THE VIEW FROM THE FRONTLINE: JOURNALIST PERCEPTIONS OF ONLINE COMMENTS AND THE MODERATION PROCESS IN THE "LESS HATE, MORE SPEECH" PROJECT

Raluca Toma, Marina Popescu, Roxana Bodea

less hate more speech

median research centre mrc

ICELAND LIECHTENSTEIN NORWAY eea grants

norway grants

THE VIEW FROM THE FRONTLINE:

JOURNALIST PERCEPTIONS OF ONLINE COMMENTS AND THE MODERATION PROCESS IN THE ”LESS HATE, MORE SPEECH“ PROJECT

Raluca Toma, Marina Popescu, Roxana Bodea

CONTENTS

INTRODUCTION

Top news outlets around the world are experimenting with ways of engaging with their audience and incorporating reader-generated content into the online experience they offer, both in order to build a more loyal readership and due to the added value that reader perspectives can bring (Masullo Chen and Pain, 2016). But in this age of polarisation, where mutual tolerance, respect and civil debate appear to be giving way to aggressive discourse and demonisation of the ”other,“ it is a major challenge to figure out how to maintain spaces for online dialogue that are free for discussion but also inclusive and ”safe“ for everybody who may wish to get involved. The needs of news organizations can appear to conflict; it is difficult to allow readers to express themselves but also stay true to organizational values, while also dealing with limited resources. These challenges have spurred creative approaches and interesting collaborations, such as the Engaging News Project or the Coral Project.

This report completes a series of reports on the online comment-focused collaboration between Median Research Centre (MRC) and major Romanian sports newspaper Gazeta Sporturilor (GSP.ro). The report offers some highlights on how the newsroom and team of moderators perceived online comments, what they thought about and how they engaged with comment moderation, how their own views shaped the process and what they feel about moderation, comments and engagement at the end of this experiment.

In the ”Less Hate, More Speech“ project, we at Median Research Centre (MRC) came together with Gazeta Sporturilor (GSP) in search of solutions for online comment moderation and the engagement of commenters.[1] In order to see whether and how online discourse can change as a result of the establishment and enforcement of new norms, we designed a moderation system through which comments on four partner websites were checked both automatically and manually, and we gradually developed and disseminated new rules for the comment section.[2] In addition to moderating comments, we also made several experiments in the comment section, to test other strategies for encouraging good behaviour and thoughtful comments. This was all done with the help of five moderators, young journalists from Gazeta Sporturilor (GSP), who worked together with the researchers of MRC, mainly by discussing comments in weekly ”moderation meetings“. A helpful infographic of what happened during this 15-month process and the moderation experiments that were implemented can be seen here. More details about how the journalists and researchers came together in this project and how the mixed teams worked can be found in the report ”Engaging with the Other“.

One of the unique aspects of the ”Less Hate“ experiment was the amount of access GSP gave the research team, making this partnership uniquely well-documented. We are grateful that the journalists responded to multiple surveys, permitted certain discussions to be recorded and wrote about their opinions or experiences on multiple occasions.

With data from several newsroom surveys, excerpts from written impressions of the moderators and with stories from comment discussions as remembered by the researchers, this report follows four main questions:

- How did the journalists see comments, the people who make them and online engagement, at the beginning and at the end of the project?
- Did the journalists ”buy into“ the ”Less Hate, More Speech“ moderation, and how did their views on it evolve?
- What were the predispositions and beliefs of the journalists, and how did they influence the moderation process?
- Beyond comment moderation, do the journalists see other ways they can foster more civil, tolerant and thoughtful discussion, and how did the moderation influence the way they see their work?

[1] The name of the project is ”Less Hate, More Speech: An Experimental and Comparative Study in Media and Political Elites' Ability to Nurture Civil, Tolerant, Pro-Democratic Citizens.“ It is financed through the EEA Financial Mechanism 2009-2014, in the programme entitled ”Research in Priority Sectors,“ and was implemented by Median Research Centre, in partnership with Central European University and the University of Bergen.

[2] The partner websites were Gazeta Sporturilor's high-traffic site GSP.ro; Tolo.ro, a popular public affairs and investigation blog by GSP's editor-in-chef Cătălin Tolontan; Blogsport.ro, GSP's op-ed hosting website; and Paginademedia.ro, a site focusing on the latest in the news and entertainment industries.

A NOTE ON DATA AND INTERPRETATION

The comment moderation that began thanks to the ”Less Hate, More Speech“ project is still ongoing. But the ”supervised“ moderation phase, during which the comment section experiments took place and the researchers observed the moderation, gave feedback and instructions to the five moderators and worked together to set up commenting rules and moderation procedures that could work in the long run, started in May 2015 and ended in June 2016.

Both before, during and after this phase, we as MRC researchers were interested in understanding the experience of both the moderators and the journalists involved in the project. Beyond documenting the moderation-related discussions and requesting written feedback on several occasions from the moderators, we also deployed surveys at certain points. The journalists filled out three surveys: one in December 2014, before the moderation process began on the main website (GSP.ro); one in May-June 2015, shortly after the introduction of the moderation system on GSP.ro; and one in February-March 2017, a few months after the end of the supervised moderation period on all websites.[3] The moderators (five people working in shifts) were included in all the surveys of the journalists, but they also answered two separate surveys, designed just for them. The first one was at the beginning of the supervised moderation period, in May 2015, and the second one, the so-called exit survey, was done in July 2016.

This report is based on survey data, analysed by the MRC researchers, as well as stories recounted by the moderators, from which we selected certain excerpts, and on the notes and memories of the researchers who participated in the moderation discussions. As a result, rather than being a first-hand account of the journalists' experience, this report presents the journalists' perceptions as recorded and interpreted by the research team.

[3] The journalists surveyed were part of the GSP.ro, Tolo.ro and Blogsport.ro newsroom.

THE NEWSROOM AND THE MODERATORS AT THE START OF SUPERVISED MODERATION

ENGAGE, BUT KEEP YOUR DISTANCE

At the start of the collaboration between MRC and GSP, there was significant, if not unanimous, newsroom interest in reader engagement. When questioned as part of the newsroom survey implemented shortly after the beginning of the moderation experiment, almost a quarter of the journalists firmly believed spending time on interaction with the readers is a worthwhile activity.[4] Another 40% were somewhat supportive of the idea.[5] Half of the newsroom agreed that for journalists today, interactivity with the readers should be just as important as producing articles. It is noteworthy, though, that only 22% of the respondents thought that interactivity was as important as writing articles in this particular newsroom. Compared to men, women saw dedicating time to interaction with the readers as more worthwhile, but younger journalists were no more interested in this than their older colleagues.[6]

A non-trivial share of the journalists saw it as important if an article became viral, got lots of clicks and likes or positive feedback in the comment section. To others, these things had very little significance. When asked to rate various journalistic accomplishments according to their level of importance or priority, 36% of the respondents to the first survey awarded having a viral article the highest level of importance. But looking at the entire spectrum of opinions on this, going viral appeared to elicit a fairly even split between appreciation and apathy. An equal share of people (44.4%) rated a viral article as being one of their top two priorities (highest rating or second highest rating on a five-point scale) and a bottom-two priority (lowest or second-to-lowest rating). By contrast, getting professional awards was more uniformly down-graded as a goal. It was rated as a top-two priority by almost 28% of the respondents, but a much greater share (almost 42%) said it had a low level of importance.

That an article should get many positive comments was most important to 19%, but an even greater share thought this was the least important thing (24%). Similarly, getting many clicks and likes was a top concern for 14% and, conversely, for 17% of the journalists it represented the least of their worries.
Younger respondents (below the age of 33) saw having an article with positive comments as a more important achievement than their older colleagues, but there were no age-based differences in the ratings of other journalistic achievements.[7]

COMMENTS: CAN'T LIVE WITH THEM, CAN'T LIVE WITHOUT THEM

A whopping 93% believed that comments distract attention from the content of an article.

Although the journalists said they didn't think it was very important for an article to get a lot of positive reader comments, most of them still read the comments pretty frequently. Three quarters of the newsroom said they looked below the line of a majority or even all of their own articles. A few admitted to reading these

[4] This was the second newsroom survey, completed by 72% of the newsroom (26 respondents) between May 19 and June 25 2015. The ″unsupervised″ moderation as part of a new comment management system developed in the project had begun in April 2015, and a month later, the researchers and moderators began meeting, thereby starting the ″supervised″ moderation phase.

[5] On the other hand, 12% were fully convinced that it was not worth trying to interact with the readers on the website, and a further 20% tended to agree with this sentiment.

[6] In wave 2, women found dedicating time to interaction with the readers more worthwhile than men (1.86 mean on 0-10 scale where 10=completely agree, compared to 4.56 mean for the men), t(23)=2.334, p=.030.

[7] In the first newsroom survey, those below the age of 33 found positive comments slightly more important than those 33 and above (mean = 2, on a scale of 1-5, where 1 means top priority, compared to mean = 3.13), t(24)= -2.021, p=.055.

submissions several times a day. When it came to other GSP articles, 46% said they read the comments at least once a day. There was also a small contingent of the newsroom that steered well away from the comment section: 16% reported they read the comments on their pieces or other GSP articles a few times a year, at most.

