What the Heck do My Scores Mean?

A Guide to MN FLL Scoring

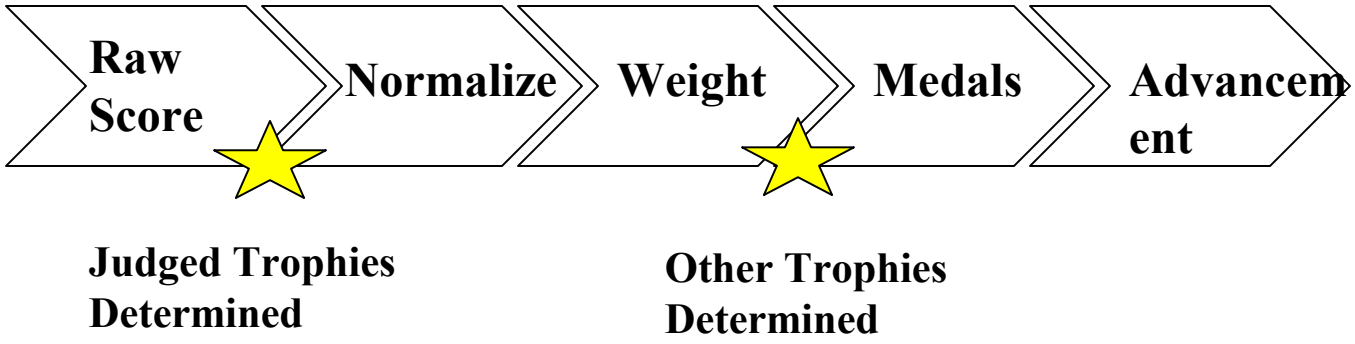
Fred Rose
Head Official, MN FLL
November 30, 2003
(updated 6/9/2006)

Overview

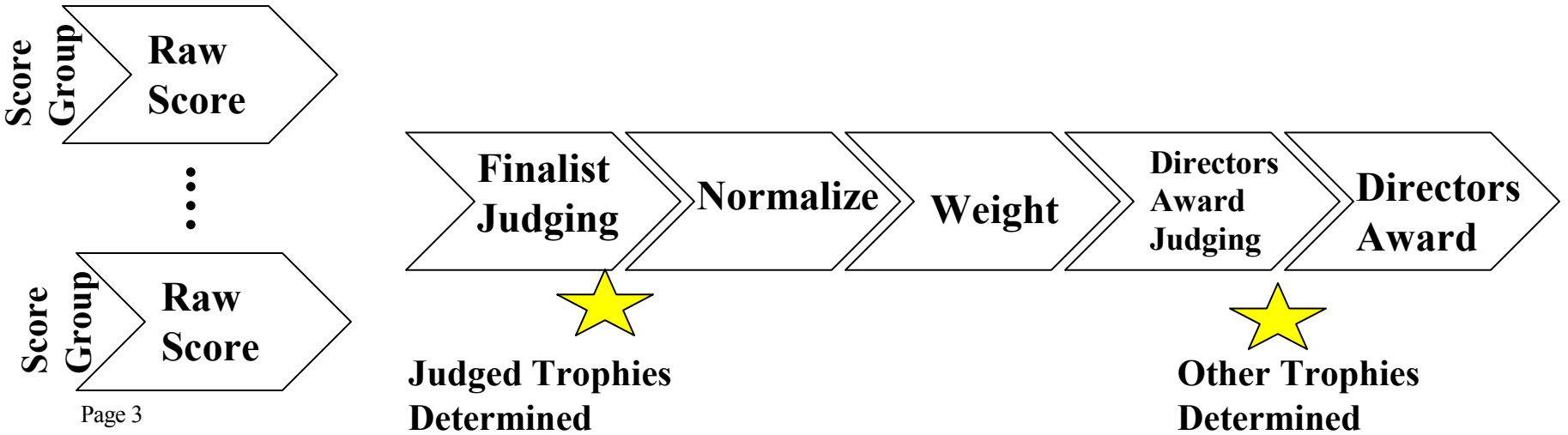
- Teams are scored on
  - Their results on the actual Mission Mars challenge. This is called the Performance Score
  - Technical Judging, which includes Robust Design and Programming. Teamwork may also be judged at Technical judging, if there are enough judges.
  - Presentation Judging, which includes Teamwork and Research Presentation.
- These 5 scores are combined into a Master Score
- This Master Score determines advancement to state and medal color for participants.

