The Washington Flag Congress 2011

On a Scale of 1 to 10

Peter Hans van den Muijzenberg

Abstract

As NAVA's 2004 American City Flags Survey has provided us with design ratings for 150 flags, comparison of any other flag with those 150 flags would allow determining a likely rating for that flag. A more general method would be to derive a set of criteria that, when applied to a flag of United States local government, would yield a rating in line with those in that survey. Are such criteria possible, and what are the limitations of such a method?

Good design—the flag of Washington, D.C.

Proceedings of the 24th International Congress of Vexillology, Washington, D.C., USA 1–5 August 2011

On a Scale of 1 to 10

Introduction

Though most people have a good idea of the difference between a good flag and a bad flag design,[1] it's less easy to express how good, or how bad, a flag's design is. A scale for rating flags would allow comparing flags that have few common elements. It could also be used as an easy-to-understand expression of the quality of a flag design. The following looks at a possible way to achieve a set of simple criteria, which can be judged by their validity, and additionally seeks to determine the relative importance of the criteria within such a set.

A Scale of Design Quality

The simple approach to creating a scale for rating flag-quality would be to take the best flag[2] and the worst flag [3] known, and to rate the quality of designs as falling somewhere in between. Ideally, personal preference would be removed by using the opinions, regarding good flags and bad flags, of as large a group of judges as practically possible. Furthermore, adding ratings for a number of flags in between would create an even scale, avoiding, for example, a limit below which a flag suddenly is considered "bad". An approximation of such a scale can be found in the results of NAVA's American City Flags Survey in 2004,[4] where 150 flags received ratings on a scale of 0 to 10 (though actually ranging from 1.48 to 9.17).

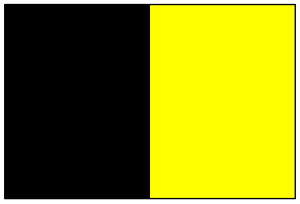

Figure 1.
Good design – the flag of Sneek.

Yet, if a flag design were to be compared directly with the flags listed, personal views might well cause a rock-paper-scissors-like triangle, where of two flags on the list plus one other design, each might be considered superior to one and inferior to another. Instead of comparing a design directly with the flags on the list and deriving a score from their ratings, we would need a way to determine the underlying criteria that resulted in those ratings, so these criteria could then be applied to the as-yet-unrated flag design. These criteria, if they could be determined, would give a fairly objective basis that would allow giving a specific rating to a flag design.

Good/Bad Flag

Opinions about the criteria for the quality of flags are mostly expressed as lists of "do"s and "don't"s. This led Ted Kaye to create *Good Flag, Bad Flag*, which condensed multiple personal lists into one five-point list of how to create a good flag design.[5] The drawback of that approach in itself, however, is that it does not quantify. It allows noting that a flag does, or doesn't, have certain characteristics, and from this an opinion can be derived on whether that makes it a good, or bad, flag. There is, however, no direct way to express how good, or how bad.

To be able to express just that, Mason Kaye proposed the K Scale, which awards 0, 1, or 2 points for each of the five principles of *Good Flag, Bad Flag*.[6] Indeed, applied to the NAVA's American City Flags Survey, the results of this method were in line with the ratings the city flags had received.[7] It does not quite match the intent given above, though, as apart from the five criteria it also requires a certain amount of judgement, to determine how well a flag matches a criterion. The criteria are not just judged by their validity, but by the level of their validity.

A Set of Criteria

Which criteria, that can be either valid or invalid, form the basis for the judgement of the flag, and will give that same accuracy the K Scale does but won't require determining a level of validity? These criteria are unknown, if they even exist: the ratings of the survey are not based on specific criteria but on collective judgement. However, one could select a set of verifiable criteria and weight them to try to match the actual results as closely as possible. Applying such a set to a different design would then rate that design on the same scale.

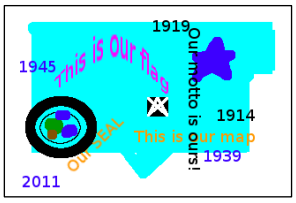

Figure 2.
Not a good design.

With the help of a spreadsheet, I selected and weighted such a set of criteria. Each criterion has a weight between 0 and 100 that indicates how important it is in rating the flag. Exceptions to this are two colour criteria that are actually weighted negatively. The percentages given in parentheses behind the weights indicate their portion of the total score generated by the criteria, and thereby the relative impact of each criterion.