Overall, the journalists seemed to find comments somewhat important for getting an idea of the opinions of the audience and feedback on their work and less important for getting information.[8] An index constructed out of their ratings of the importance of comments for providing information (Chronbach's Alpha: .849) revealed that overall, for this kind of use, the journalists gave comments an average importance (5.1 mean on a 0-10 scale).[9] Meanwhile, the index of their rating of comments in terms of offering audience feedback revealed a greater average importance rating, of 6.25 on the same 0-10 scale (Chronbach's Alpha: .743).[10]

We also asked the journalists to reflect on what experiences they had had as a result of audience feedback in the comment section. Of the possible good outcomes we presented, more than half had experienced all three at some point. Almost 68 % had found an idea for an article in the comment section at least once, and over 78% had found information that they used in articles. By contrast, only 24% admitted they had ever changed an article's structure due to anticipating negative feedback and just 30% said that an editor had changed one of their articles in order to avoid problems.

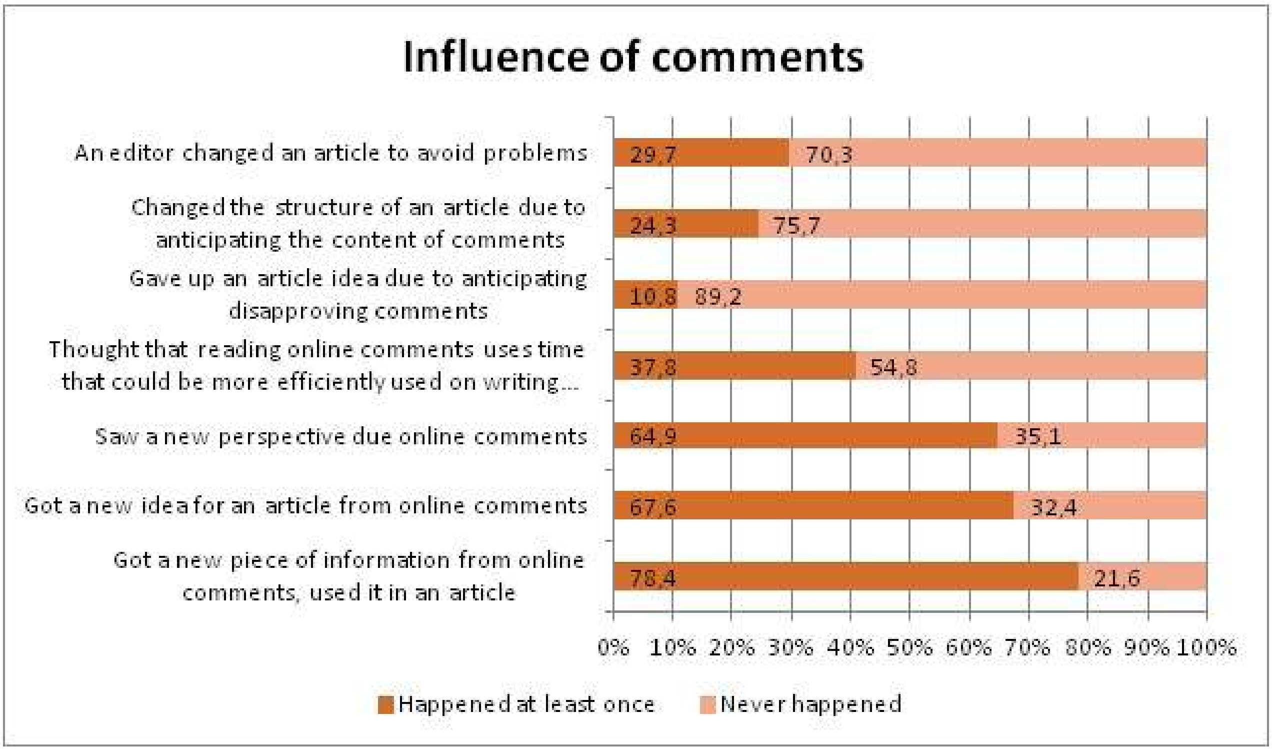

Fig. 1. Occurrence of various experiences due to existence of audience feedback. First newsroom survey (December 2014-January 2015)

[8] A factor analysis revealed that the items load fairly well on two factors, one of which seems related to getting new information and ideas and the other one to getting feedback and a feel for the readers' opinions or public opinion in general.

[9] The items included in this index assessed the importance of comments for: a) generating ideas for new articles; b) making the life of journalists more varied; c) getting useful and interesting information from readers.

[10] The items included in this second index related to the importance of comments: a) as a method of evaluating the value of an article; b) as a feedback mechanism in case you made a mistake in an article; c) as an indicator of public opinion on a subject; d) to hear people's opinions.

Still, when considering comments, the journalists seemed more preoccupied with their downsides than with their added value to the newsroom or the reading experience. A whopping 93%, meaning all but two journalists, believed that comments distract attention from the content of an article, and 68% thought that comments could change a reader's perspective on a journalistic piece.[11] Eight out of ten respondents agreed that comments polarise the readers and split them in hostile camps, and a similar share believed that comments were likely to be on the mind of interviewees and to influence how those people answer the reporters' questions.

The top complaints about the comment section referred to incivility. The great majority of the journalists mentioned the appearance of curse words, insults and aggression and hostility more broadly defined as the most bothersome aspects. They also bemoaned the general negativity of the commenters and the accusations they make against the authors, as well as the existence of user anonymity itself (78% thought comments should not be anonymous). Only a few of them also listed discriminatory and intolerant discourse towards minorities or women as being among the most frustrating aspects of comments.

There was also a significant amount of pessimism about incivility in the comment section and lack of trust in the commenters' ability to control themselves or change their behaviour. For instance, 67% of the newsroom believed that the comments are inevitably flooded by racist, violent or brutal and intolerant language. A sliver of hope emerged from the idea of a newsroom presence tamping down the nastier instincts of the commenters: 61% agreed that if a journalist was present, commenters would be more likely to act civil.

Perhaps surprisingly, though they complained about the tone of the comments and seemed pessimistic about the chances of greater civility, almost three quarters of the journalists did not actually want to get rid of them, but simply to intervene and moderate the nasty ones. Two in ten actually supported the idea of hosting comments on the website with no restrictions whatsoever **(Fig. 2)**.

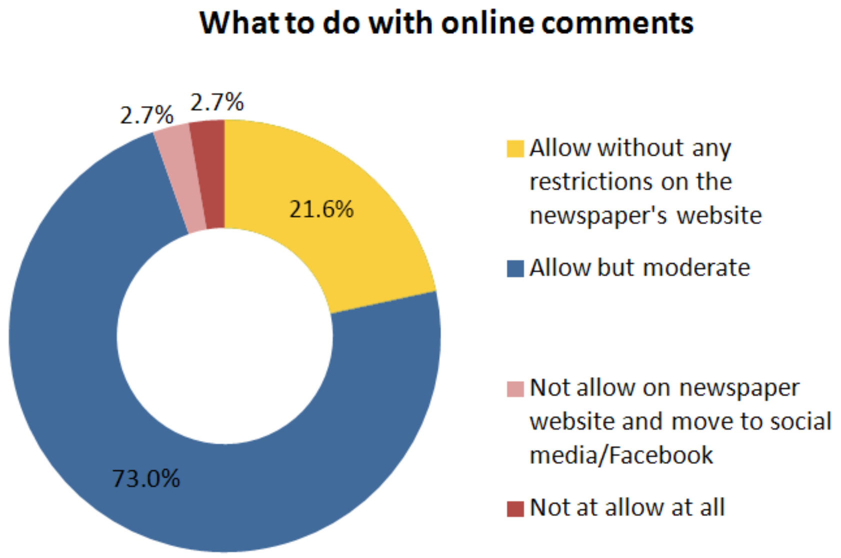

Fig. 2. Preference on what to do with online comments, when choosing between four pre-defined options. First newsroom survey (December 2014-January 2015).

[11] Research in fact backs up this perception that comments can influence how readers perceive the content of an article. See for instance Anderson et al. 2013.

VIEWS ON MODERATION

Most GSP journalists (67%) thought that the newspaper would have more readers and comments if hate speech and intolerant discourse were not allowed in the comment section.

Given the newsroom's poor assessment of the state of the comment section and their unwillingness to give up on comments altogether, one would expect to see them fully on board with the moderation experiment. Indeed, the newsroom seemed fairly supportive of the idea. Only 22% thought moderation would be against free speech and against the principles of the newsroom. This was roughly the same amount of people who thought comments should be left on the website with no restrictions.[12] Nor did most of them think an efficient moderation process would be prohibitively expensive.[13]

Most GSP journalists (67%) thought that the newspaper would have more readers and comments if hate speech and intolerant discourse were not allowed in the comment section. But they were not so confident that it was possible to distinguish between hate speech or intolerance and simply aggressive comments; only 19% thought this could be done, and another 22% were unsure one way or another **(Fig. 3)**.