Scoring Process

Regional Tournament Process

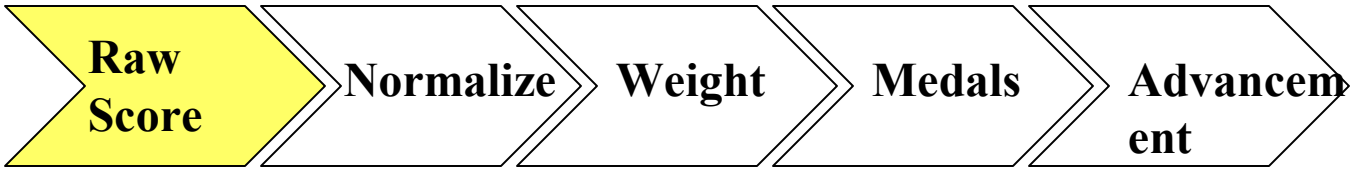


State Tournament Process



Raw Scores

- Raw Score refers to the actual score, either received on the performance table or from a judge.
- Performance:
  - In Mission Mars, a perfect score would be 400. The best score from one of your first three matches (not the head to head at the end of the day) is used.
- Judging:
  - In each of the judged categories, you will receive a score between 0-100 from the judge.
  - Don't worry too much about the absolute value of this score, judges are generally comparing teams, so it's how that score compares to the other teams that is important.
- Subjective judging allows in depth analysis of teams, rewards those teams that go the extra distance and discourages coaches from becoming "too involved" with the design. A sophisticated design/program that the kids can't explain is assumed to have been done by the coach and will score lower.



Subjective Judging Form

Design Award

Team # _____ Team Name _____
School/Org Name _____ Scoring Group _____

Award Objective: This award evaluates the mechanical design of the robot. Criteria includes a solid and reliable design, use of gears and sensors, and use of sound mechanical design principles.

Scoring Elements

1. Stuctural Design **Score (1-20):** _____

- Strength to Weight
- No parts falling off
- Modular
- Stable
- Handles environmental variation

Comments:

2. Locomotion **Score (1-20):** _____

- Speed appropriate to task
- Controllable, repeatable
- Precise
- Is MECHANICALLY capable of task
- Drive train is solid

Comments:

3. Manipulation **Score (1-20):** _____

- Manipulators appropriate for task
- Is MECHANICALLY capable of task
- Maintains balance and stability
- Controllable, repeatable
- Precise

Comments:

4. Navigation **Score (1-20):** _____

- Is MECHANICALLY capable of task
- Sophistication of control
- Controllable, repeatable
- Little wasted motion
- Precise

Comments:

5. Overall Design and Integration **Score (1-20):** _____

- Design consistent with team's plan
- Design consistent with team's scoring plan
- Shows systems thinking
- All elements work well together

Comments:

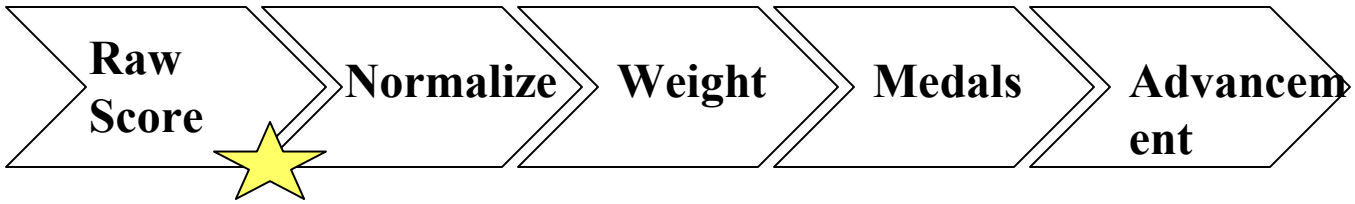
Total: _____
(100 max)

version:1.0 Judges initials _____

- One for each of the 5 judged categories
- Subcategories scored
- Sheet, with comments, given back to teams at the end.
- Judge's score, termed "raw score" is normalized.
- These forms available online, essentially unchanged from 2002/3.

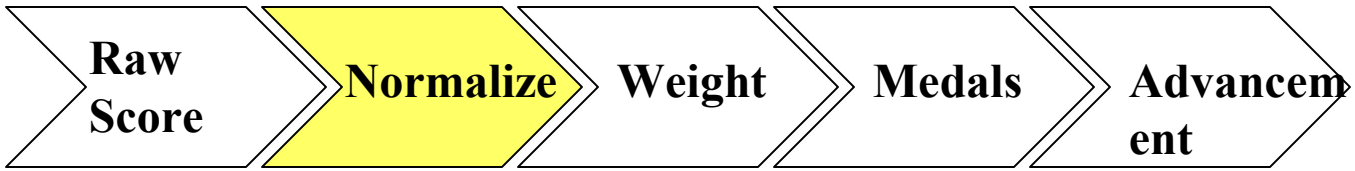
Judged Trophies

- The trophy is awarded to the highest raw score in
  - Research Presentation
  - Teamwork
  - Robust Design
  - Programming
- If there are multiple judges (for example two teamwork judges), their scores are normalized as usual and then averaged for the scaled score.
- Performance trophy goes to the winner of the head to head, not the highest score during the 3 rounds.



