The Social Construction of Rankings

Milo Schield
Augsburg College, Minneapolis, MN

Abstract

Rankings are common in the news. Superlatives abound: best, most, biggest, tallest and highest. How meaningful are these rankings? How can one assess the objectivity in a ranking? This paper examines ten factors involved in the social construction of rankings. One of these, the choice of the context or competition, is very powerful yet is often ignored. The goal is to help the reader become more aware of the different kinds of choices involved in constructing a ranking. This paper upholds the Thomas-Sewell (2008) thesis that the education of statisticians should cover the social construction of statistics and the Allen-Sharep (2005) recommendation that students should be educated on the dangers in ranking. Rankings should be used in Statistical Literacy to show students how easily everyday statistics are socially constructed.

1. Nature and Function of Ranks

Comparison is one of the most elementary human activities. Comparisons often involve comparatives (better) or superlatives (best). A more nuanced comparison orders things.

Ranks indicate order as indicated by first, second, third or by second place and third ranked. To move higher in a ranking (e.g., from second place to first) is to move to a lower-numbered rank. Ordered comparisons need not involve numbers as in top rank, in last place or in small, medium and large. For details on ties, see Appendix A.

Ranks omit the spacing between the measures. The separation between first and second place may be small or large. Nevertheless, ranks are an excellent way to become familiar with a new area. Check the web for the proliferation of top 10 lists.

Rankings can be controversial. "Hundreds of educators take part in the union-sponsored demonstration that accused the paper of unfair reporting in using a statistical analysis to rank the performance of thousands of instructors." LA Times (15 Sept, 2010).

Rankings are generally created for public consumption:

> *Americans tend to fetishize numbers, to assume that figures imply accuracy, precision, and science. Numbers seem to embody the objectivity that we seek from experts. If we give cities a livability score, and one city gets 287 points and another 286, that one-point difference provides the basis for ranking one above the other. To understand these rankings, it is necessary to take the formula apart, to understand the process that produces those scores ….* Best (2011)

As Joel Best (2001) notes so eloquently, all statistics are socially constructed: they are constructed by people with values and choices. Since people have choice, this means that rankings are more like diamonds than rocks: they can be selected, fashioned and presented to maximize their value or impact. Ranks involve many choices. In order to evaluate the objectivity and the value of a ranking, one must be able to envision how a ranking was formed, how it could have been done differently and what impact those alternate choices would have had on the ranking. For more on the social construction of statistics in general, see Schield (2010).

2. Types of Rankings

Rankings can be classified into two types based on their method of construction:

1. Pure judgment-based evaluations by experts or connoisseurs (Reviews, Ratings)
2. Purely objective ranking based on counts or measures

There can be mixtures (hybrid objective-subjective rankings) such as the U.S. News College Ranking that includes subjective ratings by college Presidents and Deans.

If the ranking is subjective, the key question is "Who did the review and what are their standards?" Blank (2007) gives an in-depth analysis of judgment-based reviews.

If the ranking is objective, the key question is what factors were included (and excluded) and how are they fashioned? This paper deals with objectively-based (non opinion-based) rankings. Objective rankings – number-based (count and measurement-based) rankings – can be divided into two types: single-factor and multi-factor.

3. Number-Based, Single-Factor Rankings

Single-factor number-based rankings seem objective since there is no weighting. But this may mask the subjectivity in choosing the factors involved in fashioning the underlying data. We will consider eight choices that can influence these single-factor rankings:

1. Choice of the factor involved (compared with others that could have been used)
2. Choice of the context or competition in the ranking
3. Choice of the denominator for ratio-based factors
4. Choice of what was not taken into account – the influence of confounders.
5. Choice of the basis for time-comparisons,
6. Choice between amount change and percentage change for time comparisons
7. Choice between score change and rank change for time comparisons
8. Choice regarding statistically-insignificant differences for all comparisons

3.1. Choice of Factor

Probably the easiest way to influence a ranking is in the selection of the factor involved.

In some cases, selection is simply impossible. The world's fastest tennis serve is simply a matter of a single measurement: According to Aneki.com, Andy Roddick (US) holds the record with a Queen's Club tournament serve (2004) at 153 mph (246.2 km/h).

In most cases, there are several ways to measure the outcome on which ranks are based.

Which country was the winner of the 2008 Olympics? Isaacson (2010) demonstrates the social construction of rankings by asking students this question. Students realize there are many ways to measure the winner: most medals (US) or most gold medals (China). Student constructed a weighted average of medals where gold=3, silver=2 and bronze=1 (China), most medals per capita (Bahamas), most gold medals per capita (Jamaica) or most medals per GDP (Zimbabwe); these uses of weights and different denominators are discussed later in this paper. After getting a different country for most criteria, students could see why the Olympics committee does not announce any country as winner.

What is the best-selling movie of all time? *Titanic* based on actual receipts; *Gone with the Wind* based on inflation-adjusted receipts or on total tickets sold. See Answer.com.

What US state is most diverse? Consider relevant data for the six largest states.

Table 1: Diversity Statistics of the Six Largest US States

Count (1,000) population	STATE	Percentage who are black	Percentage who are immigrants	Percentage who are Hispanic	Percentage who are non-native	Percentage who speak non-English
35,466	California	7% (6)	0.49% (1)	35% (1)	26% (1)	41% (1)
16,982	Florida	16% (2)	0.31% (3)	19% (3)	17% (3)	24% (3)
12,650	Illinois	15% (3)	0.25% (4)	14% (5)	13% (5)	20% (5)
19,238	New York	17% (1)	0.47% (2)	16% (4)	20% (2)	28% (4)
12,351	Penn	11% (4)	0.12% (6)	4% (6)	4% (6)	8% (6)
23,508	Texas	11% (4)	0.23% (5)	33% (2)	14% (4)	33% (2)

Source: 2006 US Statistical Abstract [Six-state rankings added in parentheses]

Depending on which factor we use to measure "diversity", California or New York could rank in first place while California can rank in first place or last.

Moral: *Small changes in the selection criteria can produce big changes in ranks.*

3.2. Choice of the Context (Limiting the Competition)

One way to stack the deck in a ranking is to limit the competition: The smaller the competition, the higher (better) the numerical ranking. This choice is usually much less obvious than picking the factor, but it is often much more powerful.