There was a healthy degree of scepticism even in the ranks of those who would become moderators. Four out of five were of the opinion that it would not be possible to totally and confidently separate hate speech and prejudice from the rest of the aggressive comments. Two of them also leaned towards the belief that moderation was against the newsroom's principles and against freedom of speech. Interestingly, men tended to agree that moderation is against freedom of speech and newsroom values more than women, although this difference was only barely statistically significant.[14]

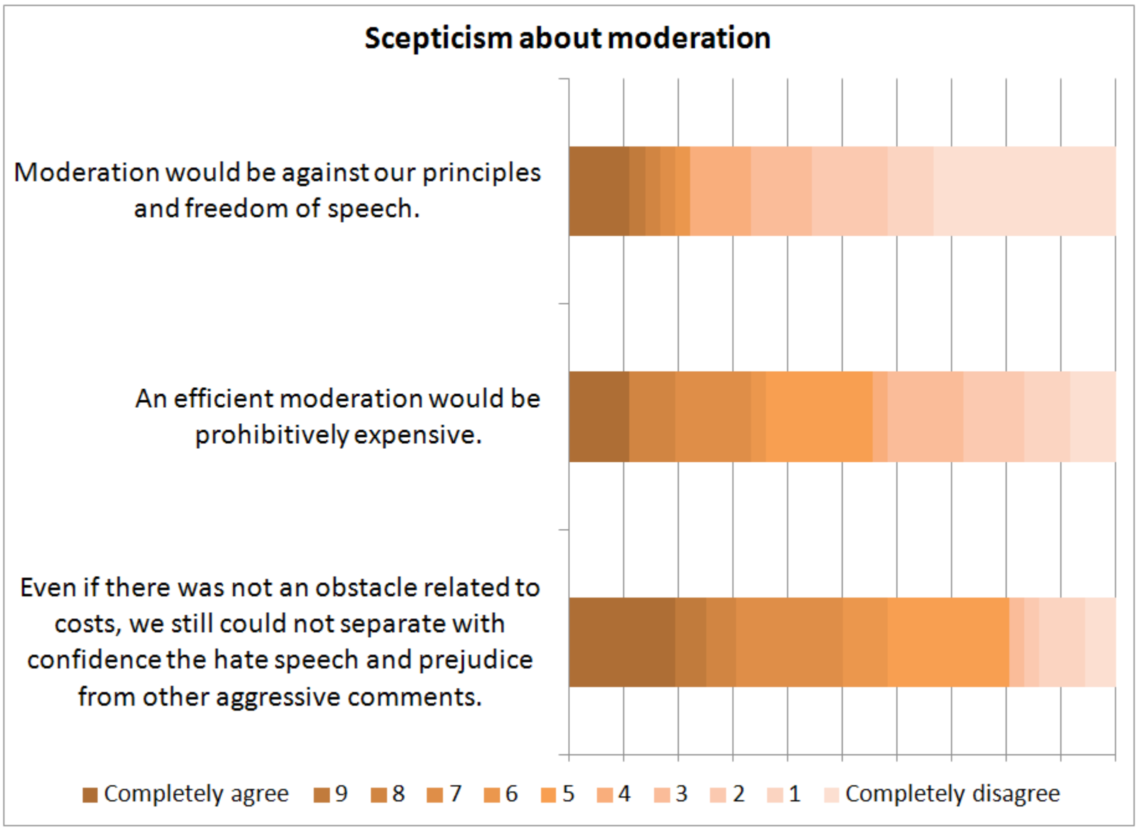

Fig. 3. Degree of scepticism about moderation, rating each statement on a scale from 0 to 10, where 0 means "completely disagree" and 10 "completely agree". First newsroom survey (December 2014-January 2015).

[12] 67% disagreed with this view, and the remaining respondents were on the fence.

[13] 44% disagreed, while 19% were undecided.

[14] In the first survey, men agreed more than women that moderation is against freedom of speech and newsroom principles. The mean agreement among women was 1.63 on a 0-10 scale, where 10 means complete agreement, compared to a mean agreement of 3.64 for the men: t(34)=1.829, p = .085.

TOLERANCE, STEREOTYPES AND OTHER TRAITS IN THE NEWSROOM

Other pre-existing views would go on to influence the moderation process and its success in the eyes of the journalists. For instance, from the beginning it appeared that many members of the newsroom were more irked by vulgarity, insults and other forms of aggression in the comment section than they were bothered by racism, sexism or other intolerant discourse, which the ”Less Hate, More Speech“ moderation was also designed to address. Particularly in the case of the moderator-journalists, their own ”baggage“ of preconceived notions or motivations shaped the collaboration with the researchers and the relationship with the rest of the newsroom. It is worthwhile, then, to note the beliefs and attitudes of the journalists regarding various out-groups, as well as other aspects like interpersonal trust.

We expected that, since the journalists are of this world, they would not think completely differently from the rest of Romanian society. In order to assess the extent to which they embraced negative stereotypes, we included a battery of statements about various groups in the second newsroom survey, taking place not long after the moderation experiment was launched.[i] As anticipated, the newsroom was not entirely free of stereotypical notions about certain groups.

Stereotypes about ethnic Hungarian people elicited the greatest agreement among the journalists, while the journalists responded least well to the poor-people related preconceptions:

- 72% of the journalists leaned towards agreement with Hungarian stereotypes. The newsroom's average agreement with a three-item index of stereotypical statements about ethnic Hungarians (Chronbach's Alpha: .824) was at 5.1 on a scale from 1-7, where 7 meant „strongly agree“ and 4 „neither agree nor disagree“;[15]

- 41% tended to agree with certain Roma-related stereotypes. For a two-item index of Roma-related stereotypes (Chronbach's Alpha: .660), the mean agreement was at 4.3 on the same 1-7 scale;[16]

- Stereotypes about gay people resonated slightly less: 33% of the journalists tended to agree. On the 1-7 scale, there was an average agreement of 3.8 in the newsroom with an index made up of two items (Chronbach's Alpha: .916);[17]

- Agreement with Jewish stereotypes, collapsed in a three-item index (Chronbach's Alpha: .877) was at a similar level, at 3.7. Only 30% of those who answered the questions tended to agree with these stereotypes;[18]

- An index of poor people-related stereotypes fared least well: only 15% of the newsroom tended to agree with these items. The mean agreement with the index was 3.09 (Chronbach's Alpha: .751).[19]

Their journalist's views did not differ statistically significantly from the views of general population, as revealed through the 2016 Less Hate More Speech national survey. The only exception was with regard to two of the three Jewish-people related stereotypes, where the newsroom was less receptive than the general population appears to be. Additionally, some journalists declined to answer the Jewish stereotype items, more than was

[15] The stereotypical statements about ethnic Hungarians were: a) Ethnic Hungarians are not more loyal/faithful to Romania than Hungary; b) Ethnic Hungarians want Transylvania to become part of Hungary; c) Ethnic Hungarians are not willing to speak in Romanian even if they know it.

[16] The stereotype items about the Roma were: a) The Roma break the law more than other people; b) When they start a job, the Roma do not work as hard as other people.

[17] The two Gay-people related stereotype items were formulated using the word «homosexual» to ensure that all respondents understood the question. They were: a) Homosexuality is abnormal; b) Homosexuals ask for certain privileges and are not content to be treated the same as everyone else.

[18] They were: a) Jewish people use the persecutions they suffered in the past to obtain certain favors; b) Jewish people are more preoccupied with money than other people; c) Jewish people control international politics and finances.

[19] Although there were three poor people-related items in the newsroom survey, we only used two of them for the index, since the third did not appear in the national survey, and we could not compare the journalists' views on it with those of the general population. The two were: a) When they get some money, poor people spend it less responsibly than others; b) Gifted people do not stay poor no matter what their circumstances are.

the case with the other stereotype items.[ii] The men agreed more than the women with the gay stereotypes,[20] while journalists younger than 33 were marginally more in agreement with the poor stereotypes than those 33 years old and above.[21]

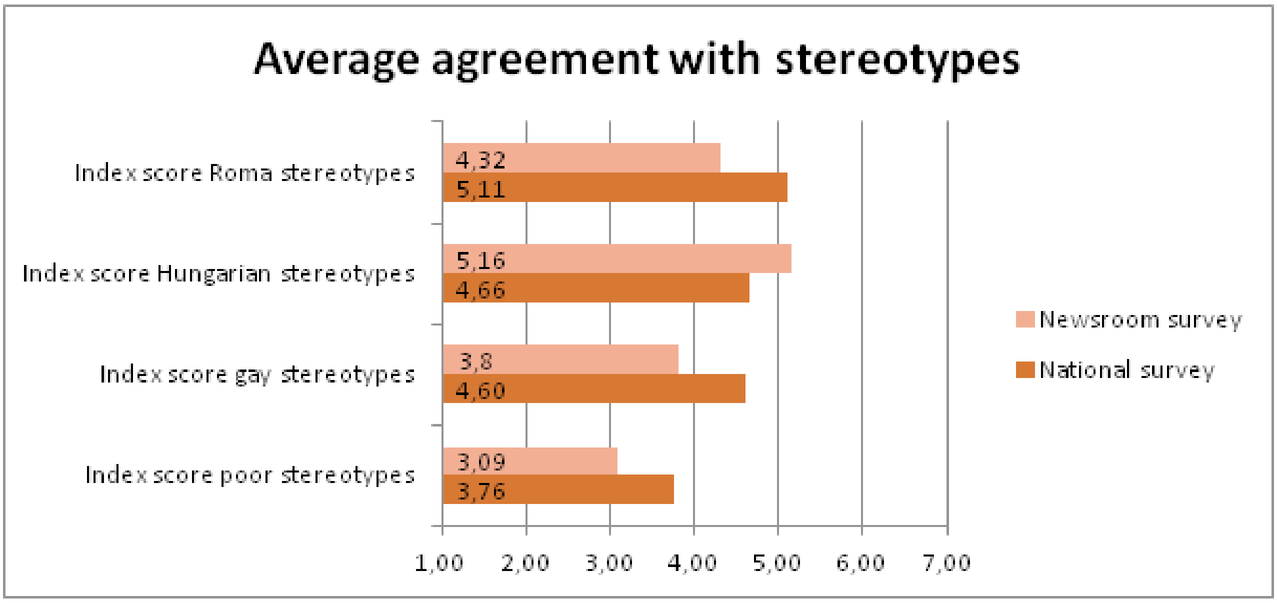

Fig. 4. Index scores of agreement with stereotype items in newsroom and national surveys, on a scale of 1 to 7, where 7 means „strongly agree". None of the differences between the two sets of scores for each index are statistically significant.