Text

- No text at all, except for letter art: 96 (24.6%)
- No more than one text element (a scroll or a ring with text counting as a single element): 64 (16.4%)
- No text written in an arc (a scroll is considered not to be arced): 33 (8.5%)

Field

- An area of non-white touches the flag edge: 58 (15.0%)
- The design includes one flag border, possibly offset, along at least half the flag edge: 26 (6.7%)
- A path of borderlines runs between two opposite edges (switching to the other border of a stripe is acceptable): 22（5.6%)
- At no point on the flag do three or more colours meet: 13 (3.0%)
- All edges of the flag are in what obviously is the field colour: 3 (0.8%)

Charge

- No element that appears to be arms, an armorial shield, or a seal: 29 (7.4%)
- The design includes a centred charge: 5 (1.3%)

Colour combination

- Two of the three colour shades of the Stars & Stripes make up more than half the flag: 13 (3.3%)
- The design includes shades of the three colours red, white and blue: 4 (1.0%)

Number of colours

- Two colours: 22 (5.6%)
- Three colours: 24 (6.2%)
- Four colours: 24 (6.2%)
- Five colours: 5 (1.3%)
- Six colours: -1 (-0.3%)
- Seven or more colours: -7 (-1.8%)

Placing these weights in a weighted formula gives a score for each of the flags in the survey.[8]

This method, with these weights, is not ideal; on average the scores are 0.67 away from the actual ratings, with a worst case difference of slightly less than 1.62. If we were to apply the same method to an unrated design with the same characteristics as the flags from the survey, we could expect a score with the same level of accuracy compared to an ideal rating system.

To verify this, I conducted a short survey on the Flags of the World mailing list[9] to obtain scores for two more flag designs, to be compared with the generated scores. The list members were asked to score the flags on the scale of the flags of the American City Flags Survey. The results, admittedly based on a rather small sample of list members, were as follows:

Figure 3.
Proposal: Washington, D.C.

The first, the proposed new flag for Washington, D.C., received 3.12 points. This puts it marginally above Lexington, Kentucky, which was 112th out of 150. The score for this flag based on the criteria was 4.27, which would place it slightly above Tampa, Florida, at 61st out of 150.

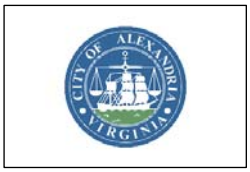

Figure 4.
Alexandria, Virginia

The second, the flag of Alexandria, Virginia, received 5.57 points. This puts it marginally below Minneapolis, Minnesota, which was 27th out of 150. The score for this flag based on the criteria was 3.87, which would place it below Fort Wayne, Indiana, and Lincoln, Nebraska, which tied at 82th out of 150.

These results may well be less than accurate due to the rather small number of votes, and may also differ because of a more international audience. For the Washington proposal a specific bias was present as the proposal was judged by some to destroy one of the best flags of the United States. But it is likely that the flags also introduce elements that so far were not included. For

the proposed flag for Washington, the effect of the lettering on stripes that don't touch the flywise edges is probably not covered by the selected criteria, nor is the relatively elegant seal of Alexandria, so simple a child could describe it from memory. They therefore may not completely fulfil the requirement of having the same characteristics as the flags from the survey, which might contribute to a deviation from the calculated scores.

Observations

The method described here, with these specific criteria, weights, and formula, is just one possibility among near-endless variations. Its criteria were chosen to address a number of widely-held views on flags, including three principles of *Good Flag, Bad Flag*, as it addresses simplicity, the number of colours, and avoiding lettering and seals. One further principle from that booklet is addressed only marginally: "Be distinctive or be related" is only addressed by the criterion for the colour combination of the Stars & Stripes.

The remaining principle of *Good Flag, Bad Flag* is not addressed at all: the meaning of the flag. Indeed, a set of criteria and weights such as these are unlikely to focus on *what* is expressed, rather than *how* it is expressed. It will address just the design, not its intentions. Likewise, the measures of relatedness and distinctiveness are not part of the design in itself, and in general are not easily included as criteria.

Apart from this general focus on the design of the flag, the particular data used may also have caused a bias in the selected criteria and weights. The American city flags may have particular design characteristics, and the audience rating them may have had a North American view on the quality of flag designs. On the other hand, this is a set of criteria that I came up with, so it's probably also biased toward my own situation. Mostly, however, the criteria are limited to characteristics of a rather narrow set of flag designs. As the two tests demonstrate, the weighting for a good match for this particular group is only relevant for flags that only show characteristics already present on the list. It's probably less relevant to other groups of flags. Also, several of the criteria still include a certain amount of interpretation, which in practice was also hampered by the limited size of the graphics available.