Sometimes the class is obvious and appropriate: "El Paso is the second-safest mid-sized city in the US."

But sometimes the choice of class is not obvious and is chosen just to improve the ranking. A potentially-devious use of this technique is indicated by the phrase "in our class" or "in its class." If "our class" or "its class" is not clearly defined, one can easily increase one's ranking without anyone realizing.

This devious usage works perfectly when the class is unstated and excludes everyone else. The webmaster of www.StatLit.org could say that this web site is "first in its class." Being first is guaranteed so long as "its class" is defined to exclude all other websites.

Moral: *Small changes in the competition can produce big changes in ranks.*

3.3. Choice of Denominator for a Ratio

A potentially dominating factor in determining the order of ranks is the choice of the denominator when the measure is a rate or ratio. With rates, the denominator is often indicated by *per*: the accident rate per person, per vehicle or per mile of road. Suppose we are comparing the auto death rates by state in the US. This statistic seems relatively immune to social construction since death is fairly well-defined.

In some cases, the choice of denominator has little effect on the rankings. Figure 1 compares auto death rates by 100,000 vehicles with those per 100,000 licensed drivers.

Normally a scatter plot shows the relation between an independent and a dependent variable. This scatter plot just shows the relation between two variables of equal status.

Normally the points on a best-fit line are considered to have the strongest connection between the two variables. Points off of the best-fit line are often viewed as error.

Rankings work differently. Rank on the horizontal axis is determined by how many dots are farther to the right; rank on the vertical axis is determined by how many dots are farther above. States with highest death rates (highest rank) are in the upper-right corner.

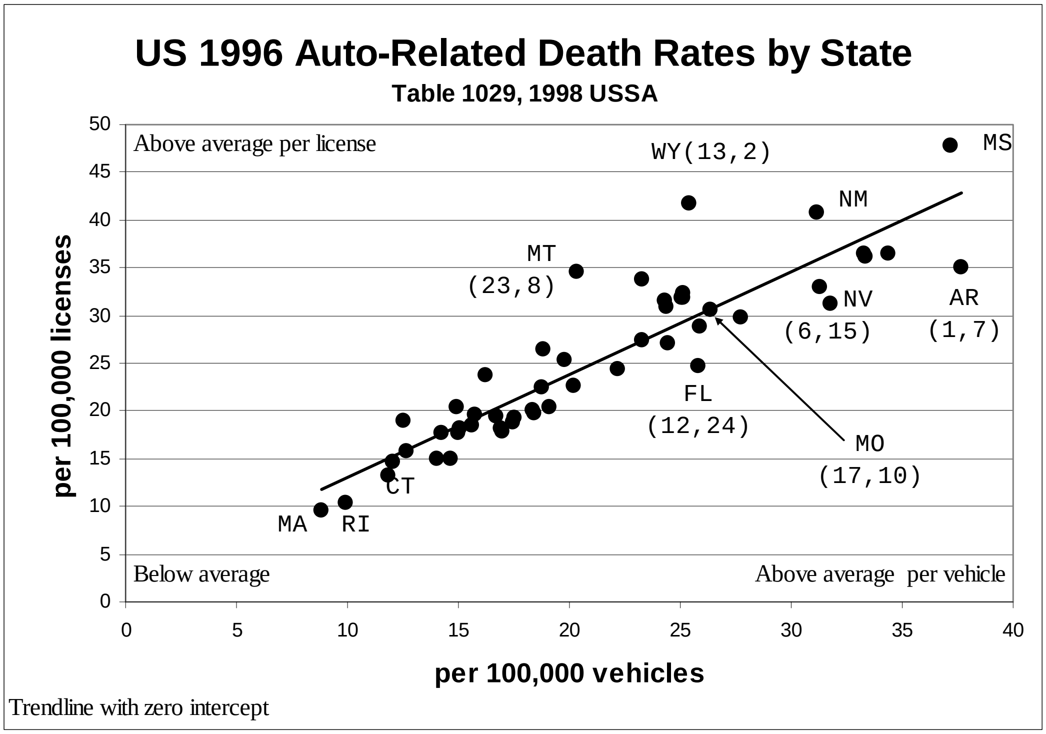

Figure 1: US auto death rates per vehicle vs. per licensed driver

In Figure 1 where the highest rate is #1, Arkansas (AR) ranks #1 per vehicle but #7 per licensed driver: a difference of six places out of 50. To see this, look at how many dots (states) are farther right (none) vs. farther above (6). The biggest difference in ranks is Montana (MT): 23rd per vehicle (22 to the right) vs. 8th per licensed driver (7 above).

Being located on the best-fit trend line is no assurance that the rank on one axis is about the same as on the other. Missouri (MO) lies on the trend line. It ranks 17th per vehicle but 10th per licensed driver: a difference of 7 places out of 50.

In Figure 1, ranks on the horizontal axis (per vehicle) were closely related to ranks on the vertical axis (per licensed driver). Mississippi (MS) was highest (in first place) on both scales. In Figure 1, 54% of the states had rank differences of 3 or less.

The choice of the denominator can have a major influence on ranks. This is shown in Figure 2. Of the 50 states shown in Figure 2, half had rank differences of 16 or more.

Figure 2 compares auto death rates per 100,000 vehicles with those per 1,000 miles of road: two different denominators. Arkansas ranks #1 (furthest right) per vehicle but ranks #32 per mile of road: a difference of 31 places. Hawaii ranks #29 per vehicle but #1 per mile of road (furthest up): a difference of 28 places.