A majority of the journalists also exhibited a certain degree of symbolic prejudice.[22] Six out of ten endorsed to some extent the view that Roma people should lift themselves up by their bootstraps. Seven out of ten tended to agree that representatives of the Hungarian minority in Romania are asking for too much, too quickly for their community. The moderators scored somewhat lower on symbolic Roma-related prejudice than their colleagues. A word of caution is necessary, however: these measurements are generally agreed to do a good job of capturing prejudice, but they must be interpreted carefully, since they are designed to address more subtle forms than other, more traditional, measurements, like stereotypes, social distance items, rights denial questions or feeling thermometers. Showing a certain degree of "symbolic prejudice" or "symbolic racism" cannot simply be translated as meaning someone is "a racist" or would behave in a discriminatory manner towards members of a certain minority.

[20] The women's mean was 1.75, compared to 4.29 for the men, on a scale of 1 to 7, where 7 means «completely agree». The mean difference was 2.544L t(19)= 2.285, p = .034.

[21] Mean difference = 1.43 on a 1-7 scale, t(15) = 1.868, p = .088, so just barely statistically significant at the .10 level.

[22] As anti-prejudice norms have taken hold in Western societies, people have become less willing to admit to holding prejudice or preferring to distance themselves from (most) minorities. As traditional measurements reveal increasingly less overt prejudice, measures of "symbolic prejudice" or racial resentment have gained in popularity. They measure prejudice less obtrusively, via opposition to pro-minority policies or beliefs about the sources of lasting inequality (Mendelberg, 2001, 130-131). Although some have argued that such measurements risk confusing conservatism or other value-driven preferences with prejudice, they do appear to capture attitudes distinct from small-government conservatism (Kinder & Mendelberg 2000).

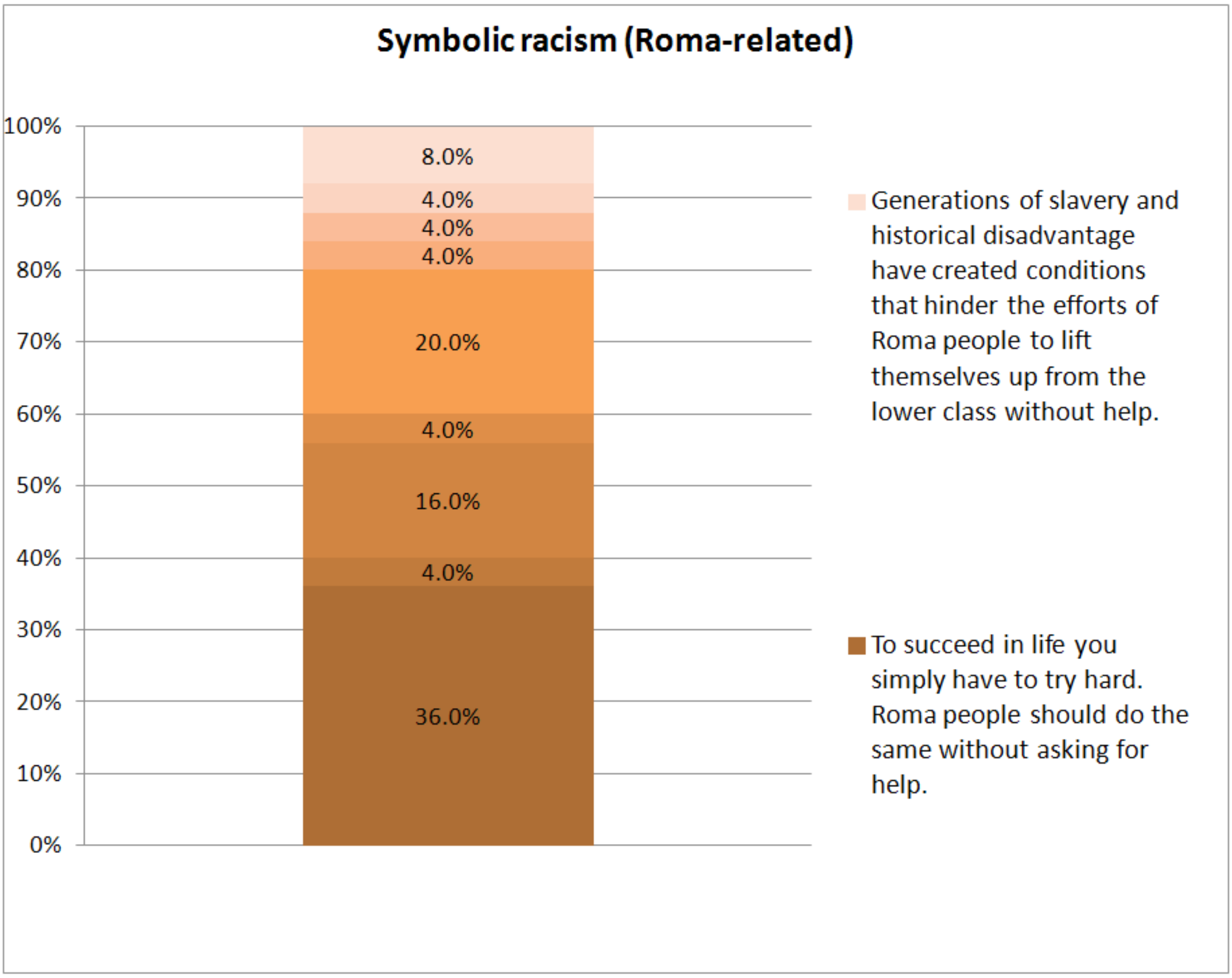

Fig. 5. Distribution of views on an item measuring symbolic racism towards the Roma. The darker the colour the closer one is to the opinion that to succeed in life Roma people should just try hard without asking for help. Second newsroom survey (May-June 2015).

Despite having certain preconceptions about particular groups, the journalists were also quite motivated to control their prejudice.[23] As part of the second newsroom survey, we applied a battery of eight items measuring their discomfort with experiencing racist thoughts or their convictions about the importance of not being prejudiced, as well as their desire to avoid being perceived as racist or prejudiced by others.[iii] While the great majority of them did possess at least a moderate motivation to control their prejudice, their levels of motivation did not differ from the general population.[24]

Of the moderators, only one had average motivation and the other four were highly motivated to control their prejudice.[iv] This, we believe, contributed to their great degree of engagement in the moderation process and to the overall success of the moderator-researcher collaboration, by helping make them particularly willing to engage in thoughtful discussions and question existing beliefs and practices.

[23] To put it simply, the motivation to control prejudice refers to one's recognition, acceptance and, ultimately, internalisation of social norms against prejudice. We can talk about both an external motivation and an internal motivation, the former referring mainly to one's desire not to be perceived as prejudiced and the latter to one's internalisation of anti-prejudice norms, leading to the desire to avoid prejudiced thoughts or actions and to discomfort with the idea of being prejudiced. In our national survey in Romania, using the same measurements we employed in the newsroom, we did not find that there were two separate factors for internal and external motivation items, so therefore we are treating all items together without distinguishing between the internal-focus and the external-focus ones. The scale we used to measure this is inspired by Blinder, Ford and Ivarsflaten (2013).

[24] Their mean agreement to the battery was 5.36, on a scale from 1 to 7, where 7 means «strongly agree,» while the average score in the national survey was 5.54.

What is more, the journalists did not embrace an exclusionary vision of national identity to the extent that the respondents to our national survey did. When asked to rate the idea that Romanians should stick together and not mix with other nations, only 16% tended to agree. Their responses as a whole were statistically significantly less supportive than those of the national survey respondents.[25] The female journalists agreed with this view less than the men.[26] Most surprisingly, those under the age of 33 were significantly more supportive of this exclusionary view. The mean difference between the two age groups was a fairly staggering 2.6-point difference on a 7-point scale - greater than the 1.15 difference between the newsroom as a whole and the national survey respondents.[27]

The newsroom exhibited a level of interpersonal trust that is fairly normal for Romania, where trust in others tends to be very low. 20% of the GSP journalists polled in 2015 agreed that you can trust most people.[v] The newsroom's mean level of interpersonal trust did not differ statistically significantly from the mean level among the respondents to the national survey.[28]

Although attitudes and personality traits are complex and intertwined, it is possible that they influence how journalists think about their audience, especially those commenters who post critical or rude comments below their article. One trait we were interested in was authoritarianism. Put simply, those with authoritarian tendencies "prioritize social order and hierarchies, which bring a sense of control to a chaotic world. Challenges to that order — diversity, influx of outsiders, breakdown of the old order — are experienced as personally threatening because they risk upending the status quo order they equate with basic security" (Taub, 2016).[29] We found low levels of authoritarianism - particularly so and statistically and substantively lower than in the general population when using an item pitting individual liberties against respect for leaders and rules.[30] On this item, the under-33 journalists appeared slightly more authoritarian, but not on the other measure.[31] On another item, the women appeared less authoritarian; specifically, the women tended to agree slightly more than the men that children should learn to think critically and make up their own minds.[32]

[25] Mean agreement was low, at 2.76 on a scale from 1 to 7, where 1 means «strongly disagree,» and it was significantly lower (by 1.153 points) than among the respondents to the national survey, t(24) = -2.968, p = .007.

