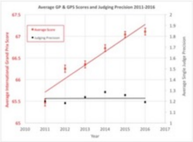
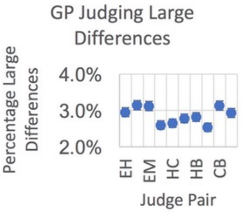
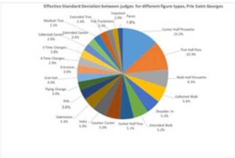
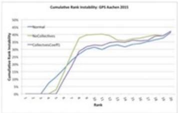


Figure Score Precision as a function of Figure Score
PSG All Figures in One Picture
Standard Deviation of Judge Differences
Figure Score


Average GP & GPS Scores and Judging Precision 2011-2016
Average Score
Judging Precision
Average International Grand Prix Score
Average Single Judge Precision
Year


GP Judging Large Differences
Percentage Large Differences
EH
EM
HC
HB
CB
Judge Pair


Effective Standard Deviation between Judges for different figure types, Prix Saint Georges
Impulsion 2.0%
Paces 1.8%
Extended Trot 2.4%
Trot Transitions 2.3%
Medium Trot 2.5%
Collected Canter 2.6%
Extended Canter 2.6%
3 Time Changes 2.8%
4 Time Changes 2.9%
Entrance 3.0%
End Halt 3.0%
Flying Change 3.0%
Aids 3.8%
Submission 4.3%
Volte 5.0%
Counter Canter 5.0%
Canter Half Pass 5.1%
Extended Walk 5.2%
Shoulder In 5.2%
Collected Walk 5.6%
Walk Half Pirouette 6.5%
Trot Half Pass 10.3%
Canter Half Pirouette 13.2%


Cumulative Rank Instability: GPS Aachen 2015
Normal
NoCollectives
CollectivesCoeff1
Cumulative Rank Instability
Rank