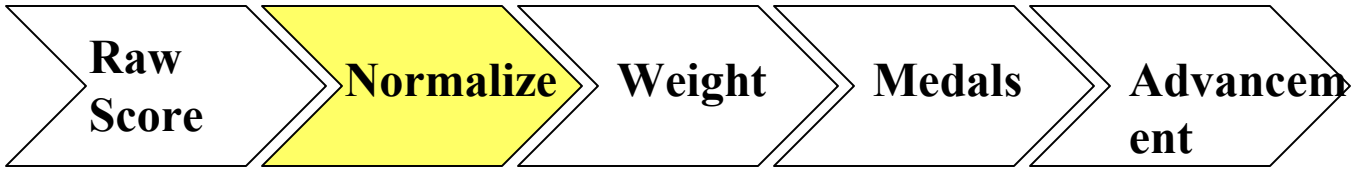
Normalized Scores

- Raw scores simply can't be added for a number of reasons:
  - All scores must be brought to the same scale (Performance ranges from 1-400, but judged scores range only from 1-100).
  - Even out judges who scale differently (i.e. the "French Figure Skating judge" judge who scores low).
- The master score is put together with weighting given to certain categories to emphasize certain aspects of the competition, normalization must be done to preserve that weighting



Normalization Details

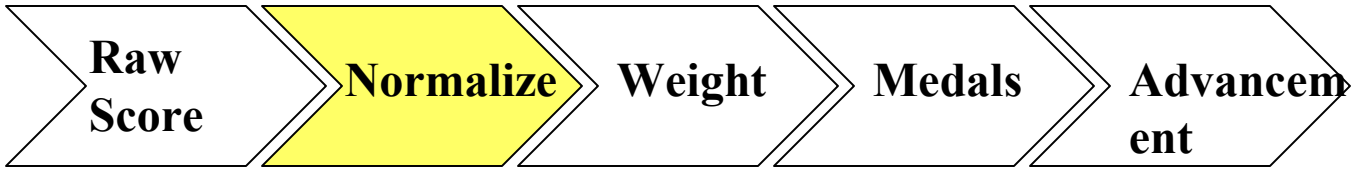
- A ratings scale is developed to normalize around an average score of 100 for each category.
  - We normalize based on how far your score is from the average score in that group of scores
  - So it's how far you are above or below the average score in each judged category that counts
  - A score of 120, is one standard deviation above the mean.
- This accounts for the amount of variability of judges scores as well as their average
- There are a number of ways to normalize, but normalization to the average is generally considered best by statisticians.



Example

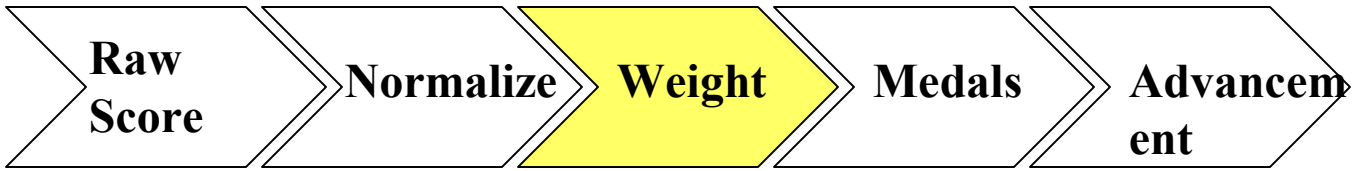
	Scores	
	Raw	Normalized
	94	147.6
	80	121.32
1 Std. Dev.	79	119.44
	76	113.81
	73	108.18
	70	102.55
	69	100.67
Average	65	93.16
	63	89.41
	62	87.53
	61	85.65
	60	83.78
1 Std. Dev.	57	78.14
	52	68.76

- This is an example from the 2003 Sandburg regional tournament. 14 teams judged for Div 1 programming.
- Average score was 68.64, with a standard deviation of 11.1.
- That means a score of 68.6 would normalize to 100, and 79.7 (average + 1 Stand. Dev) would normalize to 120.
- The top scoring team was rewarded for a having a much higher score than any other team, so the normalization preserves both the ranking and the absolute values of the judges.



Weighting

- Now that scores are all normalized, they can be weighted and added
- The Master score is comprised of
  - 25% Robot Performance
  - 25% Research Presentation
  - 25% Teamwork
  - 25% Technical Design (which breaks down to 12.5% Programming and 12.5% Robust Design)
  - This Master score is based on the weighting defined by FIRST.
- That means we can add the first 3 (performance, research and teamwork) and half the Programming and Robust Design normalized scores. According to this formula, an "average" Master Score would be 400.

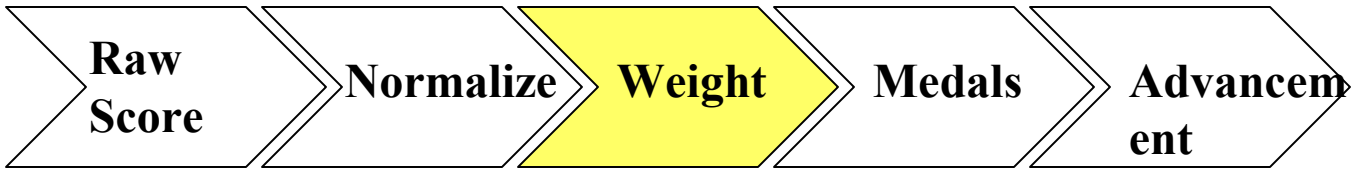


Example - Weighting

Team # / Organization / Team Name		Robust Design	Research Presentation	Programming	Teamwork	Performance	Overall Score
3680 Field Middle School	Raw:	93.00	82.00	93.00	87.00	258.00	
Severe Acute Mars Syndrome (S.A.M.S.)	Scaled:	134.74	113.27	139.10	117.16	144.33	511.68

