MAKING SOCIETIES DATA LITERATE AT LARGE SCALE BY USING ONLINE NEWS MEDIA

Pim Bellinga and Thijs Gillebaart
I Hate Statistics, Amsterdam, the Netherlands
pim@ihatestatistics.com

In informed societies, citizens require the means to inform themselves, as well as adequate skills to interpret the information. In most countries, the data literacy/numeracy of citizens is under-developed. We propose to use (online) news media as one of the channels to increase the data literacy of citizens. This paper presents one example of an interactive explainer that explains sampling variation and the need for error margins in polls. This explainer has been published in several Dutch news media. In just a few days, thousands of readers completed the explainer and reactions have been enthusiastic and encouraging. Currently, it is still unknown how well the readers now comprehend the concepts and if such explainers can be created less labor intensively. Overall, we see interactive explainers in mass media as a promising direction forward to help societies become more data literate.

NEED: INFORMED SOCIETIES REQUIRE DATA LITERACY SKILLS

In informed societies, citizens require the means to inform themselves as well as adequate skills to interpret the information. To ability to read has therefore been a central pillar in education. Interpreting quantitative information remains a challenge though, as most citizens are not able to recognize key data/statistical concepts and thus are prone to draw incorrect conclusions. (Ironically, is has been hard to find good research data to prove this claim, but given the activities and existence of – for example - the International Statistical Literacy Project, we see this as an important challenge that needs our utmost attention.)

There are also many different terms and definitions for these data and reasoning skills, such as data literacy, statistical literacy or quantitative literacy and numeracy. Although lots of papers explore these definitions to uncover useful distinctions (Ben-Zvi, 2004, Schield, 2004), in this paper we will simply use these terms interchangeably.

STRATEGY: ADD ONLINE NEWS MEDIA AS AN ADDITIONAL OUTREACH CHANNEL

In order to increase data literacy, most educators and authors on statistical literacy have focused their attention on traditional education. This makes sense: laying foundations in primary/high school and universities is essential for building a data literate society.
But focusing mainly on traditional education leaves out a large of group of citizens:

1. Everyone that has left school and currently does not master the essential concepts in statistics. In the current situation, they will not likely learn about them later in life.
2. Knowledge dissipates when not actively used (Averell, 2011). Therefore, constant revisits of the essentials concepts are required, which traditional education is currently not providing.

Therefore, we argue for a broader perspective on outreach, focusing our attention not only on traditional education, but to include all citizens who are not enrolled in schools or trainings, and are currently not inherently interested in data and statistics.

To achieve this, we should make use of multiple channels. These channels should be able to reach a large group of people that are not reached currently. This paper does not include a comprehensive overview of possible channels, so we have focused on online news media as one such channel. Online news media reach a mass audience, which includes a large group that is not enrolled in traditional education. Another possible channel could be social media, as a large part of young people now use social media as their main source of news. (Mitchell, 2015)

DESIGN CRITERIA: FIVE ESSENTIAL PROPERTIES

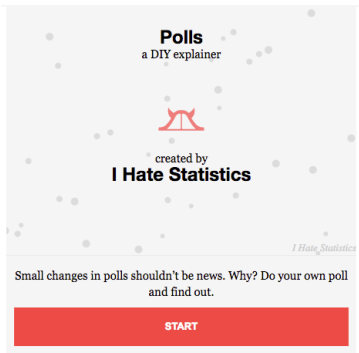
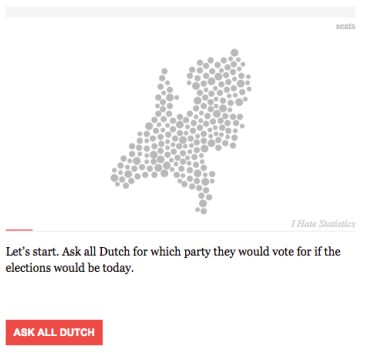
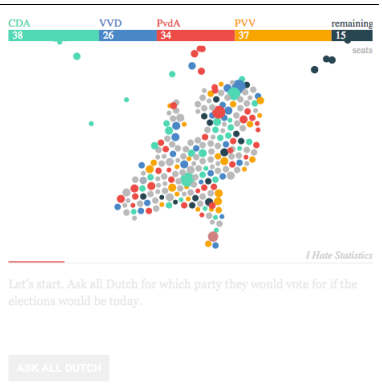
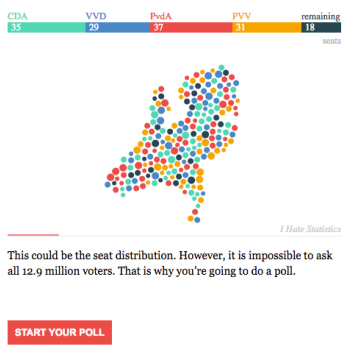
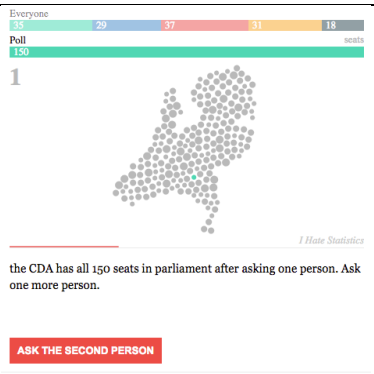
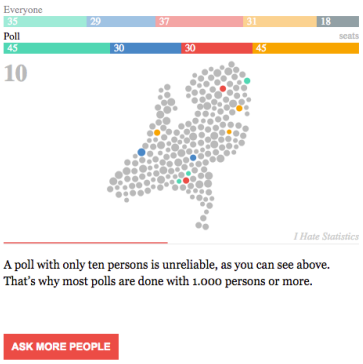
We propose to use the online news media that citizens use to inform themselves to help citizens become data literate as well. We claim that in order to do so successfully, this education should be: 1) Short, 2) Visual 3) Interactive 4) Relevant 5) Discoverable

