When Numbers Lie
Good Bad Examples
Ditlevsen, Susanne; Poulsen, Rolf

Published in:
Significance

Publication date:
2023

Document version
Early version, also known as pre-print

Document license:
Other

Citation for published version (APA):
When numbers lie
A good bad example

When the Danish government was awarding a new DAB radio licence in 2019, taxpayers assumed the selection process would be fair and mathematically rigorous. However, Susanne Ditlevsen and Rolf Poulsen show how badly handled data led to a biased decision.

Numbers don't lie, as the old adage goes. But if tortured sufficiently, they will say anything. Here we present a case study in public administration involving a number of examples of innumeracy (a lack of ability to understand and use numbers in calculations) and other bad principles. While simple, it demonstrates a wider issue in our society, where faulty use and analysis of data might lead to unwarranted and wrong conclusions.

How not to do a public tender
In autumn 2019, the Danish government solicited bids for a nationwide digital radio channel – in other words, a digital audio broadcasting (DAB) licence. The winner of the licence would receive public funding of up to 280 million Danish kroner (DKKM; around €38 million) over 4 years. Applications were assessed and the winner chosen by the Radio & TV Board, an independent seven-person board under the Ministry of Culture. The board members’ backgrounds were in law or economics – none in publishing or broadcast media. Applications were scored in three main categories with weighting and subcategories, where each subcategory could receive an integer score between 0 and 8:

- operating plan (administrative aspects), with overall weight 35% and three subcategories;
- budget (amount applied for), with overall weight 25% and no subcategories;
- programme plan (creative aspects), with overall weight 40% and 11 subcategories, with one subcategory given double weight relative to the other subcategories.

By the deadline on 22 September 2019, there were three applicants: Radio24syv, LOUD, and dk4. On 22 October, the Board announced its decision in a report, whose final scores and rankings are shown in Table 1.

From the total scores it follows that LOUD was the winner – and seemingly a clear one at that. This came as a surprise to many. Radio24syv was regarded as the favourite by the general public, having successfully run a similar channel on FM radio for the previous 8 years. However, problems began to emerge. It was revealed that a cocktail of numerical errors and dubious algorithms had been systematically implemented in the evaluation, without which a different outcome could have been reached.
**Rounding in intermediate calculations**

For each applicant and within each category, the points from the subcategories were added up, divided by the number of subcategories, rounded to the nearest integer, and multiplied by the category’s weight factor. That was the number used for the ultimate score for each category in Table 1.

However, this way of calculating the score violates one of the most basic numerical rules: do not round off in intermediate calculations. If – as here – intermediate rounding is performed, you lose significant accuracy in the final product. In the programme plan category, for example, the raw numbers were that Radio24syv scored 6.3 points, LOUD got 6.6 points and dk4 got 4 points. There is only a difference of 0.3 points between Radio24syv and LOUD. With 40% weighting, this becomes a difference of 0.12 points. But that is not how the Board calculated the final category score. They rounded to the nearest whole number before applying the weights – therefore Radio24syv got 6 points and LOUD 7 points. Now there is a difference of one whole point – which after weighting becomes 0.4 points.

This is the difference used in the calculation of the final score. So there ends up being a larger difference in absolute numbers between two applicants, after down-weighting the result by 40%, than in the raw, unweighted numbers!

If all three applicants had scored either 0.2 points more or less, corresponding to a translation of the scale, then Radio24syv and LOUD would have received the same score – either 7 points (if they had all received 0.2 points more) or 6 points (if they had all received 0.2 points less). Thus, the relative difference in the final scores between applicants depends on the chosen scale. In this case, the rounding happens to exacerbate the difference between LOUD and Radio24syv.

**Averaging within subcategories**

The rounding issue was not the only case of flawed numerics. One of the programme plan subcategories (culture) was to be given double weight relative to the other 10 subcategories. The Board did this by multiplying the applicants’ point scores in this subcategory by 2. When the averages were calculated, the sum totals were divided by 11, the number of subcategories. However, because the culture subcategory effectively enters twice, the sum totals should be divided by 12.

It might be less obvious why this is wrong. Consider an example. You take two exams, both scored from 0 to 100. One has double weight. You score 80 on both. Of course, your weighted average score should be 80. By the Board’s method, however, your weighted average score would be (2 × 80 + 80)/2 = 120 out of 100.

The incorrect weighting scheme did indeed change the scores. The raw scores of 6.3, 6.6 and 4 points for Radio24syv, LOUD and dk4 should have been 5.8, 6.1 and 3.7, obtained by multiplying the previous scores by 11/12. Applying the previously discussed incorrect intermediate rounding and then down-weighting, the final scores would have been 2.4, 2.4 and 1.6. In this case, Radio24syv and LOUD would have obtained the same scores in the programme category. Applying down-weighting without intermediate rounding, the final scores should have been 2.3, 2.4 and 1.5 for Radio24syv, LOUD and dk4, respectively. Thus, the difference in LOUD’s favour should only be 0.1, and not 0.4.