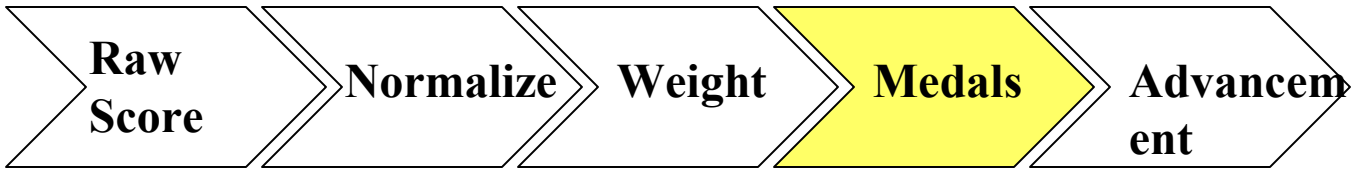
Master Score Equation = **134.74*0.5 + 113.27 + 139.10*0.5 + 117.16 + 144.33 = 511.68**

- The SAMS team Master score is the sum of the individual normalized scores, properly weighted. The raw scores are printed for reference only.
- According to this formula, an "average" Master Score would be 400. The SAMS team was overall a very strong team, scoring well above average in every category.



Medals

- At MN Regional tournaments, we give out a participation medal to every team member.
- Trophies are given to teams and generally sit on a shelf somewhere, but the medal is something each student gets to keep and shows how their team did. We make it more than a mere participation medal by using the Olympic style gold, silver, bronze approach.
- The color medal gives the students a sense of how they performed overall and a measuring stick for improvement.
- The Master Scores are divided approximately into thirds to determine medal color for each team.



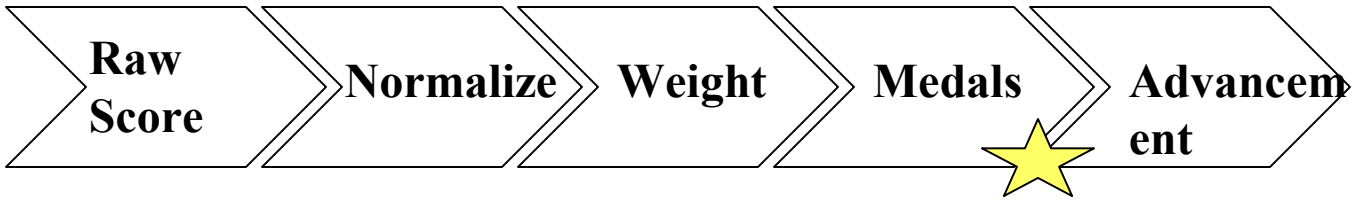
Example - Medals

	Team # / Organization / Team Name		Robust Design	Research Presentation	Programming	Teamwork	Performance	Overall Score
Gold	3680 Field Middle School	Raw:	93.00	82.00	93.00	87.00	258.00	
	Severe Acute Mars Syndrome (S.A.M.S.)	Scaled:	134.74	113.27	139.10	117.16	144.33	511.68
	3319 Marcy Open School	Raw:	86.00	92.00	59.00	89.00	154.00	
	Basement Builders, Inc.	Scaled:	119.35	129.07	100.57	120.98	89.98	449.99
	2530 Anwatin Middle School	Raw:	84.00	74.00	91.00	88.00	132.00	
	Strangers here Ourselves	Scaled:	114.95	100.63	136.83	119.07	78.49	424.08
Silver	2393 Monticello Middle School	Raw:	77.00	75.00	56.00	84.00	185.00	
		Scaled:	99.56	102.21	97.17	111.44	106.18	418.20
	1691 Heritage Christian Academy	Raw:	73.00	91.00	50.00	78.00	172.00	
	LegoMeisters	Scaled:	90.77	127.49	90.37	100.00	99.39	417.45
	1819 Grandview Middle School	Raw:	65.00	78.00	42.00	90.00	158.00	
	Tonka-Bots	Scaled:	73.18	106.95	81.30	122.88	92.08	399.15
	1981 TLC	Raw:	86.00	58.00	55.00	61.00	235.00	
	Psychotic Maniacs	Scaled:	119.35	75.35	96.03	67.58	132.31	382.93
Bronze	263 W. Harry Davis Academy	Raw:	66.00	61.00	49.00	70.00	186.00	
	Milky Ways	Scaled:	75.38	80.09	89.23	84.74	106.71	353.85
	2558 Litchfield Middle School	Raw:	69.00	73.00	40.00	66.00	159.00	
	Mars Maniacs	Scaled:	81.97	99.05	79.03	77.12	92.60	349.27
	1978 sandburg middle school	Raw:	73.00	52.00	50.00	67.00	170.00	
	lions	Scaled:	90.77	65.87	90.37	79.02	98.35	333.81

This Div. 2 group at Sandburg regional consisted of 10 teams. The break lines for medals are determined by the head judge and generally break like grading on the curve. While the Monticello and Heritage teams were close to the top, the head judge felt that given the number of teams, and that 3 teams were advancing to state, the breaks fell where they did. While moving the gold/silver break down two teams, or up one, wouldn't be wrong, it's a judgment call.