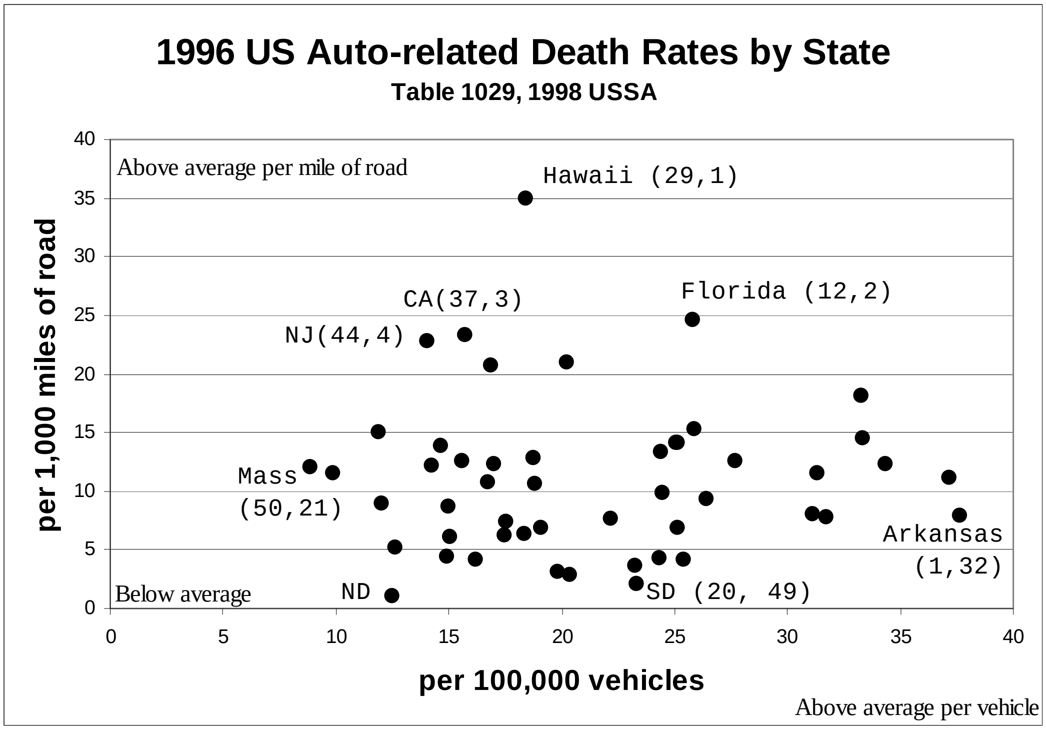

Figure 2: US auto death rates per vehicle vs. per mile of road

Moral: Small differences in the choice of denominator can produce big differences in ranks. For more on this topic, see Thomas and Sewell (2008).

3.4. Influence of Potential Confounders

While potential confounders can influence all kinds of statistics, the influence may seem less obvious when dealing with ratios.

Controlling for confounders can change the values of ratios (averages) which in turn can cause changes in rankings. A simple way to control for a confounder is to compare the overall association with the associations in each of the subgroups. Terwilliger and Schield (2004) did this for state rankings using NAEP grade 4 mathematics data for 2000.

Table 2 shows NAEP state scores broken out by family income. As shown in Table 2A, the state score is two points lower for Oklahoma (OK) than for Utah (UT). Yet when classified on family income (based on school lunch payment status), the state score for each subgroup is higher for Oklahoma than for Utah. Note that the percentage of high income families is larger in Utah (64%) than in Oklahoma (45%) and students from high income families tend to score higher than those from low income families. After taking into account family income, OK ranks higher than UT: a reversal.

Table 2: NAEP 2000 Grade 4 Math Scores Classified by Family Income

State	All	High $	Low $
UT	227	233	216
OK	↓225↓	↑234↑	↑218↑

Table 2A UT vs. OK

State	All	High $	Low $
MD	222	233	207
LA	↓218↓	233	↑211↑

Table 2B: MD vs. LA

As shown in Table 2B, the state score is four points lower for Louisiana (LA) than for Maryland (MD). Yet when classified on family income, the state score for each subgroup is at least as high for Louisiana as for Maryland. The percentage of high income families is greater in Maryland (58%) than Louisiana (32%); students from such families tend to score higher. After controlling for family income, LA ranks higher than MO: a reversal.

Table 3 shows state scores broken out by school location. As shown in Table 3A, the state score is two points lower for New York (NY) than for Missouri (MO). Yet when classified by school location, the state score for each subgroup is at least as high for New York as for Missouri. The percentage of students who attend non-city schools is higher in Missouri (78%) than in New York (54%) and that those attending such schools tend to do better. After controlling for school location, NY ranks higher than MO: a reversal.

Table 3 NAEP 2000 Grade 4 Math Scores Classified by School Location

State	All	City	Non-City
MO	229	216	233
NY	↓227↓	216	↑236↑

Table 3A: MO vs. NY.

State	All	City	Non-City
GA	220	208	222
TN	220	↑213↑	↑224↑

Table 3B: GA vs. TN

As shown in Table 3B, the state score is the same for Tennessee (TN) as for Georgia (GA). Yet when classified by school location, the state score for each subgroup is two to five points higher for Tennessee than for Georgia. Note that the percentage of students who attend non-city schools is higher in Georgia (85%) than in Tennessee (71%) and that the students who attend non-city schools tend to do better. After taking into account school location, Tennessee ranks higher than Georgia: a change from a tie.

In all four cases, the relative ranking of the two states changes after taking into account the influence of a confounder: family income or school location. The failure to take into account relevant factors is perhaps the most important influence in ranking large groups with heterogeneous subjects such as schools and hospitals. See Schield (2006).

Moral: *What is – or is not – taken into account can easily change a ranking.*

3.5. Time-Based Comparison: Choice of the Basis

A common factor in ratings is change – improvement: Most improved team, etc. But any measure of change depends critically on the basis chosen for the comparison.

Figure 3 plots the percentage change in state lottery revenue. The horizontal axis shows a two-year basis; the vertical axis shows a one-year basis. States with negative changes or inordinately large changes are not shown. In the following discussion, there are two kinds of changes: change of basis (over different times) versus change in ranks.

In many states, the choice of the basis has little consequence. West Virginia ranked higher than any other state shown on both the one-year and the two-year basis.

For some states, the choice of the basis has major consequences. In going from a one year basis (vertical scale) to a two-year comparison (horizontal scale), Florida jumped 23 places from 28^{th} to 5^{th}, New York jumped 15 places from 25^{th} to 10^{th} , Missouri jumped 13 places from 17^{th} to 3^{rd}, and Oregon jumped 12 places from 31^{st} to 19^{th}. Meanwhile, Vermont dropped 26 places from 8^{th} to 34^{th}. These changes of 12 to 26 places are major.