[26] Women endorsed this exclusionary view significantly less than the men, by 2.048 points on a 7-point scale, t(23) = 4.045, p= . 001.

[27] The over-33 group mean was 1.8, compared to a 4.44 for the under-33), (17)= 3.859, p = .001

[28] The mean answer in the newsroom was 2.32 on a scale from 0 to 10, while the population mean was 3.25. T-test was not significant: t(24) = -1.578, p = .128.

[29] There are a few different ways of trying to get at a subject's authoritarian tendencies, from questions about politics and law and order to questions about child-rearing. In the newsroom survey, we used both a question about children learning critical thinking skills and one pitting respect for rules and leaders against individual liberties. Both of those were also found in our national representative survey implemented in the same year. The disparity between the newsroom and the population was far greater on one of the measurements than on the other. In the newsroom, 16% agreed that «It is more important to encourage respect for existing rules and leaders in society than for individual liberties,» while in the general population 53% agreed. In the newsroom, only 4% disagreed with the statement «It is very important for children to think critically and form their own opinions,» and in the national survey, 5.6% disagreed.

[30] The disparity between the newsroom and the population was far greater on one of the measurements than on the other. In the newsroom, 16% agreed that «It is more important to encourage respect for existing rules and leaders in society than for individual liberties,» while in the general population 53% agreed. Mean agreement in the newsroom was low (3.16), and it was significantly lower, by 1.736, than among the national survey respondents (4.89), t(24) = -4.773, p <.001.
As for the other item, in the newsroom, only 4% disagreed with the statement «It is very important for children to think critically and form their own opinions,» and in the national survey, 5.6% disagreed. The difference between the national survey mean and the journalist survey mean was not statistically significant.

[31] Mean difference between over-33 and under-33 = 1.489, t(17) = 2.031, p=.058.

[32] The women's mean was 7 on the 1-7 scale, where 7 means completely agree, compared to a mean of 6.56 for the men (t(23) = -1.810, p = .088).

Another trait we looked at was the social dominance orientation (SDO). It refers to a propensity to believe in (natural) social hierarchies and to prefer the preservation of current hierarchies or the upwards advancement of oneself and one's group to the detriment of others. One can also think of SDO as the opposite of egalitarianism, since those who display a strong social dominance orientation believe that it is normal and right that some should be on top in society and others at the bottom (Pratto et al., 1994). The journalists had moderate levels of SDO, but we were not able to compare these with SDO levels in the national survey, since these items were not included in that survey, due to limited space in the questionnaire.[33]

When we tested for a few associations between these traits and their outlook on comments, we found that higher levels of authoritarianism may be associated with a stronger belief that journalists have no reason to respond to online comments.[34] Also, higher levels of the social dominance orientation may be associated with a tendency to believe that it is not worth trying to engage with the readers on the website.[35] Other beliefs and attitudes in the newsroom and their impact on the project will be addressed in more detail in future publications.

[33] For our newsroom survey, we used two fairly «soft» items, which did not have the strongest possible anti-egalitarian interpretations, since we were looking to capture whether people think social hierarchies are normal and whether they perceive life as a competition between different people. Among the journalists, 44% tended to endorse the view that «It is natural for some to have better chances in life than others» rather than the view that «Everyone should have equal chances in life.» The second item asked people whether they think that «To make it in life, people have to help each other» or that «To make it in life, people sometimes have to be individualistic and use others in their own interest.» 32% of the newsroom tended towards the latter perspective.

[34] This association only appeared when looking at one of our measures of authoritarianism in this survey, namely a propensity to disagree that it is important for children learning to think critically for themselves. No such association was found when looking at the other measure (a tendency to agree that "It is more important to encourage respect for existing rules and leaders in society than for individual liberties".)

[35] Like the previously-mentioned correlation, this one also only worked with one of the social dominance measurements, namely agreeing that to make it in life you have to sometimes take advantage of others.

MODERATOR EXPECTATIONS

"I think of it as a sort of re-education of the users."
- Moderator

"We are trying to turn the comment section in a safe area, where everyone can express their opinion in a civil way without being attacked or verbally harassed, and at the same time to study the behaviour of those who post comments."
- Moderator

The moderators spent the first month of the moderation experiment checking comments and making decisions about what can be published and what cannot without much direct feedback or instruction from the researcher team. It was only after this initial month that we began to have weekly meetings in which we addressed dilemmas and gradually drew up some rules and criteria on how to moderate.

After this first month of "unsupervised" moderation, we asked the team some questions about their mission, their impressions on the comment section and their hopes and fears regarding the project. When pondering what the newsroom may be trying to accomplish through the comment moderation, they first and foremost cited the ideal of removing and discouraging violent expressions, name-calling and insults, as well as racism, xenophobia and general intolerance in the comment section.

Some of them also talked about the idea of encouraging more well-argued debate and creating an environment where people feel more comfortable expressing themselves. Along the same lines, they defined their role mostly as "cleanup" agents, whose main job is to sanitise the comment section and to thereby show users what goes and what does not.

While their hopes revolved around seeing more genuine debate and more civility, their fears were also about themselves, not just the commenters. On the one hand, they were concerned that people may flock to other sports websites, something that some commenters were threatening to do already. On the other hand, they also worried about being psychologically affected by spending many hours reading hateful comments. Their comments would prove quite prescient, at least with regard to the latter aspect.

MODERATION DEBATES

In order to help show why certain changes in the moderators' views towards comments, the idea of comment moderation and even attitudes towards intolerance and certain groups may have happened, we cannot avoid addressing some moderation challenges and debates that occurred as part of this collaboration. In addition to explaining how the moderators thought and engaged with the task, it may also highlight connections to wider societal attitudes and challenges like defining and combating racism.

The process of drawing up the rules and procedures for dealing with the comments turned out to be more lengthy, exciting, and difficult than initially anticipated. After the initial month of "unsupervised" moderation, during which the moderator team checked and moderated all comments without having had any in-depth discussion with the researchers, the MRC researchers and the team of moderators began to meet to discuss challenging comments and make joint decisions about what can and cannot be allowed on the website. We continued to do so for the better part of the experimental moderation period.

The discussions started from a weekly field report submitted by the moderator team, containing moderation dilemmas and observations on the comment section (how the vocabulary of the commenter is evolving, how people react to the moderation, how users relate to each other, what topics generate nasty comments, what kind of intolerant or uncivil speech is more pernicious, etc). The researchers then generated a weekly "handbook," to be discussed with the moderators, the coordinator and, sometimes, other journalists in attendance. The handout contained advice on how to recognise various types of intolerance or incivility, feedback on comment moderation and the researchers' position on how to deal with specific comments. Decisions were usually the result of reaching a consensus, but heated debates sometimes occurred, and there were situations were consensus was not reached.

Faced with some new concepts, new imperatives (like combating intolerance in the comment section), and often difficult, "borderline" comments, the moderator team engaged very intensely with the issues. They constantly questioned if the moderation was too harsh or too soft or how, exactly, key concepts like racism, religious intolerance or freedom of opinion should be applied when faced with the challenges the commenters threw at them.

Everyone now has a greater appreciation for how complicated judging speech can be and how apparently simple concepts become much more complicated when you have to apply them to real-life moderation cases, or your own language choices or reactions. Is saying something disparaging or mocking about Christianity less bad than saying the same thing about Islam, if the country is majority Christian? If many Romanians consider "gypsy" to be an acceptable term for the Roma, should we allow it on the website? Should Romanians worry about offending Asians with problematic language, even though our country has never had a sizeable Asian minority or perpetrated systemic discrimination or abuse of Asians? All of these questions, and others, were discussed, some of them intentionally and repeatedly; others came up almost accidentally. They reflect the difficulty of striking the appropriate balance between protecting the commenters' freedom of speech and "curating" a comment section more in accordance with democratic values and respect for all people. Below we detail a couple of moderation challenges and stories from our debates.

"ŢIGAN" - REGULAR WORD OR SLUR?

Before the project, the gsp.ro comment management software already scanned all comments, flagging words based on a "bad words" dictionary. It contained 162 variations on "ţigan" (Gypsy) many being the result of commenters attempting to bypass the system by using symbols or numbers.

In the early phase of moderation in the Less Hate project, the team debated whether to allow the word "ţigan" when no other part of the comment would require moderation, and the comment is not using the word with an obvious intent to offend. We resolved to pursue a blanket moderation of all instances of "ţigan," regardless of context, pending a final decision based on further research and the observation of user behavior.