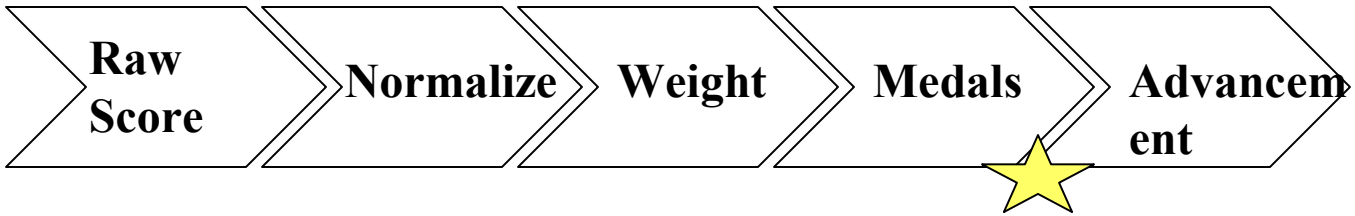
Other Trophies

- The referees provide main input on the Team Spirit winning team.
- The judges and referees throughout the day are asking and looking for interesting background stories about the teams as input into the Against All Odds and Judges' Awards.
- The Innovative Design scores are tabulated and the head judge looks at these as well. There is one trophy per division, unlike the other trophies on this page, which are one per tournament.
- The head judge makes the final decisions on all these awards, attempting to balance awards to as many teams as possible yet still fairly rewarding teams.



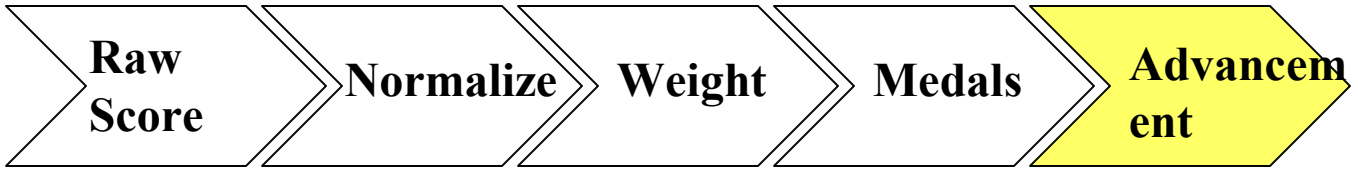
More on Trophies

- Against All Odds-
  - May be a story of overcoming something during the season or tournament.
  - Often given to a new school/team or a team starting in a remote location away from much support.
  - Intent is to encourage teams of all abilities and locations.
- Judges Award
  - May be a particularly striking example of technical prowess, teamwork, or community service.
  - Often awarded to a team that does very well in a number of categories but has not won another trophy.
- Statistical analysis of regional trophies for the past 2 MN FLL seasons showed that 50% of all teams won at least one trophy at the regional level.



Advancement

- Advancement is based on percentages of teams registered in Div 1 or 2. We aim for a 32 team event for each division at state. This year we have 95 Div 1, 112 Div 2 teams, and about 35-40 Div 3 (High School) teams. That means about 34% (32/95) of Div 1 teams advance and about 29% (32/112) of Div 2 teams advance. I say about because of ties, and rounding issues for individual events.
- For our Sandburg example, we had 16 Div 1 teams registered, which makes the equation = (32/95)*16 = 5.39. 5.39 is rounded up to 6, because many tournaments have 16 teams, so when looking at the overall tournaments and numbers, that makes more sense than rounding down. It means more than 32 teams at state but in these kinds of decisions, the benefit is always given to more teams rather than less.
- The numbers stay the same even if some teams don't show up for the event.
- A tie is determined to be within 1.0 point.
- The Head Judge always has the ability to advance teams under special circumstances (judging mixups, data entry errors, etc.).