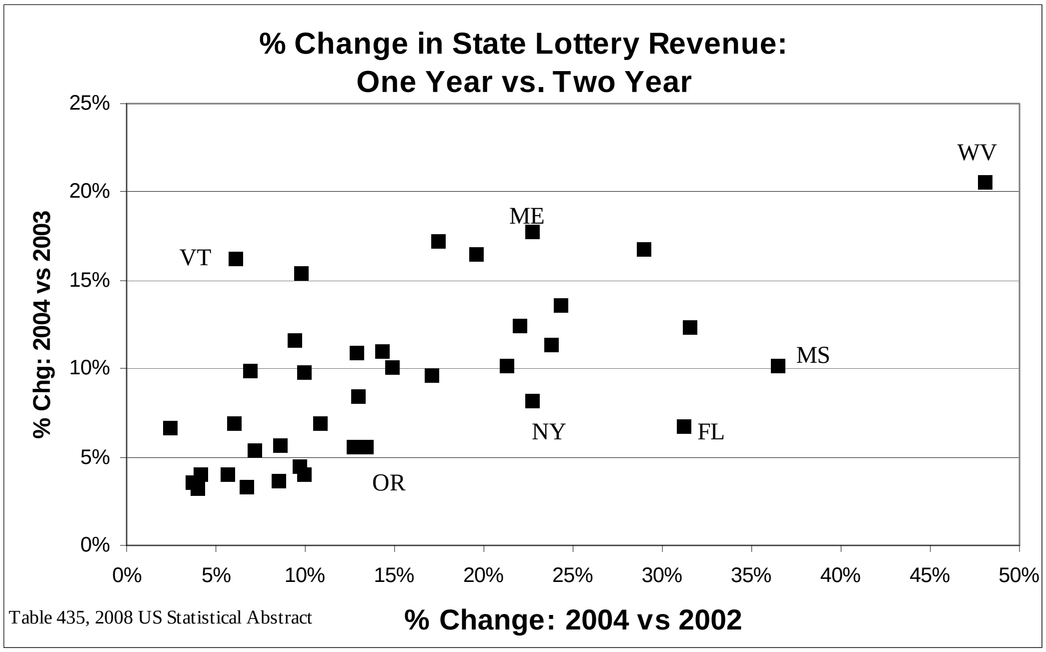

Figure 3: State Lotteries: Percent Change—One Year vs. Two Year

Moral: *Small changes in the basis of a comparison can produce big changes in ranks.*

3.6. Time-Based Comparison: Amount Change vs. Percentage Change

A related way of getting a higher rank in measuring improvement is to choose the better of amount change or percent change. To see how this works, compare the increase in state lottery revenues between 2003 and 2004 as shown in Figure 4.

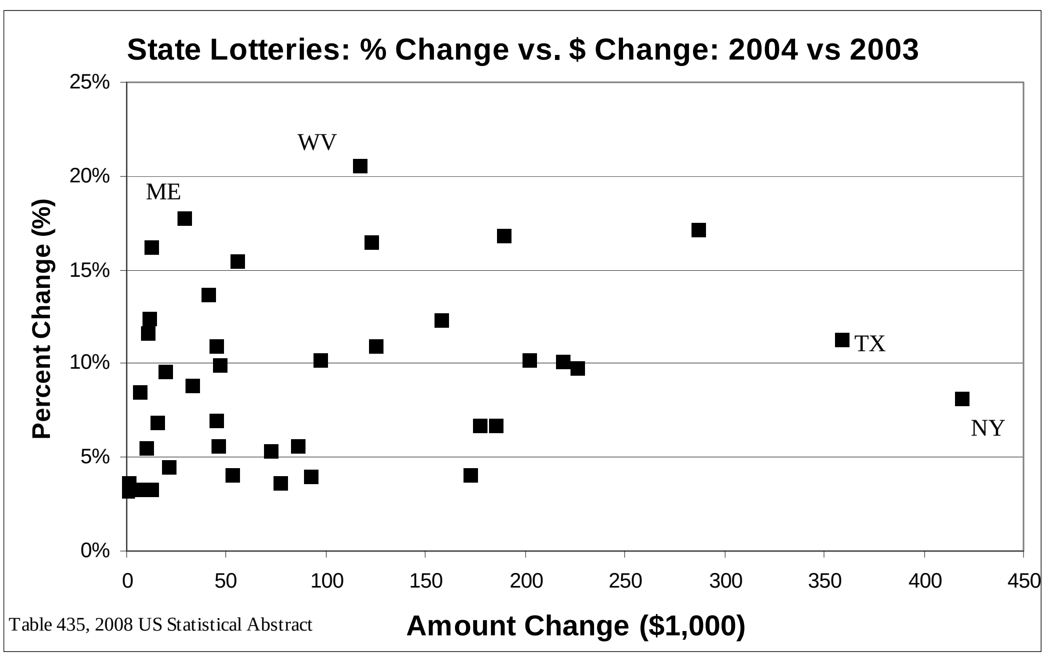

Figure 4: State Lotteries: Percent Change vs. Amount Change in Dollars

New York ranks first based on amount increase; West Virginia ranks highest based on percent increase. To improve their rank, New York will measure improvement using dollar increase whereas West Virginia will use percent increase. Both New York and West Virginia can honestly say "we ranked first in increasing lottery revenue between 2003 and 2004". The ambiguity is in the choice of what is being increased.

Moral: *A small difference in the type of change can produce big difference in ranks based on change.*

3.7. Time-Based Comparison: Score Change vs. Ranks Change

Figure 5 illustrates NAEP 8th grade math ranks by state for 2007 and 2009. A higher ranking (a higher score) has a lower-numbered rank (lower left corner).