In: A Molnar (Ed.), *Teaching Statistics in a Data Rich World*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2017, Rabat, Morocco. iase-web.org/Conference_Proceedings.php

1. Short; because people have short attention spans online
2. Visual: because people do not read text carefully when casually browsing
3. Interactive: because people have different speeds at which they can comprehend new material
4. Relevant: because the target group will only be interested in the possibilities and solutions that statistics offers, instead of learning about abstract statistics.
5. Discoverable: because the target group will not be actively searching for material on statistics and do not have teachers suggesting material.

DESIGN EXAMPLE: INTERACTIVE EXPLAINERS IN ONLINE NEWS MEDIA

In March 2017, we collaborated with a Dutch online journalistic platform (www.thecorrespondent.com/) with the goal to educate their readers on the limitations of election polls and the importance of the error margins. The 2-3 minute interactive explainer was embedded in the online article and accessible on all major devices, including mobile phones. Screenshots from the explainer can be found in Table 1.

<table>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>1</td><td>2</td><td>3</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>4</td><td>5</td><td>6</td></tr>
</table>

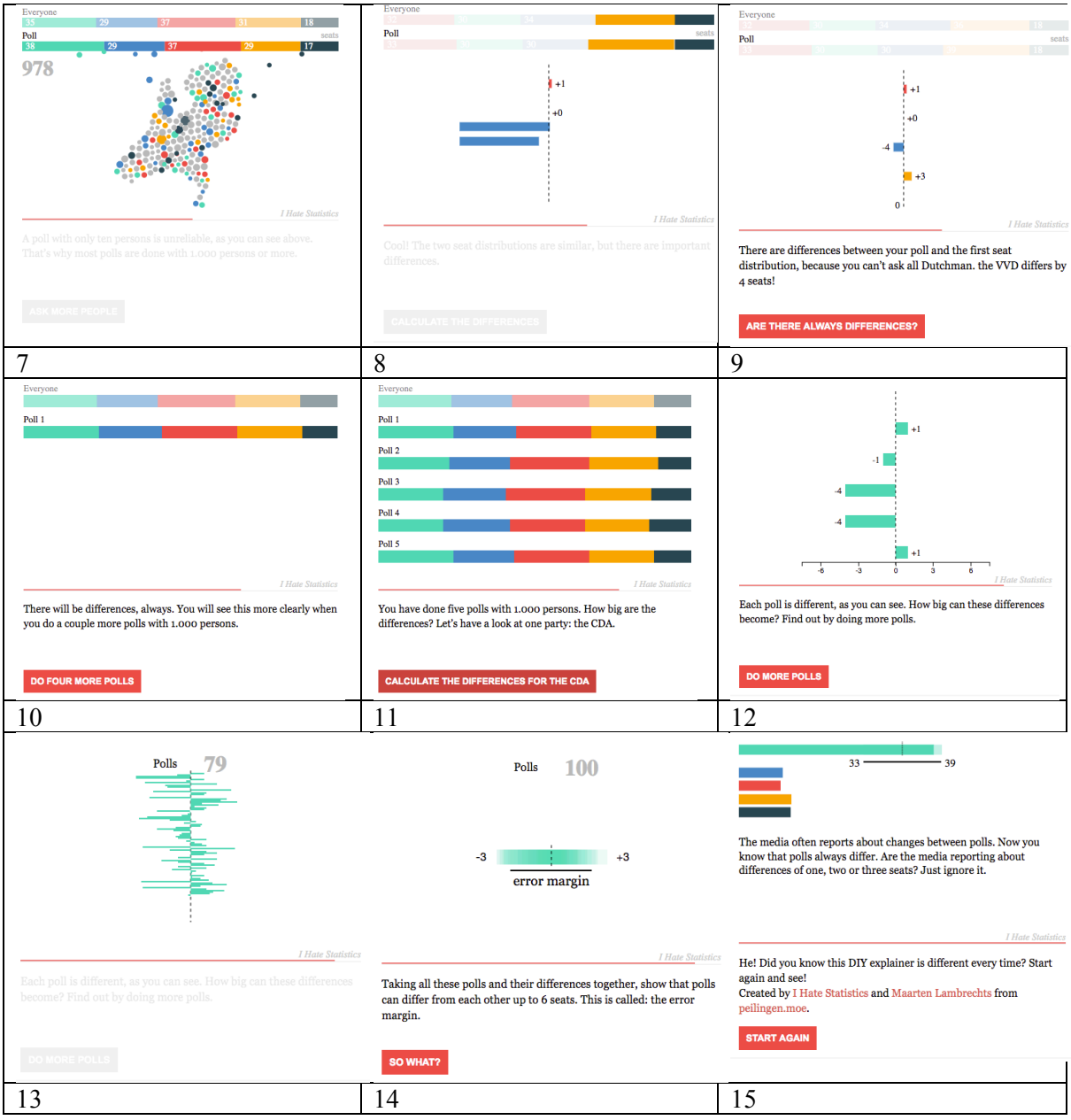


Table 1: screenshots of the interactive explainer on election polls, covering sampling error and the need for margins of error. The explainer uses motion to support important steps. This is hard to map to paper. An online demo of the interactive explainer can be found at www.snapstat.org

RESULTS: OUTCOMES OF THIS PILOT ON SAMPLING AND ERROR MARGINS

Here are some results of the first pilot:

- 5051 readers started the explainer (32%, out of 15.715 who read the article)
- 3366 readers finished the explainer (67%)