Example - Advancement

Advance →

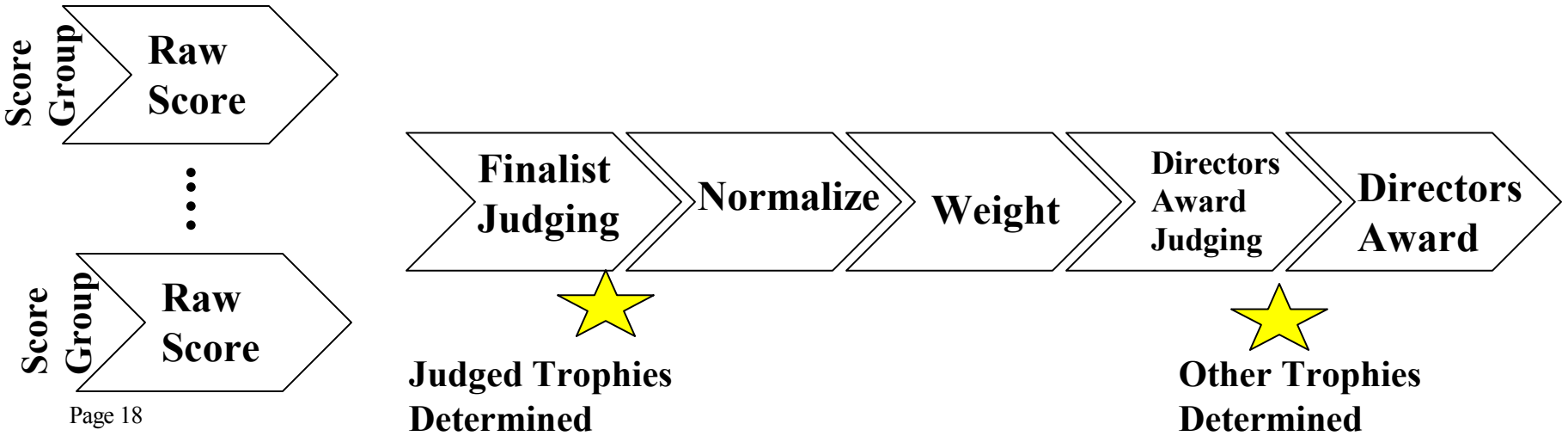
Team # / Organization / Team Name		Robust Design	Research Presentation	Programming	Teamwork	Performance	Overall Score
3680 Field Middle School	Raw:	93.00	82.00	93.00	87.00	258.00	
Severe Acute Mars Syndrome (S.A.M.S.)	Scaled:	134.74	113.27	139.10	117.16	144.33	511.68
3319 Marcy Open School	Raw:	86.00	92.00	59.00	89.00	154.00	
Basement Builders, Inc.	Scaled:	119.35	129.07	100.57	120.98	89.98	449.99
2530 Anwatin Middle School	Raw:	84.00	74.00	91.00	88.00	132.00	
Strangers here Ourselves	Scaled:	114.95	100.63	136.83	119.07	78.49	424.08
2393 Monticello Middle School	Raw:	77.00	75.00	56.00	84.00	185.00	
	Scaled:	99.56	102.21	97.17	111.44	106.18	418.20
1691 Heritage Christian Academy	Raw:	73.00	91.00	50.00	78.00	172.00	
LegoMeisters	Scaled:	90.77	127.49	90.37	100.00	99.39	417.45
1819 Grandview Middle School	Raw:	65.00	78.00	42.00	90.00	158.00	
Tonka-Bots	Scaled:	73.18	106.95	81.30	122.88	92.08	399.15
1981 TLC	Raw:	86.00	58.00	55.00	61.00	235.00	
Psychotic Maniacs	Scaled:	119.35	75.35	96.03	67.58	132.31	382.93
263 W. Harry Davis Academy	Raw:	66.00	61.00	49.00	70.00	186.00	
Milky Ways	Scaled:	75.38	80.09	89.23	84.74	106.71	353.85
2558 Litchfield Middle School	Raw:	69.00	73.00	40.00	66.00	159.00	
Mars Maniacs	Scaled:	81.97	99.05	79.03	77.12	92.60	349.27
1978 sandburg middle school	Raw:	73.00	52.00	50.00	67.00	170.00	
lions	Scaled:	90.77	65.87	90.37	79.02	98.35	333.81

This Div. 2 group at Sandburg regional consisted of 11 teams (1 no show due to weather). Advancement equation = (32/112)*11 = 3.14. This is rounded down to 3 teams. Gold medals do not always correlate with advancing to state, but it can help to draw the break for medals, depending on the number of teams.

State Tournament

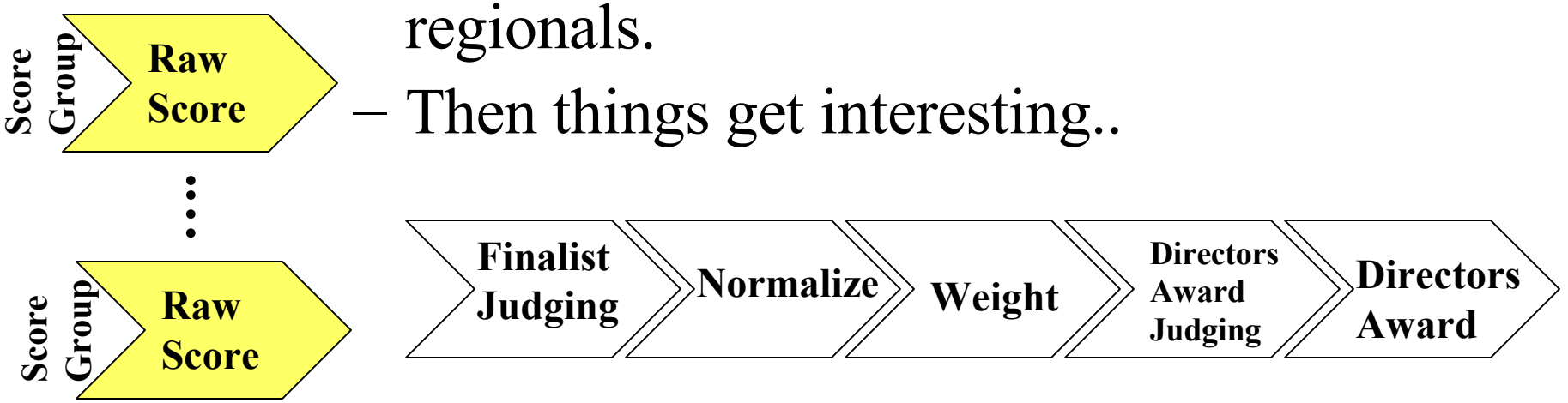
- The state tournament is larger which changes the process a bit. This process, with the exception of the Director's Award, could also be used for any large regional event.