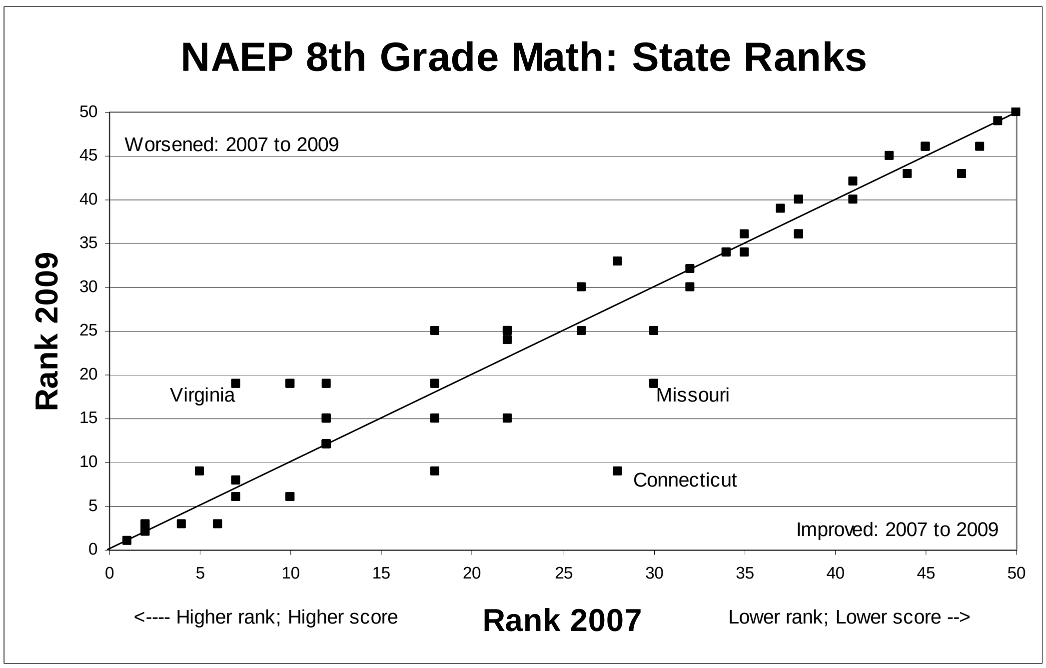

Figure 5: NAEP 8th Grade Math: State Ranks

States having the same rank in 2009 as in 2007 lie on the diagonal line. States that improved with higher ranks (lower numbers) in 2009 than 2007 are in the lower-right. States that worsened with lower ranks (higher numbers) are in the upper left.

Based on their change in ranks, Connecticut was the "most improved state" by rising 19 places from 28th to 9th place while Missouri was second by rising 11 places from 30th to 19th. Virginia (upper-left) was in last place dropping 12 places from 7th to 19th place.

A very deft (underhanded?) way of getting higher ranks in measuring improvement is to focus on the change in ranks – not the change in the underlying scores. We say how to focus in change in ranks in Figure 5. Now we will relate those changes in rank to the underlying changes in score.

An improvement in scores may generate a worsening in ranks, a worsening in scores may generate an improvement in ranks, and ranks can change without any change in scores – all because the scores of others are changing. To see how this works, consider Figure 6.

Figure 6 shows how these changes in ranks compare with the changes in the underlying scores. Connecticut's 19 place improvement in rank was a result of a 7 point increase in scores from 282 to 289. Missouri's 11 place improvement in rank was a result of a 5 point increase in scores from 281 to 286. Virginia's 12 place drop in ranks was a result of a 2 point drop in scores from 288 to 286.

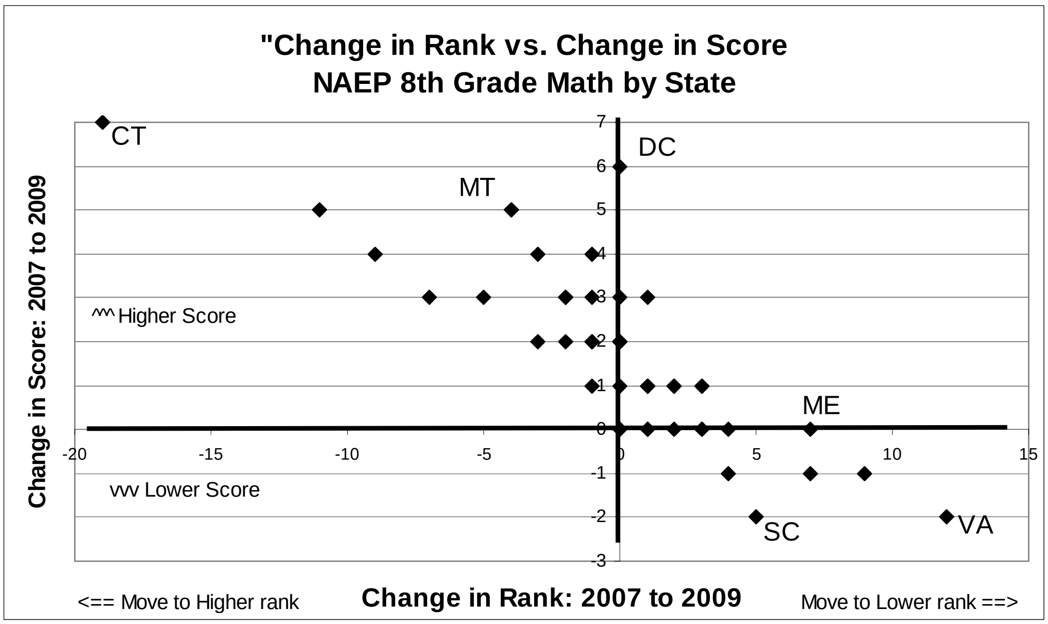

Figure 6: NAEP 8th Grade Math: Change in Rank vs. Change in Score

Note that Virginia (VA) dropped only 2 points but dropped 12 ranks. Connecticut improved by 7 points but moved up 19 ranks. Moral: *Small changes in scores may result in big changes in ranks*.

Consider Washington DC. Their score improved by 6 points (from 248 to 254) but their rank remained unchanged. Moral: *Large changes in scores may result in small or no changes in ranks.*

Maine scored the same in 2009 as in 2007 (286), yet they dropped by 7 places from 12th to 19th. Moral: *A score that remains unchanged can still result in a change in ranks.*

3.8. Comparison: Statistical Significance

Ignoring the lack of statistical significance is a choice that allows one to capitalize on noise – on bogus or spurious differences. The NAEP data included the standard error for each state. These ranged from 0.5 for Delaware and South Dakota to 1.6 for Michigan and Louisiana. Of the 50 state changes in scores, only 18 were statistically significant:: the change exceeded the 95% margin of error (1.96 times the state's standard error). While standard errors varied by state, the average of all state SEs is shown in Figure 7.