- we have spent approximately 240 hours with a team of 2-3 people developing and testing the explainer.
- reactions from readers were positive. One example from Twitter, translated from Dutch: "Thanks! This is the clearest explanation on sampling that I have seen so far!" - Stefan Breet

DESIGN CONSIDERATIONS

We had to make a lot of design decisions along the way. We will briefly explain the most important decisions:

Why we chose election polls as subject: in democratic countries, election polls are major news items. They are covered extensively in the news (see criterion 4: relevance).

However, they are also often incorrectly interpreted, by the public and journalists as well. Although we do not have numbers on this, we feel confident to state that not a lot of people understand the need for margins of errors. In addition, concepts such as sampling variation and margins of error hard to explain in words (hence we feel there is a need for criteria 2 and 3: visual and interactive). Finally, as election polls are conducted and repeated regularly, the explainer can potentially be reused on multiple occasions.

Why we did not address representativeness: a frequently asked question has been: "an even bigger source of error/bias in election polls is unbalanced polls. Why did you not address representativeness?" We agree that this is also an important concept to understand and take into consideration. The reason we left it out is that the explainer would simply become too lengthy (see criterion 1).

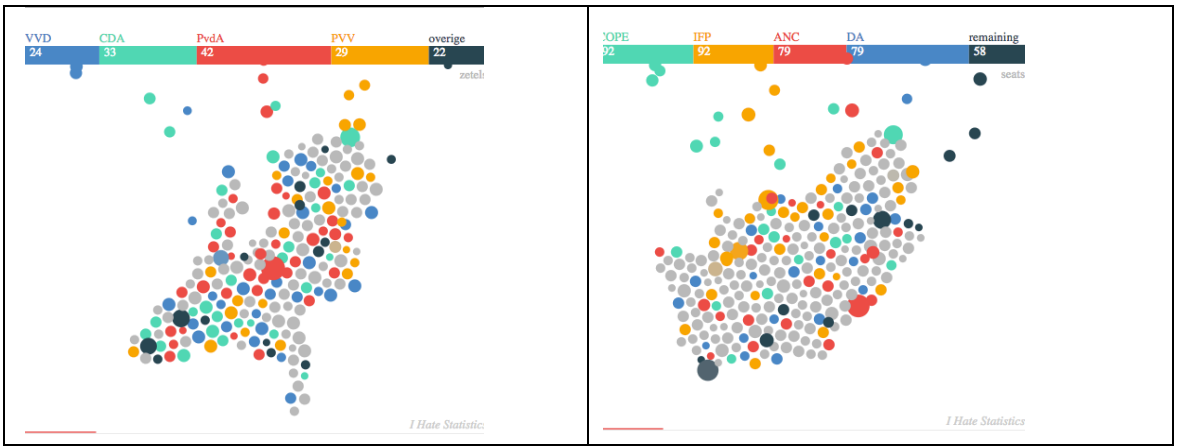
Why we used parliament seats instead of percentages: in the Netherlands, the results of election polls are always communicated in parliament seats. So to make our explainer feel more familiar, we used parliament seats as well. This choice will likely be different for other countries. Our collaborator Maarten Lambrechts pointed out that in Belgium, poll results are always communicated in percentages.

