A Guide to Category Weighting

Everyone wants to know how much each category "counts" in the FPA judging system. The answer is: It depends.

The weight of each category is the spread of that category divided by the sum of the spread of all categories.

$$Weight\ of\ CategoryX = \frac{Spread\ of\ CategoryX}{Sum\ of\ Spreads\ in\ Each\ Category}$$

The spread of each category is determined by the difference between the highest and the lowest score in that category.

$$Spread\ if\ CategoryX = HighestCategoryXScore - LowestCategoryXScore$$

Suppose there are three categories—Difficulty, Artistry, and Execution—scored between 0 or 10.

Each of these categories has the same *potential* weight. Each team can score up to 10 points in that category. But the *actual* weight of each category depends on how the judges score the round.

Example 1: Suppose there are three teams with the following scores:

Team	Difficulty	Artistry	Execution	Total
Blue	7	4	2	13
Red	8	2	9	19
Green	2	3	1	6

The *spread in each category* is calculated:

- Difficulty: 8 – 2 = 6
- Artistry: 4 – 2 = 2
- Execution: 9 – 1 = 8

The *total spread among all categories* is: 6 + 2 + 8 = 16.

We can then calculate the weight of each category by dividing each category spread over the total spread:

- Difficulty: 6/16= 37.5%
- Artistry: 2/16=12.5%
- Execution: 8/16: 50%

Of course, if the judges scored the round differently, the weight of each category changes.

Example 2: Suppose the teams scored as follows:

Team	Difficulty	Artistry	Execution	Total
Blue	5	0	7	12
Red	5	10	9	24
Green	5	3	7	15

The spread in each category in Example 2 is calculated:
- Difficulty: 5 – 5 = 0
- Artistry: 10 – 3 = 7
- Execution: 9 – 7 = 2
- Total Spread: 0 + 7 + 2 = 9

The weight in each category is calculated:
- Difficulty: 0/9= 0%
- Artistry: 7/9=78%
- Execution: 2/9: 22%

In this case, because every team received the same difficulty score, difficulty did not have any weight. In other words, no team gained any advantage over the other teams in the difficulty category. Of the 9 points separating the teams, none of them came from difficulty.

The sum of the category spread does not equal the team score spread.

The spread between the highest scoring and lowest scoring *team* is *not* the same as the sum of the spread among all categories. In Example 1, the spread between the highest scoring team—Red Team with 19—and the lowest scoring team—Green Team with 6—is 13 (not 16).

This is not relevant to calculating the weight of each category. Suppose the following scores:

Team	Difficulty	Artistry	Execution	Total
Blue	5	0	0	5
Red	0	5	0	5
Green	0	0	5	5

Each Team scored 5 in one category and 0 in the other categories. Each team's total score is 5 and the spread between total scores is 0. The sum of the spread of each category, however, is 15, with each category having equal weight.

Note on Negative Scores

The presence of negative scores does not affect the calculation of category weight. Suppose the execution scores in example 1 were negative.

Team	Difficulty	Artistry	Execution	Total
Blue	7	4	-2	9
Red	8	2	-9	1
Green	2	3	-1	4

The *spread in each category* is calculated:

- Difficulty: 8 – 2 = 6
- Artistry: 4 – 2 = 2
- Execution: -1 (the highest value) - -9 (the lowest value) = +8

The total spread among all categories and the weight of each category is the same.

The *total spread among all categories* is: 6 + 2 + 8 = 16.

We can then calculate the weight of each category by dividing each category spread over the total spread:

- Difficulty: 6/16= 37.5%
- Artistry: 2/16=12.5%
- Execution: 8/16: 50%

The "proportion" of a category to a team's score is not relevant.

The proportion of a category score to that team's total score does not affect the weight of that category. Put another way, it does not matter if one category (i.e., difficulty) is 90% of a team's total score.

In the second example above, the proportion of the category scores to the team's total score would be calculated as follows:

Team	Difficulty	Proportion	Artistry	Proportion	Execution	Proportion	Total
Blue	5	5/12=41.6%	0	0/12=0%	7	7/12=58.3%	12
Red	5	5/24=20.8%	10	10/24=41.6%	9	9/24=37.5%	24
Green	5	5/15=33.3%	3	3/15=20%	7	7/15=46.6%	15

These proportions do not tell us anything about the ranking of these teams. For the Blue Team, artistry was 0% of their total score, but artistry the category counted for 78% of the outcome of their ranking.

These proportions are also unique to each team. Although artistry accounted for 78% of the difference among *all* teams, for the Blue Team specifically, artistry was 0% of their total score. For the Red Team specifically, artistry was 41.6% of their total score.

These proportions also have **no relationship** to the calculation of the category weight. Suppose that only difficulty scores were calculated between 5,000 and 5,010, and the teams received the following scores:

Team	Difficulty	Artistry	Execution	Total
Blue	5,007	4	2	5,013
Red	5,008	2	9	5,019
Green	5,002	3	1	5,006

The *proportions* of each category score to the teams total score would be calculated as follows:

Team	Difficulty	Proportion	Artistry	Proportion	Execution	Proportion	Total
Blue	5,007	99.9%	4	0%	2	0%	5,013
Red	5,008	99.9%	2	0%	9	0%	5,019
Green	5,002	99.9%	3	0%	1	0%	5,006

Nevertheless, the *weight* of each category score would be the same as the first example because the spreads are identical.