The attitudes towards the word "ţigan" among the moderators reflected the ambivalent relationship of Romanians towards the term. While readily acknowledging that most comments about the Roma (whether using ţigan or other terms) were intolerant, and while highly motivated to moderate racism in the comment section and, thus, to intervene whenever ţigan occurred in a derogatory or prejudiced context, some moderators argued that suppressing the term itself, regardless of context, was excessive or even counter-productive.

The main arguments and concerns this subset of the team put forth were the following:

- Ţigan could occasionally occur in a positive context. So the question they put forth was, what if someone says something nice but they use the word "ţigan"? For instance in a comment that says "Banel, you're a cool gypsy". (The question was phrased in the very first meeting by a journalist in attendance but it resurfaced over time as a concern shared by some moderators).
- The word can occur in a neutral context or in a quote, for instance if a Romanian referenced chants used by foreigners at football games. So the question was, what do we do if someone just says, "the Italians yelled ‚ţigani, ţigani'"?
- Some argued that ţigan itself is not comparable to racial or ethnic slurs directed at other groups because by itself it is not always meant to insult or humiliate. Therefore ţigan should only be moderated when meant in a derogatory or intolerant way, rather than in all contexts.
- One person argued that, in his opinion, "Roma" is often used ironically by the majority population, so in itself is not better than ţigan but, rather, context should dictate the interpretation of the word. The same moderator also brought up the fact that there are many Roma who prefer to call themselves "ţigan". He also argued that suppression of the use of ţigan does nothing to discourage intolerance of the Roma, especially considering that it is often used generically to designate that group, by Romanians who mean nothing in particular when using that word. "Obviously many use 'ţigani' to denote all that they consider subhuman, but that does not mean we should change the name, rather we should penalise that perception," he felt.

To sum up, not everyone accepted that a) "ţigan," in regular speech if not in the comment section, is mostly used in a negative context, where the person using it means to insult by the very use of the term or says other insulting things while using the term; and b) the term itself is "sullied" by its overwhelming association with negative stereotypes and prejudice about the Roma, and as such, can no longer be sanctioned as being a neutral designation for the group. This reflects overall discourse about the term among Romanians, in the sense that, while only a very few campaign for the exclusive use of "ţigan" (some MP proposing last year that the official term be designated "ţigan" rather than Roma), many continue to reject what they see as excessive political correctness and claim the use of "ţigan" is appropriate, neutral language rather than a term that intrinsically transmits a negative message.

In the absence of survey data about how Roma themselves perceive the term, it is naturally difficult to assert that "Roma dislike the term," except by appeal to authority, explaining that Roma civil society organizations tend to reject the term. The researchers argued that promoting the use of "Roma" by moderating occurrences of ţigan was a way of "showing respect" to a minority, by allowing it to determine what to call itself and using the more polite term if speaking as a member of the white majority. Appealing to outside norms, the MRC team reiterated that while African-Americans may call themselves whatever they like, the majority population is not, according to norms of civility and tolerance, allowed to use what are now considered derogatory terms. These arguments seemed to work somewhat, but with those who are unconvinced of the offensiveness of "ţigan" in the eyes of the Roma or according to democratic norms, they can only go so far. In other words, someone who is very motivated to suppress discriminatory or offensive speech must believe that a certain term is offensive or discriminatory in order to actually be motivated to renounce it.

In the end, we decided to continue to moderate "ţigan" in all circumstances, replacing it with "***". As the Council of Europe recognised, the term "is felt to be pejorative and insulting by most of the people concerned (although it is true that it may depend significantly on the context in which it is used)" (Council of Europe, 2012). Furthermore, based on our observations, in the comment section "ţigan" is almost always used as a slur, either to put down the Roma or to insult others by labelling them as such. This usage indicates that commenters also mostly perceive "ţigan" as a negative label, rather than a neutral one.

Interestingly, in the evolving conversation about "ţigan," some of the more skeptical moderators have proven open to emerging evidence about the negative associations around the term. Thus, a moderator recounted that he has been asking Roma acquaintances how they perceive the term. One of these people told him that he did perceive the term as offensive and that what bothered him, as a Roma, the most, was that whenever you searched the term online you found only definitions like "dirty person," which persuaded him that the term itself should be out of use. The moderator said that he had found this persuasive. He also found it persuasive when evidence was brought up that public figures do distinguish between ”ţigan“ and Roma and use that distinction to make points about members of the group. The example used was a television personality explaining how Roma children often have no choice ”but to become gypsies“ due to lack of opportunity, meaning, implicitly, that Roma = normal citizen but ţigan = delinquent person, beggar, etc.[36]

WHEN LABELS ARE HARMLESS AND WHEN THEY ARE OFFENSIVE

One of the most heated moderation debates began from a hypothetical moderator question: ”What should we do if a commenter uses the term ‘slanted eyes‘ as shorthand for ‘Asians‘?“ While the research team saw it as a racial slur and vigorously argued that this term cannot be allowed under any circumstances, not all moderators immediately agreed with this.

Among other arguments the researchers brought up were the following:

- The term represents a caricaturing of stereotypical traits attributed to a set of ethnicities, similar to depicting all Jewish persons with curly hair and a crooked nose or black persons with big facial features. Such caricaturing or shorthand for how a racial or ethnic group supposedly looks, we argued, is no longer acceptable in the West because it is recognised to be inaccurate and offensive, as it inherently positions oneself as ”normal“ and the target as somehow deviant or even clownish.
- Slanted eyes, much like the crooked nose or big lips, are features that have been historically used to either make fun of certain racial groups (blackface and old-school caricatures or ads with African Americans, for instance) or to depict them as threatening (World War Two imagery of the Japanese or Jewish persons), both being forms that frequently dehumanise. We argued that even if there is no history of systemic discrimination or abuse of Asians in Romania, we must be mindful of the greater historical context which can influence how discourse or imagery are perceived.
- We also argued that Western outlets would never consider this term acceptable and likened ”slanted eyes“ to the n-word, echoing the arguments that people at the New York Times evoked as contributing to the decision to stop printing the n-word in the 1970s: ”Paul did not think they were excusable at all. The point he made to me, patiently and persuasively, was essentially the one that came to be embodied in the stylebook entry on slurs: that even when supposedly innocuous or newsworthy, appearances of the word erode a barrier against its promiscuous use“ (Dunlap, 2015).

Citing situations where this term had been used in the press without apparent intent to offend or put down, some moderators essentially argued that it is not a slur and could be seen as a harmless label:

- Some argued is a fact that many Asians have ”slanted eyes“ and referencing a prominent or prevalent feature should not be considered offensive, insofar as calling people ”black“ is not offensive (most did, however, acknowledge that calling people ”yellow“ is different). Along the same lines, a couple of people argued that white, blue-eyed, blonde-haired people would likely not mind being referred to based on those features, asking why it should be different for other races.
- When confronted with the history of racial caricaturing, some said that while it was natural for Americans to be mindful of it, it was not so obvious why in a small country like Romania, which has few Asians and is far less developed and important on the global stage than China or Japan, people should conduct themselves as if there was some historical ”blame“ that they ought to account for or some unequal power-relation that would require taking special precautions not to offend Asians.
- They also pointed out that European publications tend to be more lax about offending minorities than American publications.

We eventually resolved that if ”slanted eyes“ or similar labels should ever occur in the comment section, they would be moderated, even though not everyone was initially persuaded of this necessity on normative grounds.

[36] Pundit Moise Guran: ”The problem of those born in Roma families is that they have no chances of becoming anything other than Gypsies.” Guran, M. (2016). Retrieved from http://www.biziday.ro/2016/02/10/sa-furi-de-la-saraci-nu-e-haiducie-e-tiganie-aferim-dna/.

THE MODERATION DEBATES - A MIRROR OF WIDER SOCIETAL STRUGGLES

The discussion on ″slanted eyes“ detailed in the previous section echoed other dilemmas on how to treat different statements or labels, depending on who they referred to. We had to also consider whether comments mocking Christianity are the same as comments mocking Islam, or whether the latter should be treated more harshly. And though we had decided not to suppress inter-team competition, playfulness and even mockery when it came to sports teams and their supporters, we repeatedly had to revisit the question of how much the commenters should be able to say about their adversaries and whether teams and their supporters as a group should be treated as if they were a racial group or another social group.

Although this may not be immediately obvious, in a way these discussions reflected wider societal debates about the best way to define and tackle intolerance. The moderation debates frequently ran along two different lines that could occasionally clash, highlighting what could be inherent tensions between two approaches to combating intolerance and discrimination: a) focusing on groups and targets and taking into consideration social and historical context, calling attention to systemic and pervasive injustices or discrimination and b) focusing on general principles and the promotion of their application to any and every one.

As our survey among the moderators had confirmed, the team rated quite high on motivation to control prejudice, and throughout the project they were motivated to promote tolerance, although they themselves were not totally free of preconceptions about various groups. The survey and our discussions highlighted something that public opinion studies and experiments on attitudes towards various social groups have suggested: anti-prejudice norms do not apply to the same extent to all groups. Moreover, everyone has certain sensibilities or, conversely, blind spots, though they are not always immediately obvious based on their group identity. The two female moderators tended to bristle at sexist comments more than the three male moderators. One moderator appeared particularly perceptive about comments that disparaged people of certain socio-economic situations, although this person claimed this was one area where they were more lenient than other moderators. Two moderators were particularly religious, but one of them argued for a more lenient treatment of comments mocking religion, while the other thought any mockery of pious people could fuel religious intolerance and normalise disparagement of religious persons.