State Tournament Process



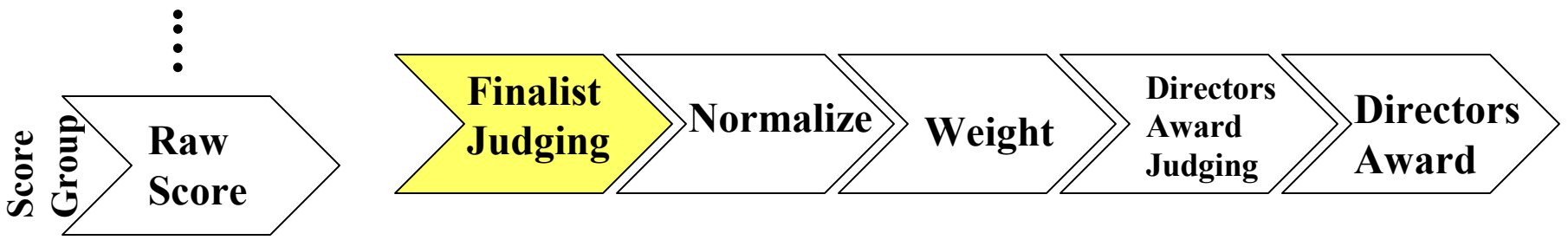
Scoring Groups

- Judges can realistically only judge 15-16 teams in one group. The limitation is time, and generally the eye-glazed factor.
- So how do we judge a group of 32 or more teams?
  - Each division is broken down into 3 scoring groups.
  - These groups are judged, as they are at regionals.
  - Then things get interesting..



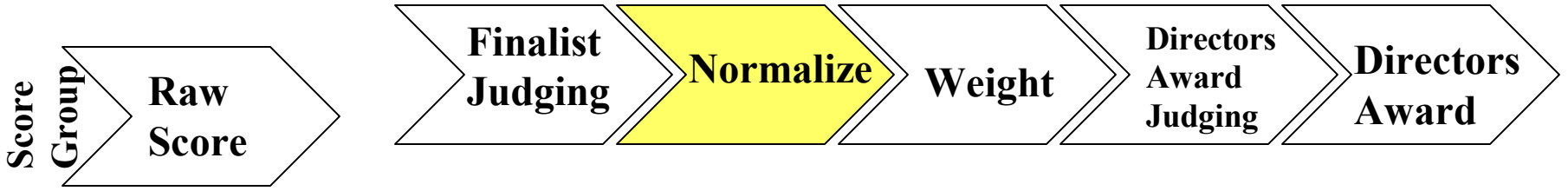
Finalist Judging

- The raw scores are input into the computer and normalized as before.
- The top scoring team from each scoring group advances to Finalist judging. Here a group of judges or a single judge will look at these top scoring teams from each group, and rejudge them to determine which one is best. Having the same judge (s) look at the same teams is the only fair way to determine the absolute best team for the state award.
- The Normalized scores are then added as before to get a Master Score. However, because of the normalization, the category results may look weird.



Normalization

- How does normalization work with multiple judges?
  - We must normalize across judges as well. Which simply means we compare the normalized score across judging groups, not the raw score.
  - What that means is a team potentially may finish with the highest normalized score but not win the trophy.
    - How can that be?
    - Teams are assigned randomly to judging groups. A team may be judged to be far better than other teams in their judging group, and ends up with a higher score above average than the other two top scoring teams in the other two judging groups. Comparing raw scores across judging groups means very little. Different judges have different scales. What's important to a judge is consistency. We normalize in order to try and bring the scales together.
    - However because judging is subjective, that's why we bring the top three teams together for another round of judging. The same judges see the same teams and compare them. Subjective judging is all about comparing teams.
    - The bottom line is, if you have a team that had the highest normalized score in the category, you are in the finals for judging, but you may not win the trophy. You need to think of the two rounds of judging as a preliminary round, and then the final round.