The spread in each category is calculated:

- Difficulty: 5,008 – 5,002 = 6
- Artistry: 4 – 2 = 2
- Execution: 9 – 1 = 8
- Total Spread: 6 + 2 + 8 = 16

The weight of each category is calculated:

- Difficulty: 6/16= 37.5%
- Artistry: 2/16=12.5%
- Execution: 8/16: 50%

Accordingly, *all that matters to the weight of a category is the spread of the category compared to the sum of the spread of all categories. The proportion of that category to the team's total score is irrelevant.*

<u>But does the potential spread of categories matter?</u>

The potential spread of a category is at least theoretically relevant to how much weight that category will have. *All else being equal*, two categories with the same potential spread are more likely to have the same actual weight. For example, if there are two categories, execution and artistry, each with a potential spread of 0-10, they are—again, *all else being equal*—more likely to have the same weight.

On the other hand, if one category had a potential spread of 0-5000 and the other had a potential spread of 0-10, the categories are—*all else being equal*—more likely to have different weights. If, for instance, judges gave artistry scores as low as 0 and as high as 5,000, the weight of the artistry category would be ≈99% regardless of how execution is scored.

However, different categories tend to be judged differently, and some categories with the same potential spread are calculated in ways that cause the scores to have typically narrower or wider spreads. For instance, under the old FPA judging system, difficulty scores were calculated by taking the *average* of difficulty scores awarded in time blocks, while execution scores were calculated by *adding* together execution deduction. Averaging had the effect of creating narrower difficulty score ranges (5-6), whereas adding had the effect of creating wider execution deductions (0-10). Accordingly, even though both categories had the same potential spread (0-10), the average *actual* spread of the two categories was very different. (On the other hand, two categories may have different potential spreads, but similar actual spreads.)

How can we "balance" the weight of each category?

Normalization

There is only one method for ensuring the weight of each category is consistently the same: normalization. But even normalization can be imperfect.

Normalization involves adjusting the value of raw scores to equalize the spread between two categories. There are different ways to achieve this: either by requiring the judges to use the entire range or through mathematical formulas that equalize the raw score spreads in each category.

The main problem with normalization is that it can cause small differences in routines to have significant impact on the rankings. Suppose the judging system includes two categories, execution and artistry, and the execution category is judged by providing a deduction each time the disc touches the ground. Suppose then that each team is "dropless" and the disc never touches the ground. In that instance, there is: (1) no meaningful way to mathematically normalize their execution score; and (2) no normative reason to distinguish the teams in the execution category because each team performed equally well. Simply put, it is appropriate for execution to have no weight in that round even if the judging system otherwise seeks to weigh execution and artistry equally.

Similarly, suppose that one team went dropless and every other team had one drop. After normalization, the dropless routine receives 10 execution points while all other teams receive 0 execution points. In most scenarios, that would result in a small difference in routines—a single drop—having an outsize influence on the team rankings. Indeed, although in a mathematical sense, normalization would result in execution having "equal" weight as other categories, in a practical sense, a very minor execution error would have far more weight than any other component in the routine.

Example: Normalized scores can make small differences in routines very significant.

Raw Scores

Team	Artistry	Execution	Difficulty	Total	Rank
Blue	1	9	1	11	2
Red	9	7	0	16	1
Green	1	8	1	10	3

Normalized scores

Team	Artistry	Execution	Difficulty	Total	Rank
Blue	0	10	10	20	1
Red	10	0	0	10	3
Green	0	5	10	15	2

Scalars

Scalars are another tool to at least help balance the weight of categories in a judging system.

Suppose, for example, there are two categories: execution and artistry. Both are worth up to 10 points. But on average, execution scores range between 0 and 9 and artistry scores range between 5 and 6. On average, the sum of the spread of each category is 10, and 90% of that spread—that is, 90% of the weight—is caused by execution.

If the goal of the judging system were to value execution and artistry equally, a scalar could be used to make artistry count more relative to execution. In this example, if artistry scores were multiplied by 9, the average spread in artistry would be 9—the same as execution—and therefore the weight of artistry, *on average*, would equal execution.

Using scalars to balance categories to desired weights requires analyzing real-life results to determine what scalar is appropriate. The process is imperfect for many reasons. For one, as explained above, the weight of each category changes each round and often for legitimate reasons. Every round is determined by a different combination of categories based on how the teams play—the weight turns on what categories the teams distinguish themselves in. In addition, there is high variance in how judges judge, which means there is high variance in the spread in each category. Finally, judges often react to changes in the judging system that amplify or negate the intended effects of scalars or other tools—including judging education—meant to balance category spreads.

Simply put, scalars rely on understanding the average weight of each category, but those averages are moving targets.