**The budget evaluation – a tale of many errors**

The errors described in the previous section are equivalent to adding random noise, in that it is not systematic which applicants or category winners will benefit from them. However, the errors we discuss next are different and more severe, with the ability to systematically blow up even tiny differences.

**How to give more weight to the budget category than intended**

Table 2 takes a closer look at the budget category. Radio24syv applied for 100% of the maximum possible funding and got almost the minimum score, 1 out of 8. LOUD and dk4 applied for, respectively, 93% and 92% of the maximum funding and received scores of 7 and the maximum 8, respectively. Intuitively, that seems like a large swing in points based on differences in amounts that are small compared to the total budget, less than 9%.

This was a deliberate component of the application scoring algorithm. The lowest bid will always obtain a maximum score of 8, independently of how small the difference is to the highest bid. The other bids will get scores that depend on the minimum realistic bid, which in this case was decided to be the lowest bid rounded down to the nearest 10 DKKM. This forces the lowest point to be close to zero, as explained in the equations below.

Constructing an evaluation that forces this category to use the entire scale from a value close to 0 up to the maximum of 8, as opposed to the other categories which only use the central part of the scale, will effectively give the budget category a weight that is much larger than the intended 25%.

This is quite a subtle, but important, point, and it is an easy mistake to make. Intuitively, the scales between the categories should be comparable since they theoretically can take all values; however, in practice they rarely will. Since only differences between scores matter when determining the winner of the bid, categories with forced larger differences will inadvertently matter the most.

Let us formalise this comparison by introducing the explicit mathematical formulae used by the Board. The points awarded for a given bid $b$ are determined by a linear function $p(b)$ where the lowest bid $b_{min}$ is always awarded 8 points. The other bids are then awarded points in decreasing order as bids are increasing. How many points each bid $b$ is awarded (and, thus, how close the
highest bid is to the point scale minimum) depends on the slope in the linear function \( p(b) \). The slope depends on four values: the lowest and the highest bid, the maximum possible funding and the minimum realistic bid, which is fixed by the Board. Thus, a bid of \( b \) is awarded \( p(b) \) points, where \( p(b) \) is given by the linear interpolation function

\[
p(b) = \left(1 - \frac{b - b_{\text{min}}}{b_{\text{max}} - b_{\text{min}}} \right) \times \left( \frac{8 - 0}{8} \right)
\]

where 8 and 0 are the upper and lower point limits and \( b_{\text{min}} \) is the smallest budget applied for across all applicants, and the so-called slope coefficient \( s \) is determined as the simple average of two terms,

\[
s = \frac{\text{maximum possible funding} (280 \text{ DKKM}) - \text{minimum realistic bid} (250 \text{ DKKM})}{\text{minimum realistic bid} (250 \text{ DKKM}) - \text{lowest bid} (256.25 \text{ DKKM})}
\]

Table 2: Summary of applied budgets and the points assigned to each applicant in a Danish public tender round for a nationwide digital radio channel. The total scores the numbers appearing in the budget category of Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Radio24syv</th>
<th>LOUD</th>
<th>dk4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding applied for (DKKM)</td>
<td>280.00</td>
<td>260.82</td>
<td>256.25</td>
</tr>
<tr>
<td>Points, interpolation formula</td>
<td>1.03</td>
<td>6.66</td>
<td>8.00</td>
</tr>
<tr>
<td>Points, rounded</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Total score, weighted 25%</td>
<td>0.25</td>
<td>1.75</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 1 shows the interpolation formula, using the actual minimum realistic bid used by the Board (89.3% of the maximum possible funding), as well as minimum realistic bids of 50% and 80% of the maximum possible funding, respectively. The figure illustrates how crucial the choice of this value is for the individual score distribution, where the resulting budget scores for these different choices of minimum realistic bid are given. Changing the value of the minimum realistic bid will not change the order of the scores, but it will hugely influence the magnitude of the differences in scores between applicants, which will be decisive when weighted and summed with the other categories. For example, LOUD scored 6 points higher that 24syw with the Board’s choice of minimum bid, which would be reduced to only a 1 point advantage using 50% as the minimum realistic bid. Choosing a minimum realistic bid that is approximately equal to the minimum bid will induce a large range between scores even if the bids only differ by small relative amounts, and thus the budget category will most likely be the decisive factor even if weighted low. Notice in Table 1 how the supposedly most important category weighted with 40%, the programme plan, only has scores varying between 4 and 7, and thus, after weighting, only varies by 1.2 points, whereas the budget category, with only 25% weighting, provides differences in final points of 1.75.