Example from Last Year

Tournament: STATE

InnovativeDesign Division: 1

	Team # / Organization / Team Name		Raw Scaled
Group A	370 Sunny Hollow Elementary	F.B.I.	97.00 127.84
Group A	3603 Al-Amal School	City Surfers	96.00 126.33
Group C	2996 Snail Lake School	Metro Mechanics	92.00 126.30
Group B	1989 Chippewa Middle School	Dead Batteries	89.00 124.71
Group C	1452 Crystal Lake Elementary	K2 Katz	90.00 123.84
Group B	155 Alice Smith Elementary	Alice Smith Utility Workers	88.00 123.29
Group B	1704 Minnie Neighborhood	G-TRAMS	88.00 123.29
Group A	1988 Chippewa Middle School	Bots in Black	92.00 120.32

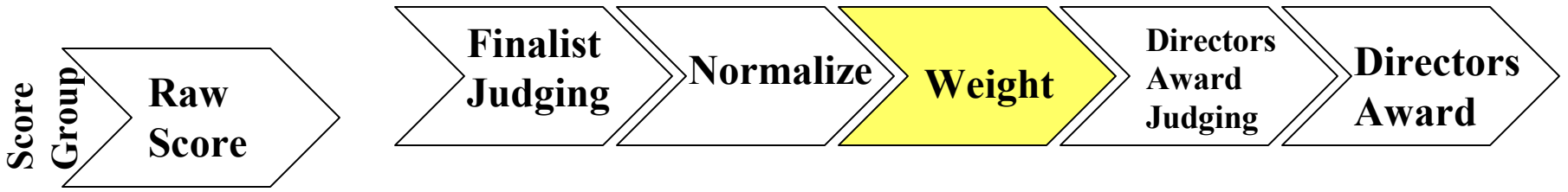
- Innovative Design, Div 1, 2002/3 State Tournament.
- The finalists were: Sunny Hollow FBI, Chippewa Middle School Dead Batteries, Snail Lake Metro Mechanics. Each finished first in their respective judging group.
- Continued on next page..

Example from Last Year

- Continued from previous page
- The raw scores were 97 for FBI, 89 for the Dead Batteries, and 92 for Metro Mechanics. Those raw scores were given by 3 different judges based on their judging of 3 completely different groups of teams. But they were the highest raw score in each of their respective judging groups. Al-Amal City Surfers had a 96 for a raw score but they were in the same group as the FBI, so they finished 2nd in their group. The difference of one point is an indication that the judge deemed the teams virtually the same but had to pick one, and found one to be superior in a particular way.
- Does the 96 of Al Amal indicate they are a better team than the Metro Mechanics, who scored 92? No. Normalized, their scores are virtually identical (126.33 to 126.30). And in head to head judging, the Metro Mechanics were judged to be the winner in Innovative Design.
- Do we want judges to use the same scale? Of course but it's impossible to calibrate that exactly, and especially at this level of competition. That's why we normalize to get a master score.

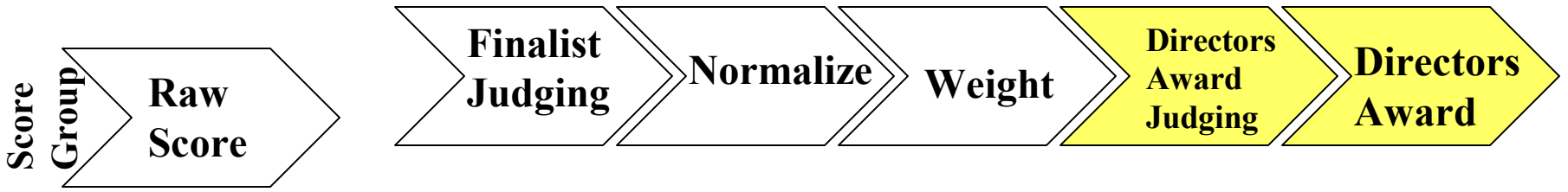
Weight

- The weighting for Master Score is the same as for regionals.
- For state, at this point, Master score is really only used in Director's Award criteria and to give you a sense of how your team ranked overall.
- All medals to teams at state are the same. They happen to be bronze because that is the color FIRST makes them.



Director's Award

- Considered the Grand Champion and is only given at State tournament.
- Process
  - After first round of judging, evaluate master score
  - Interview top 2-4 teams
  - There are also judges scattered throughout the building during the day watching the teams.
  - Look for lowest score (want a team that scores high in all areas, not a couple overpowering ones)
  - Look for kid-driven teams
  - The highest master score does not always win
    - Has happened a couple times, once when team had only 2 kids and below average research presentation (but a heck of a robot).
- Still interview to ensure a kid driven team and "intangibles"
  - FIRST calls these FLL Values
  - Ultimately it's choice of the Director's Award judges and the head judges.
  - It's hard to quantify the subjective percentage but it's 10-20% of decision.



Normalization Details

- {Won't be discussed, for reference only}
- Compute a z score for each team for each event
  - A z score is the number of standard deviations the team's score is away from the mean of the group being averaged
  - Z-scores by definition have a mean of zero and a standard deviation of 1
- To get a mean of 100 and a standard deviation of 20, we multiply the Z score by 20 and add that result to 100.
- This not only centers the mean at 100, but it makes sure that a score that is better than 66% of the competitors in that category is always 120 and a score that is better than 95% of the competitors in that category is always 140, etc.
- This accounts for the amount of variability of judges scores as well as their average.