Using exact standard-errors by state, all nine of the zero-point changes were statistically insignificant. All 14 of the one-point changes were statistically insignificant. Nine of the 12 two-point changes were statistically insignificant. Thus, most (32) of the 50 NAEP state score changes between 2007 and 2009 were statistically insignificant.

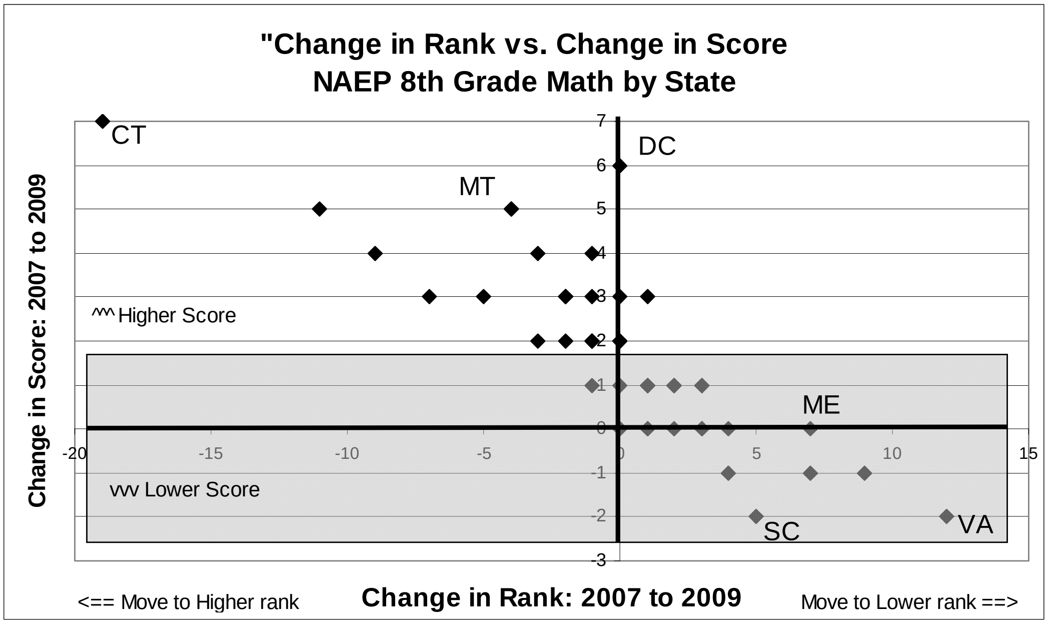

Figure 7: NAEP 8th Grade Math Change. Statistically Insignificant

Moral: *When the scores being compared are very close together, many of the score changes from one moment to another may be statistically insignificant.*

The area in grey contains all the changes that are statistically insignificant. It also includes four two-point changes that are statistically significant. In the lower-right corner, note the 12 rank drop by Virginia, the eight rank drop by Maine and the five rank drop by South Carolina – all statistically insignificant.

But the issue of statistical insignificance also applies to comparisons between states within a given year as well as to changes in a state's score over time.

Consider the NAEP 8th grade math scores by state along with their 95% margin of error bars as shown in Figure 8. Only one difference between adjacent state scores is statistically significant (non-overlapping Margins of Error bars): the difference between Massachusetts (#1) and Minnesota (#2). All other adjacent differences are statistically insignificant. If all states had a 2 point 95% margin of error, then 18 states ranking from 12 to 29 would have scores whose differences were not statistically significant.

In summary, for this NAEP data most of the year-to-year differences in scores for a given state are not statistically-significant and almost all of the differences between adjacent state scores are statistically insignificant. In general, ignoring the influence of randomness allows one to treat small differences as real rank differences when in fact they both may be spurious.

Moral: *When the scores being compared are very close, many of the differences in ranks at a given time – and changes in ranks over time – may not be statistically significant.*

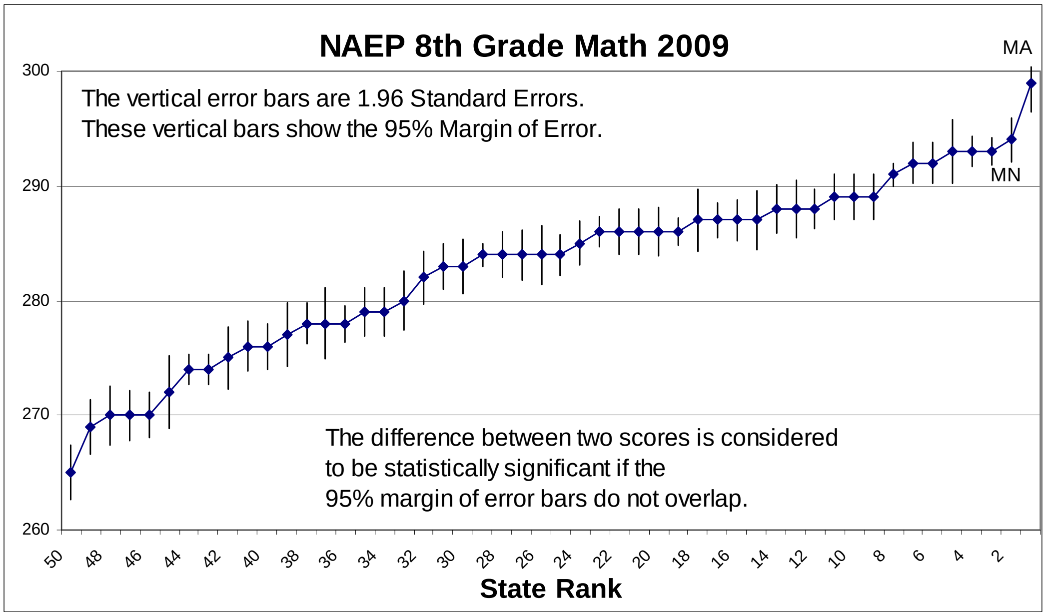

Figure 8: NAEP 8th Grade Math Scores with 95% Margin of Error Bars

3.9. Summary of Influences on an Objective, Single-Factor Ranking

The influence of eight choices on a single-factor objective ranking has been investigated. Most of these influences might be summarized as answers to two questions: "Compared to what?" (2, 3, 5, 7) or "compared by what?" (1, 4, 6, 7 or 8).

We now investigate the last two sources of influence that apply to multi-factor rankings.