The Board did not specify the minimum realistic bid in advance. This was considered part of the competition. However, by defining the minimum realistic bid based on the actual bids received (instead of in relation to the known maximum possible funding), the resulting scores depend heavily on this value. In fact, the scores received can be manipulated by varying this value, see Figure 1. Nevertheless, the Board did not ensure that their minimum realistic bid, while not revealed in advance to the applicants, could be verified independently afterwards by a third party. In fact, the Board set the minimum realistic bid size (at 250 DKKM) at its meeting on 22 October after having seen and scored all the applications, which they stated openly. They further argued that the ordering of the scores would be the same regardless of the value of the minimum realistic bid, completely ignoring that the differences between scores matter when weighted together with the other categories. It is simply not an acceptable procedure. It is unknown how the minimum realistic bid was decided upon, but it is worrying that it was exactly the lowest bid rounded down to nearest 10 DKKM, which is the choice that exaggerates the differences the most.

Choosing essential parameters (here the threshold for a minimum realistic bid) that influence the outcome of an analysis after having seen the data is usually suspect in statistical inference; one of the most recognisable and notorious examples of this is \( p \)-hacking, where data is iteratively subsetted and manipulated until an analysis yields a significant \( p \)-value of 0.05. Even if done with the best of intentions and even if...
Each error in itself might seem innocent. However, they ended up having a possibly huge impact on how taxpayers’ money was spent.

It intuitively makes sense to let choices be guided by information in the data, it can have huge and unforeseen consequences that jeopardise the validity and objectivity of the conclusions drawn from the analysis.

How the magnitude of the difference between two applicants depends on the bid of a third applicant

There is another subtle problem built into the construction of the budget evaluation through the interpolation formula. Since dk4 scored low on all categories other than the budget, it is clear they could never win the bid. The interest therefore gathered around Radio24syv and LOUD, which were in a close race. They differed by 1.75 points after weighting in the budget category. Assume that the minimum realistic bid is fixed at 250 DKKM. How would points have been allocated to Radio24syv and LOUD if dk4 had bid something slightly lower, for example 240 DKKM, which is only 6% lower than its actual 256.25 DKKM bid? Since the smallest bid is included in the calculation of the slope, the point allocation after weighting would now be 0 points for Radio24syv, 0.8 points for LOUD and of course 2 points for dk4 (it is still the smallest bid). Now the difference between Radio24syv and LOUD is only 0.8 points in the budget, against 1.75 before.

Let us be clear: the relative rating of two of the bids depends on the third bid! Readers wanting to fully explore how changes in the full algorithm may affect point distributions and bid outcome can use the spreadsheet in our repository (bit.ly/415zOGT), where all relevant documents can be found.

Correcting all the errors changes the winner

Let us now see what happens when we correct all the errors and poor choices discussed above. To recap, this means:

- Doing rounding, weighting and averaging correctly.
- Calculating the slope coefficient in a way that avoids “irrelevant third party effects”.
- Specifying the minimum realistic bid exogenously, in advance, and at a reasonable level. As discussed, using a value that is close to the maximum available budget creates instability (a “winner takes all” effect among applicants) and implicitly leads to an over weighting of the budget category.

The result is shown in Figure 2, which comes from the spreadsheet implementation where the user can fully explore different choices and combinations. We find that implementing these changes does in fact change the outcome – the winner is now Radio24syv.

This shows that even if (as is the case here) each individual error on its own does not alter the final outcome (i.e., the winner), the combined effect of fixing all the errors does.

Postscript: Light in the gloom

The DAB licence decision caused quite a stir in the Danish media, and the points raised above (and others related to the content of the evaluations) were presented to the Board. One of the Board’s rebuttals was: “We acknowledge that the calculations could have been more mathematically correct, ... but it would not have mattered for the final result.” As we saw in the previous section, you cannot argue in this marginal, adjust-things-one-at-a-time way; you must fix everything that is wrong and then see what the combined effect is. The flaws in the tender process eventually gained political traction. Nobody would (or could) change the decision in the DAB case, but in late 2020 a number of changes (different procedures, new board members) were implemented after the criticisms were raised.

This story is a particularly stark example of a series of numerical errors, where each error in itself might seem innocent and our criticisms might appear pedantic. However, as shown above, they ended up having a possibly huge impact on how taxpayers’ money was spent. Our aim is not to say that Radio24syv is the only possible winner, but to show that errors matter and, importantly, that it is crucial for a fair evaluation that values such as the minimum realistic bid are determined before the bids are known. When all other errors are corrected, it turns out that the choice of this minimum bid is the decisive factor for who wins the bid. In a democratic society with a public tender, obviously this value should be decided beforehand, to rule out any biased decisions.

We hope to have demonstrated that numerical literacy is an important skill, one that we all need in order to understand, evaluate and participate in democratic discussions and decisions, and for that we need to be trained in understanding numbers as much as being able to read and write.

Acknowledgement

We thank Carl Winsløw for providing the example of the exam scores.