How we chose the different steps: we tested the explainer with potential readers dozens of times. Iteratively we adapted the explainer until most people said they understood the content and did not become bored while engaging with the material. At some points, we invested heavily in visual scaffolding. One example is subtracting the sample from the population to obtain differences (see Table 1, screen 8). Another example is the step to generalize the obtained differences to all possible polls, which we spread out over four screens (Table 1, screens 10-14)

FREE TO USE AND ADAPTABLE FOR OTHER COUNTRIES

A key feature from this explainer is that it is relatively easy to adapt it to a new country (for example South Africa, see Figure 1) The bubbles in the country are generated based on a shape file and the number of parties and their names are variables. Teachers or organizations that want to use the explainer are free to do so. Interested teachers or organizations can contact us if they want us to adapt the explainer to a new country.

One remark is that currently only systems of proportional representation are supported. In countries with simple plurality voting systems (such as the USA), this explainer might feel less relevant and realistic, as end-results in those elections are often communicated differently than the form used in this explainer.



Dutch version | South African version

Figure 1: the explainer can be used for other countries with proportional voting systems.

LIMITATIONS: UNKNOWN POST-MASTERY AND BIASED TEST GROUP

We see two main limitations regarding these results:

1. In this setup we cannot check whether the readers really grasp the concepts sampling error and error margins or whether they only say they do. It would be interesting to see if we could embed a pre- and post-test within the explainer. But without losing readers, this will be hard to achieve. Replicating this pilot study in a lab setting where participants are required to take a separate pre- and post-test would be interesting direction for research.
2. The audience of this particular news outlet we collaborated with is relatively highly educated. This introduces a concern: are these results representative for the rest of the population? This concern is mainly on two fronts:
   a. would the rest of the population find, read and engage with this explainer in a similar fashion as the readers from this particular news outlet?
   b. would those readers – say they – grasp the steps and concepts as easily as the current group?

Performing experiments with pre- and post-tests and pilots with different groups would be needed to confirm or negate these concerns.

DISCUSSION: APPLICABILITY TO OTHER STATISTICAL CONCEPTS

We see interactive explainers – published in online news media and perhaps in social media as well - as a promising method to improve the statistical literacy of citizens. This pilot focused on sampling variation and margins of error. One question we have is: which other statistical concepts could benefit from an interactive explainer as well? Possible examples:

- What are confounding factors?
- Why is it so important that assignment is random in experiments?
- What is the median income and what is the difference with the mean income?
- What is regression and how does it work?

We are interested to hear which data/statistical concepts and numeracy skills the field deems important that all citizens should be aware of them and master them.

In the light of possibly developing more explainers we should also discuss the significant time required to develop these explainers. If they are being reused often and in multiple contexts, then the investment could be well worth it. However, if these explainers work and are being requested for more statistical concepts, then the labor involved will be high. We would warmly welcome suggestions on how to decrease the amount of time required to develop these explainers.

FUTURE RESEARCH

More research is required to investigate the extent to which these types of explainers would appeal to mass audiences and whether they really help people to grasp statistical concepts. We see three main areas of research that would push this investigation further:

1. Find out whether people really learn from the explainer. For example using well-calibrated pre- and post-tests.
2. A limitation of this pilot was the possibly unrepresentative audience. It would be interesting to see whether these results can be generalized to larger and more diverse audiences.
3. A major disadvantage of these types of explainers are the labor-intensive nature. If these explainers turn out to be useful, we should explore ways to reduce the amount of time required to create and adapt them.

IMPLICATIONS: EDUCATING SOCIETIES, LIFE-LONG, AT SCALE

We suggest that in addition to educational efforts at schools and universities, effort should be spent educating citizens as well. We propose (online) news media as a good channel for reaching citizens: the reach can be large, and the audience will include members who might otherwise not get into contact with data literacy concepts. Open questions are: 1) how well do the readers understand the margin of error now, 2) can these findings be extended to a broader audience 3) whether material can be developed less labor intensively. To conclude, we see explainers on data literacy concepts for citizens using online news media as a channel a promising method of creating more informed societies that should be explored further.

REFERENCES

Averell, L. & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology (55),* 25–35.

Ben-Zvi, D. & Garfield, J. (2004). Statistical Literacy, Reasoning and Thinking: Goals, Definitions and Challenges. Kluwer Academic Publishers, Dordrecht.

Mitchell, A. (Ed.). (2015). Millennials and Political News. *Pew Research Center, Journalism & Media*. www.journalism.org/2015/06/01/millennials-political-news/

Schield, M. (2004). Information Literacy, Statistical Literacy and Data Literacy. *IASSIST Quarterly Summer/Fall 2004*, 6–11.