Regardless of the limitations of the specific group, the relations of the general principles that also were covered in *Good Flag, Bad Flag* do, where easy to describe, indeed show up as important. Their relative weights suggest that they are not all equally important, though, with most prominent a group of criteria covering writing on flags, which combined represent almost half the score.

If a more general method of scoring is possible, selecting and calibrating it will require a wider distribution of data and a wider selection of criteria, which should be particularly well-determinable. Ultimately, a better understanding would be needed of what makes a flag design better or worse. Without such an understanding, we are limited to scoring characteristics often displayed by good flag designs, rather than scoring the characteristics that make these designs good to begin with.

Conclusions

By using a series of rated flags, it's possible to calibrate a set of weighted criteria to provide ratings for flag designs similar to the ratings given by human judges. The resulting method can then be used to assign a consistent rating to other flag designs. This offers a more specific argument than just recognising an adherence to general principles of good flag design.

However, a particular set of criteria and weights optimised for a certain group of flags may become very specific to that group, limited by the variation occurring in that group of flags. To strive for a generally valid rating, the criteria will have to represent the characteristics that are universally considered to make a flag's design better or worse. The assumption in all of this, then, is that there are such criteria, which are global enough to make such a rating meaningful for all specific situations. Alternatively, different sets of criteria could be developed for different situations.

Acknowledgements

- Image of the flag of Washington, D.C. (U.S),[10] by *António Martins-Tuválkin* (2008), as published on the Flags of the World website.[11]
- Image of the flag proposal for Washington, D.C. (U.S.),[12] by *António Martins-Tuválkin* (2008), as published on the Flags of the World website.
- Image of the flag of Alexandria, Virginia (U.S.),[13] by *Joe McMillan* (2004), as published on the Flags of the World website.
- Image of the author, from a photograph taken by *Tineke Veltman* (2011).
- Information regarding the K Scale provided by *Ted Kaye* (2011).

End Notes

1 This same general grasp of flag knowledge is the basis for a method often used informally to identify a good flag, which Graham Bartram mentioned at the presentation of this paper: Since most people recognise a good or bad flag when they see one, a flag is likely to be a good design if it is actually flown frequently by the people it represents.

2 For the presentation, proving the flag of Sneek to be the best design known was used as a vehicle. (Figure 1)

3 As a stand-in for this, at the presentation an imaginary design was used that would receive the lowest possible score according to the criteria from this paper. (Figure 2)

4 "American City Flags Survey Results", North American Vexillological Association, 2004, http://www.nava.org/Flag%20Design/city_survey.htm.

5 Kaye, Ted, *Good Flag, Bad Flag*, North American Vexillological Association, 2001; http://www.nava.org/Flag%20Design/GFBF.

6 Kaye, Mason, "The Flags of Portland Oregon, (1916-2002)", *Proceedings of the XX International Congress of Vexillology*, Nordic Flag Society, Bergen, 2004, pp. 416-417.

7 Kaye, Edward B., "The American City Flag Survey of 2004", *Raven* 12, 2005, pp. 41–43, http://www.nava.org/documents/raven/vol12/NAVA_Raven_v12_2005_p027-062.pdf.

8 In detail, this is done as follows:

- The weights for the criteria that are met are added up, plus an extra weight of 27;
- The weights of all the criteria are added up, including the extra 27, and a further 100. As the weight for the number of colours, the highest of the weights for the number of colours is added once;
- The ratio is calculated as the first divided by the second;
- This ratio is multiplied by 9 and then 1 is added, for values in the range of 1–10.

The two constants, 27 and 100, modify the height of the scores, and their range. Changing these can improve the match for a given set of criteria.

9 FOTW (Flags of the World) mailing list, as described on its website: http://flagspot.net/flags/mailing.html.

10 FOTW, Washington, D.C. (US): http://flagspot.net/flags/us-dc.html.

11 FOTW website at: http://flagspot.net/flags/.

12 FOTW, D.C. Taxation Without Representation flag (U.S.): http://flagspot.net/flags/us-dc-nt.html.

13 FOTW, Alexandria, Virginia (U.S.): http://flagspot.net/flags/us-va-al.html.

About the Author

Peter Hans van den Muijzenberg, Frisian Herald of the Isles, was born in 1961 in Hilversum, but has spent most of his life in Sneek, where he currently resides. The formal education he received doesn't support his interest in flags, and he approaches the field mostly with general knowledge and skills in information retrieval. Though any problem in the field tends to have his interest, more than in details of specific flags he is interested in the development of the science as a whole.