Our experience with the moderators suggests that people think about tolerance and intolerance intuitively in terms of out-groups and in-groups. At the outset, in discussing particular comments that posed challenges in moderation, we naturally focused on the groups those referred to (women, the Roma, working class people or poor people, etc). Especially when confronted with calls for ″free speech,“ both the moderators and the researchers naturally tended to grasp for arguments in favour of moderation that related to the social context in which we placed ourselves. Indeed, discussing why it's important to ″not disparage a broad group of people“ or ″not allow any slurs“ can be quite difficult without referring to specific situations and possible effects. And it can be hard to explain why a certain statement is prejudiced (as opposed to true) without appealing to the wider social context for justification. In discussions with the moderators, we found that before they began to use phrases like ″it's not okay to generalise“ or ″dehumanising insults of any kind are unacceptable“ all the time, they more readily said things like ″Jews have been exterminated and so we have to be careful“ or ″the Roma are disadvantaged so we have to protect them.“ This worked quite well when it came to vulnerable minorities (like the Roma) or those groups that everyone agreed tend to be the target of hostility and mistrust (Hungarians or Jewish people). Where we ran into the biggest trouble was groups who are not perceived by everyone to be in need of particularly sensitive treatment.

What does this mean for how moderation teams tackle discourse in the comment section, and for how we should think about intolerance in general? On the one hand, when it comes to historically marginalised or disadvantaged groups, the project of promoting tolerance and egalitarian norms takes place in a context in which societal structures themselves can be stacked against the same groups that face prejudice from regular citizens. This is why many define racism not as simple racial prejudice, but as the combination between beliefs about the characteristics of a group, the belief that those characteristics make the group inferior to one's own and, crucially, ″the social power that enables these to translate into disparate outcomes that disadvantage other groups or offer unique advantages to one's own at the expense of others“ (Dovidio, Gaertner, Kawakami, 2010). From this point of view, it can be said that the prejudice the Roma face and the prejudice that Romanians may hold about, say, Americans, is not exactly the same. Thus, a comment containing a negative stereotype about the Roma (″they're all lazy and thieves“) does not have the same possible social impact as a nasty comment about Americans (″they're all fat and dumb“).

On the other hand, successfully promoting tolerance by pointing to inequalities and the wider social context in which intolerant or offensive discourse occurs requires that the interlocutor actually accept the existence of such inequalities and view the context and the causes in a certain way (i.e. not by believing the relative powerlessness of a certain group is due to their own inferiority or that historic injustice should have no bearing on how we decide to treat or talk about certain people in the present). And, as we saw, arguing that tolerant speech is called for because a certain group is already vulnerable can get you into trouble when you encounter groups that are not necessarily vulnerable, but also deserve to be treated with respect.

In the end, we tried to ”square the circle“ by finding definitions and criteria for key concepts like intolerance that would help make moderation decisions easier regardless of the group. These are detailed in our report on the moderation. However, we did take tackle comments exhibiting intolerance based on categories that are usually discriminated against more aggressively than comments that discussed categories such as football team membership. These challenges underscore the difficulty of promoting tolerance and egalitarian norms in a society where rules about the acceptability of various forms of speech are loose and almost anything goes, and they show how difficult it can be to reach the ideal point where equal respect is afforded to all groups.

MODERATION PROCESS RIPPLES IN THE NEWSROOM

The comment moderation and discussions that accompanied it also made some ripples in the newsroom at large. The moderators occasionally reported on reactions from the newsroom, especially situations where their colleagues felt that the commenting rules were not sufficiently harsh, and they pointed out insults or forms of aggression that were not moderated, such as disparaging comments about the articles or certain types of name-calling.

What is more interesting, however, is that as the meetings went on and the researchers and moderators continued to debate and learn together, the latter admitted to becoming increasingly aware of framing effects, when choosing certain subjects, wordings or pictures. Due to their exposure to the comment section, they also grew more aware of what tended to generate the most hate or incivility. Because of this, they reported occasional disagreements and even heated debates in the newsroom when the moderators drew the alarm about problematic coverage. Some of these instances are detailed below.

The ”magic carpet“ and other problematic framing of religious difference

The treatment of religion generated several newsroom discussions, according to the moderators. Attitudes and approaches differed significantly among the staff, and the same situation could be covered quite differently in a blog versus in a print article or on the front page of the newspaper.

The researchers generally avoided being the initiators of discussions on newspaper or online articles in the moderation meetings, as whatever the authors chose to write was outside of the scope of the moderation experiment. Still, as time went by the moderators came to bring up these topics themselves and ask for feedback or report on ”battles“ won or lost. On several occasions they complained about editorials that trafficked in implicit prejudice. One such example was an article following the Paris terrorist attacks where the editorialist, writing about French footballer Karim Benzema, who is of Algerian descent, said that if not his friends, then ”friends of his friends“ could be terrorists. The team also occasionally mentioned newspaper caricatures they found offensive, such as a drawing of half-naked sirens wearing headscarves depicted trying to entice Victor Piturca, then manager of the Romanian national football team, to leave the team and go manage a Middle Eastern team instead.

Several moderators reported arguing against a front page of the print version of Gazeta Sporturilor which used a photo of a praying Sulley Muniru, a Steaua Bucharest football player who is Muslim. Steaua had just defeated Viitorul 3 to 1, and Muniru was photographed praying on a small rug in a corridor at the stadium. The front page featured the words ”magic carpet,“ and this word choice was the major, if not the only, point of contention. ”Despite some objection in the newsroom, this title was kept for the front page of the newspaper,“ a moderator wrote.

> ”[A] question still lingers: is it ok to make a front page subject out of the fact that a man is praying? Could anything scream ‘look at the freak’ louder than that?“
>
> \- Moderator

Perhaps ironically, the editor-in-chief of the newspaper referenced the same situation on his blog (Tolo.ro), with arguably different purposes and effects. In the wake of the Brussels terrorist attacks, as public discourse with regard to Muslims and security measures got increasingly heated in Romania, Cătălin Tolontan opened an article pleading for mutual tolerance and respect for civil liberties with the story of Muniru praying at the stadium. He noted that this private moment occurred in a relatively public setting but fans were respectful and walked on without making a big fuss. Going on to cite other examples of Romanians being civil despite encountering ”difference,“ he made a point that people can live together and share common joys without being exactly the same in creed or choices.

The case of the holy paint roller

Other editorial choices also generated discussions. One front-page cover that referenced the religious beliefs of Anghel Iordănescu, the former manager of the national football team, split the moderator team in unexpected ways. Iordănescu, a Romanian Christian Orthodox, is known to be very religious and is occasionally mocked for his very public displays of devoutness or superstition with regard to football game outcomes.[37] The cover in question was based on a viral snapshot of the Church's foremost leader, the so-called Beatific Patriarch Daniel (in the Romanian, ″the very happy“). While doing a blessing of the offices of Trinitas, the Church's media company, he was photographed in a series of comical positions, holding a paint roller that had been dunked in holy water and using the very long handle to reach every corner of the rooms. The Gazeta cover in question spoofed these images by depicting football manager Iordănescu in the Patriarch's luxurious garb and using a similar long-handled paint roller to ″bless“ the pots for the Euro 2016 group draw. The cover read ″The very lucky“ in large print, accompanied by ″Iordănescu prays for an easy group in the Euro 2016“.

The moderators reported that the cover had been hotly debated in the newsroom. Some argued that the cover mocked Iordănescu's devoutness and normalised the disparagement of religion in the public sphere. They argued that in the comment section and elsewhere, not just Muslims, but even people of Christian faith were increasingly labelled as backward, and that the cover normalised this attitude. Some also said that the Patriarch functioned as a religious symbol, and thus it was offensive to mock him, albeit indirectly. They also questioned whether it would have been appropriate to depict any other religion in this manner and whether there wasn't, in fact, a double standard, with more liberty afforded to mockery of Christianity than of other religions. The other side disputed that the cover mocked Christianity or religious people in general. They countered that the Church's leaders were fair game because they are closer to political appointees than religious symbols, with all the accompanying public scrutiny that this entails. Making fun of them could not be likened to mocking the Church or faith itself, they argued. Some also said that the cover, rather than mocking Iordănescu's faith, made light of his superstition, in the form of over-reliance on ritual, icons and other objects meant to secure good fortune. Interestingly, the moderators and other members of the newsroom did not split along religious lines on this matter: two of the most openly religious persons on the moderator team were in opposite camps.

Sexism blind spots

Sexist comments were very frequently discussed in the moderation meetings. By contrast, the moderators did not report any coverage of women in the print or online version of the newspaper to be problematic from their point of view. However, some moderators volunteered that they are occasionally bothered by their colleagues' failure to recognise and take seriously instances of sexism in the sports industry, even when they took pains to explain the issue.

> ″I think the idea of discrimination is sometimes taken as a joke. Sexism cases in particular. The most revelatory example that comes to mind is the reaction to a news piece about Figo. […] The reaction in the newsroom was, 'You're exaggerating, a man can't even tell you you're beautiful anymore because you immediately accuse us of sexism.'“
>
> \- Moderator

[37] Though over 90 % of Romanians are Christian Orthodox, the Church and its leaders can be quite divisive subjects, with even some religious folk criticizing the upper ranks for their opulence and involvement in politics and public affairs.