4. Multi-Factor Rankings (Number-Based)

When evaluating a multi-factor ranking, all eight single-factor influences remain, but some new factors emerge. We will consider two:

1. Choice of weighting
2. Choice of scaling

4.1. Choice of Weights

The choice of weights is essentially subjective. When used to combine primary measures that are objective, they are treated as objective since their impact is clearly visible.

Myers and Robe (2009) provide an excellent history of college rankings. They note:

> *One specific result of a highly-precise, ordinal ranking is that changes in the methodology of the ranking system can produce changes in an institution's rank without any change in the institution's quality. Minute adjustments to what criteria are used, how these criteria are weighted, and how institutions are classified and chosen for ranking can result in wide swings or even complete disappearance of a school's rank, even though the characteristics of the institution remain unchanged. In fact, one study published in May of 2008 by two mathematicians at the University of California, Berkeley, revealed that the U.S. News rankings are highly volatile. By*

adjusting the different criteria weighting and using the same data as U.S. News, these researchers concluded that a school's "specific placement…is essentially arbitrary."

A real-world example of the importance of weighting is when two rankings of the same outcome disagree – even though they are using the same inputs.

Baroness Young has resigned as Chair of the Care Quality Commission (CQC) after reports of conflicts with Health Secretary Andy Burnham concerning procedures for monitoring the NHS [National Health Services].

In particular, there has been strong media attention on the apparent conflict between the ratings provided by the CQC and those released last week by Dr Foster in their Good Hospital Guide: nine of the 12 hospitals identified by Dr Foster as "significantly underperforming" for patient safety had been rated as "good" or "excellent" by CQC for Quality of Services.

Both sets of ratings rely on the mathematical combination of performance indicators to create an overall rating and so are essentially based on statistical analysis. So Baroness Young can perhaps be seen as a statistical casualty.

Source: Straight Statistics (UK): www.straightstatistics.org/article/statistical-casualty

To examine the influence of weights, consider the MacLean's Personalized University Ranking Tool for Canadian Colleges. This tool allows the user to rank Canadian colleges by selecting their own weights and seeing what difference their choice makes. Go to: http://oncampus.macleans.ca/education/category/rankings/rankings-tool/ Click on "CLICK HERE TO GO TO THE TOOL" Select up to seven of the 13 indicators shown.

On a national basis, Laurentian ranks in the bottom half on 11 of the 13 MacLean's single-factor college indicators for 48 Canadian colleges. The following fictional story illustrates how easy it may be to generate higher ratings. Laurentian is used in this very informal example.

The Laurentian President picked the four categories that seemed most important to prospective students and parents. The categories and Laurentian's national ranking were Reputation (43^{rd}), Library expenses, % (35^{th}), Awards to faculty per faculty (39^{th}) and Student awards per student (40^{th}). The President weighted these four equally and MacLean's ranked Laurentian as #45. Admissions was not impressed.

The Dean picked the four categories in which Laurentian ranked highest nationally: Operating Budget, $ per student (7^{th}), Student/Faculty Ratio (15^{th}), Scholarships & Bursaries: % of budget (26^{th}) and Student Services, % of budget (27^{th}). The Dean ranked each of the four equally and MacLean's ranked Laurentian as #10 nationally. The President was very impressed but Admissions was not.

The PR officer took the President's four indicators, added Student-Faculty ratio and found that Laurentian ranked #4 *in its class*. "Class" was Ontario-based colleges with student-faculty ratio weighed at 60% and the President's four indicators ranked equally at 10% each. Doing this raised Laurentian from #40 nationally to #4 locally.

The PR officer took the Dean's four indicators, ranked them equally and found that Laurentian ranked #2 *in its class*. As before, "class" was Ontario-based colleges. Restricting the competition moved Laurentian from #10 nationally to #2 locally.

> The PR officer made small (10 pt) changes in the weights in the Dean's model (35-35-15-15) and found that "Laurentian is #1 *in its class*." Everyone was impressed.
>
> As a result, the President featured "we're #4", the Dean featured "we're #2" and Admissions featured "we're #1." No one ever mentioned the choices involved in "our class" and – sadly but not unsurprisingly – no parent or student ever asked.

This very-informal make-believe story clearly illustrates how much flexibility there is in calculating the ranking of a given Canadian college. While the primary focus in this story is on the influence of weights (#9), this story also involved the influence of selecting the factors (#1) and restricting the competition (#2). Readers usually recognize the importance of choosing the factors (#1) and the weights (#9), but may be totally oblivious to the importance of restricting the competition (#2).

4.2. Choice of Scaling

Any ranking involving multiple factors requires that scores with different ranges be scaled. To see why consider a hypothetical ranking where the three factors average 1, 10 and 100 respectively with a 20% variation around each average. If the three factors are just added together, the biggest factor will dominate. Thus, some scaling is required.

The first step is to put positive and negative scores on the same basis. In some cases, higher scores are positive (percentage of the faculty who have graduate degrees); in other cases, higher scores are negative (class sizes and golf). Weights are assumed to add to one. To avoid negative weights, negative scores must be converted into positive scores. There are several ways to do this; here are two:

1. "Complement" the negative scores. For example, if the highest golf score is 120, then a score of 70 would become 50, while a score of 90 would become 30.
2. Negate negative scores so that higher (farther right on the number line) is better.

The second step is to scale disparate amounts assuming higher is better. Consider these:

1. **Average Scaling**: NewScore = 100*OldScore / AveScore.
2. **Maximum Scaling**: NewScore = 100*OldScore / MaxScore.
3. **Max-Min (Range) Scaling**: NewScore = 100*[(OldScore-Min) / (Max-Min)].
4. **Z-score Scaling**: NewScore (Z) = (OldScore – Average) / Standard deviation

McLean (2010) appears to use Max-Min (Range) scaling. Since each scaling of a single factor preserves the underlying ranking, one might wonder why the choice really matters. The choice of scaling matters when the scaled values are combined. Showing this in general is beyond the scope of this paper. But it can easily be seen in a special case.

Suppose that we want to rank student athletes on size. Size involves a combination of height and weight. Suppose that one athlete is average height and one standard deviation above average in weight while the second is the reverse: one standard deviation above average in height and average weight. The z-scores for the first athlete are zero and one; the z-scores for the second are one and zero. The average z-score for each athlete is 0.5 and so they tie in rank.

Consider average scaling. Suppose the average height is 67" with a 3" standard deviation; the average weight is 160# with a standard deviation of 20#. On this basis, the first student is 67" tall and weighs 180#; the second student is 70" tall and weighs 160#.

The first student has a scaled height of 100 (100*67/67) and a scaled weight of 112.5 (100*180/160). The second student has a scaled height of 105 (100*70/67) and a scaled weight of 100 (100*160/160). Assume equal weights. The composite score of the first is 106.25; of the second is 102.5. In this average-based scaling, the first student will rank higher in size than the second.

Moral: Subjects with the same composite z-scores may differ in their composite average-based scores and thus may differ in their rankings.

For more background on rankings, see Allen and Sharep (2005) and Becker, et al (1987). For more details on the US News and World Report College Rankings, see the articles by Morse (2010), by Myers and Robe (2009) and by US News and World Report (2010).

5. Teaching Implications

We strongly agree with the teaching implications noted by Allen and Sharep (2005):

> *Rankings are a common tool used in the comparison of towns, institutions, teams, individuals, countries – and almost every entity that can be quantified in some way. As a result, they are frequently used to make financial, policy, and political decisions. It is well known that most rankings can be improved – either by investigating the ranking method, the composite indices, or the raw data used to create the indices. Thus it is important that our students learn about ranking issues in their statistics courses. Since our students are future consumers, and perhaps creators, of surveys and rankings, it is essential that students are educated in the dangers of rankings. It is important that our students understand that there exist multiple ranking methods – each of which will yield different (although perhaps similar) results. … Perhaps by including a discussion of ranking methods in our statistics courses, we can enhance the ethical creation and consumption of rankings, in general.*

6. Conclusion

Whenever a rank is given, *take care*: recognize that all rankings are socially constructed. The idea that all statistics are socially constructed is a big idea – perhaps the biggest in all of statistics. Teaching this principle may be most productive when analyzing ranks. The mathematics are minimal so the influence of social construction is more obvious.

7. Recommendations

Given the importance – the fundamentality – of the social nature of statistics, statistical educators should support the Thomas-Sewell (2008) thesis that the education of statisticians be extended to cover the social construction of statistics and the Allen-Sharep (2005) recommendation that our students should be educated on the dangers in ranking. Rankings should be used in Statistical Literacy to show students how easily statistics are socially constructed.

Acknowledgments

To the W. M. Keck Foundation for their grant “to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts.” To Joel Best for identifying and communicating the importance of the fact that all statistics are “socially constructed.” To Cynthia Schield, Tom Burnham, Robert Raymond, Marc Isaacson, John Schmit, Larry Lesser and Larry Copes for their comments and suggestions.

References

Allen, I. Elaine and Norean R. Sharep (2005). Demonstration of Ranking Issues for Students: A Case Study *Journal of Statistics Education* Volume 13, Number 3 (2005), www.amstat.org/publications/jse/v13n3/sharpe.html

Answers (2010) http://wiki.answers.com/Q/What_are_the_top_ten_grossing_movies_of_all_time

Becker, R.A., Denby, L., McGill, R., and Wilks, A.R. (1987), "Analysis of data from the Places Rated Almanac," *The American Statistician*, 41(3), 169-186.

Best, Joel (2001). *Lies, Damned Lies and Statistics*. University of California Press.

Best, Joel (2011). *Everyone's a Winner: Life in Our Congratulatory Culture*. University of California Press. [To be released in February]

Blank, Grant (2007). *Critics, Ratings and Society: The Sociology of Reviews*. Rowman and Littlefield.

Isaacson, Marc (2010). Selecting the Winner in the 2008 Olympics. Data at http://simon.forsyth.net/olympics.html and http://simon.forsyth.net/olympicsGDP2008.html.

Los Angeles Times (2010). L.A. Unified teachers stage protest at Times building. www.latimes.com/news/local/la-me-teachers-protest-20100915,0,750979,print.story

MacLean (2010). MacLean's Personalized University Ranking Tool at http://oncampus.macleans.ca/education/2008/02/11/the-macleans-personalized-university-ranking-tool/

Morse, Robert (2010). Methodology: Undergraduate Ranking Criteria and Weights. http://www.usnews.com/articles/education/best-colleges/2010/08/17/methodology-undergraduate-ranking-criteria-and-weights-2011.html

Myers, Luke and Jonathan Robe (2009). *College Rankings: History, Criticism and Reform.* Center for College Affordability and Productivity. Copy available at www.centerforcollegeaffordability.org/uploads/College_Rankings_History.pdf.

Schield, Milo (2006). Presenting Confounding and Standardization Graphically. STATS Magazine, American Statistical Association. Fall 2006. pp. 14-18. Draft at www.StatLit.org/pdf/2006SchieldSTATS.pdf.

Schield, Milo (2010). Assessing Statistical Literacy: Take CARE. Ch 11 in *Assessment Methods in Statistical Education: An International Perspective*, pp. 133-152, Wiley.

Terwilliger, Jim and Milo Schield (2004). Frequency of Simpson's Paradox in NAEP data. Presented at the American Educational Research Association. Copy at www.statlit.org/pdf/2004TerwilligerSchieldAERA.pdf

Thomas, Ray and Martin Sewell (2008). The Qualities of Statistics as Facts about Society. *DataCrítica: International Journal of Critical Statistics*, Vol. 1, No. 2 (2008). Copy at http://www.datacritica.info/ojs/index.php/datacritica/article/view/13

US News and World Report (2010). How US News Calculates the Rankings. http://www.usnews.com/articles/education/best-colleges/2010/08/17/how-us-news-calculates-the-college-rankings.html

Appendix A: Ranking Ties

Here are four ways to rank ties with four players where 2 and 3 tie.

1. Standard competition ranking: Both ties rank high – "1, 2, 2, 4" ranking
2. Modified competition ranking: Both ties rank low – "1, 3, 3, 4" ranking
3. Dense ranking: "1, 2, 2, 3" ranking
4. Fractional ranking: Both ties rank middle – "1, 2.5, 2.5, 4" ranking

Microsoft Excel uses standard competition ranking.