> ”When I wrote an article about the former footballer Figo, who told a female reporter at a press conference, ‘Normally I don’t answer questions from the Catalonian press, but because you’re beautiful, I will,’ and I labeled this remark as sexist, I had a discussion with some of my colleagues, who said that ‘Now you can’t even compliment a woman because it’s sexism/misogyny, harassment. If you tell her she’s ugly it’s not ok, if you tell her she’s beautiful it’s not ok. Nothing is ok anymore.’ We discussed the matter a little, and I explained the context in which the statement took place and why the remark was inappropriate, but I don’t think they were convinced, in the end they were left with the impression that we have something against men that compliment women.“
>
> - Moderator

Racial insensitivity and reporting on discriminatory incidents

Several moderators also came to believe that some of their colleagues were not sufficiently sensitive about racial or ethnic issues. For instance, at one point the newspaper used the words ”yellow invasion“ to describe the success of Asian sports teams. After all of the moderation discussions about ”slanted eyes,“ months later, the moderators seemed to have come around to considering this kind of expression inappropriate, and they somewhat bemusedly mentioned that a print reporter had written about ”stadium stands full of slanted eyes“ in an article.

The moderators also reported that, in previous years, the newsroom had wrestled with how to react to racist incidents in the football world. Some of them still bristled at the memory of a front page cover from years before, which read ”Suspend us, we love crows.“ The cover was in response to the ”Suspend us, we hate crows“ chants of Steaua supporters, directed at the Rapid team, one that is associated with the Roma minority.

> ”Many years ago, the front page of the newspaper, after a big Rapid game, used the title ‘Suspend us, we love crows!’ in response to a racist Steaua chant, I think, ‘Suspend us, we hate crows!’
>
> I understand there was no racist intent, but just paraphrasing a racist slogan is totally wrong. In the discussions about that front page with the people in the newsroom I was always told I didn’t see ‘the big picture’ and that I am subjective because I’m a Rapid fan. I still think it has nothing to do with that, and this opinion has been cemented since doing moderation. I don’t think we’ll be able to change anything by using a negative example, regardless of our good intentions.“
>
> - Moderator

According to the journalists, the newspaper currently has a policy of reporting when a racist or discriminatory incident took place in the sports world, all the while explicitly labelling it as such in order to avoid any appearance of endorsing such behaviour. Still, accidents or mistakes did happen. One such situation occurred after some football fans used the word ”crow“ on a giant banner in order to offend rival fans. The newspaper reported that the club had been penalised for its fans’ racist behaviour, which the online article explicitly called out as such. However, on Facebook, the article was arguably inappropriately framed, as it was promoted using just a photo of the racist banner and a caption to the effect that ”Team is facing penalisation.“ Some activists saw the Facebook post and took the journalists to task for failing to appropriately condemn racism.

Moderators influencing their colleagues

There were also a few stories from moderators who managed to change the minds of colleagues regarding the appropriateness of certain topics or angles.

”After Dinamo won a friendly with two goals from African players, a colleague asked me if it was ok to use a title like ‘Black Dogs’ (n.b. Dinamo players and fans called themselves “dogs” or “red dogs” because of the team’s insignia). We discussed this subject a little, without reaching a clear conclusion. There were arguments for and against, that the expression itself was not racist, but the fact that it was highlighting their skin color, in a context where this had no relevance, could be perceived differently. In the end he used a different title.“

- Moderator

”When the international football section wrote an article for the newspaper titled ‘Schweinsteiger, Nazi soldier! The Chinese made Nazi statuettes that look like the United player,’ we discussed this a little, about the angle, the subject being extremely delicate. We adopted a position that penalised the gesture of the Asian company and left as little room for interpretation as possible.“

- Moderator

”A colleague found a video of Liverpool fans with disabilities online, at a game, behind the goal posts, in a specially designated area, where it looked like they got up from their wheelchairs when a goal happened, to celebrate. He spread it all over the newsroom, until it reached our department too. I looked at it more carefully and noticed that, in fact, those who were getting up after the goal were the attendants, as each person with a disability had to be accompanied in that part of the stadium. The general reaction was, ‘Eh, so what, put it online! You can’t see it clearly anyway, only you saw it. It looks like they’re all getting up. On the website where we found it why did they share it?’ I explained that it’s not ok to insinuate that persons with disabilities get up from their chairs when they are not actually doing that and to make ironic comments about it too (because they proposed an ironic title like ‘The Liverpool miracle‘). I also explained that if a person with disabilities is in the specially designated area of the stadium, it doesn’t necessarily mean they are paralysed and cannot move / get off their wheelchair for a brief while. I noticed that most did not accept the explanations and were left with the impression that we did not take advantage of a very good video, which was worth posting.“

- Moderator

MODERATOR VIEWS AT THE END OF SUPERVISED MODERATION

HOW THE MODERATION AFFECTED THEM

The moderation meetings were a common learning process, for journalists and researchers alike. When asked to write about their experience a year after the start of the moderation experiment, the moderators confessed to gaining a greater sensitivity to intolerant speech and awareness of the pitfalls of careless word choice or playing to the audience's prejudices.

> "I would like to say that I have a different opinion compared to April 2015. It's been a year and now I, at least, think twice or three times before I write a piece of news that could appear racist or lead to violence. I admit that before I didn't have this perception, now all five of us are more careful about this type of news, and I think we notice problems much more quickly."
>
> - Moderator

> "Personally, the moderation made me see much more clearly than before what's not ok in formal or informal speech, and I penalise the behaviour of those around me much more often than before, explaining what the problem is and why they shouldn't repeat it."
>
> - Moderator

We did not expect to see any quantifiable shifts in the moderators' attitudes as a result of the moderation exercise, and indeed, in the last moderator survey of June 2016, we did not measure any significant changes in their attachment to stereotypes. Yet we did find that the team exhibited lower levels of symbolic prejudice towards the Roma, compared to where they were at in 2015. The measurement for this had been a choice between two statements, placed on either end of a 0-10 scale. The first statement, indicating symbolic Roma-related prejudice, was that "To succeed in life you simply have to try hard. Roma people should do the same without asking for help." The second statement was "Generations of slavery and historical disadvantage have created conditions that hinder the efforts of Roma people to lift themselves up from the lower class without help." In the first moderator survey, two of them tended to agree with the first statement, two were in the middle, and one agreed with the second statement. A year later, everyone had shifted closer to the second statement, with one now totally in agreement with it.

In addition to positives like greater awareness, there were also downsides to the moderator job, something that the journalists had anticipated, when voicing concerns at the outset about coping with the negativity and hate in the comment section. In the exit survey, four out of five said that moderating comments had affected them personally or psychologically, and the team agreed that moderation work would be better combined with other activities in the newsroom.

VIEWS ON MODERATION

The moderators' views of what their main role was in the "Less Hate, More Speech" project and of what the ideal role of a comment management team should be reflected what the bulk of their activity had been. When asked both about their real role and the ideal role of such a team, they put gate-keeping, acting as a filter and making decisions about what comments should appear on the website, first. They also thought acting as a facilitator, encouraging interactivity and dialogue, for example by posting comments or highlighting the best contributions was important -- an activity that also featured in the project but had far less prominence in the experimental moderation process. Two alternative roles, being an expert, someone who answers commenters' questions or helps them navigate the website or learn how to use it and how to behave, and being a messenger, someone who relays information to the newsroom or their bosses, were considered much less important by the end.

By the end of the supervised moderation period, a time when the moderators were about to set off on their own and work without the feedback and advice of the researchers, they seemed quite confident in their ability to separate comments that need moderation from those that do not, be it by identifying the most aggressive comments or those that exhibit hate speech and prejudice.[38] All of them thought the moderation could be applied systematically.

By and large, the moderators had a positive appraisal of the moderation process and its effects on the comment section, although their views on more specific, nitty-gritty details, were mixed. On the positive side, while before the moderation experiment, two of those who would become moderator-journalists believed moderation went against newsroom principles and free speech, only one held on to this opinion to the end. All of the moderators agreed that rather than being eliminated altogether or being allowed with no restrictions, it is better for comments to be moderated.

One of the moderators thought the moderation had been too harsh and another thought it was too lenient (the rest appeared to be satisfied with the current rules). The moderators also continued to believe in the necessity of a banning option for certain users, something that had specifically been avoided during the supervised moderation phase. The fact that banning had not been enabled was cited as a major frustration.

The moderators did not, however, seem supportive of the idea of forcing users to give up their anonymity. Only three out of five believed this would make conversations better, and all of them agreed that when people comment anonymously they can communicate ideas that they might otherwise be afraid to express and that forcing commenters to use their real names may expose them to certain risks. All things considered, only one of them believed anonymity should be eliminated, another was undecided, and the other two disagreed.

As far as the effects of moderation go, in July 2016, all said they noted a reduction in the number of vulgar, aggressive or intolerant comments. But by March 2017, they had become split on this: three leaned towards a no, one was completely sure it had, and the last one leaned towards yes.[39] They did not notice any reduction in the overall number of comments or commenters, although the number of comments did in fact suffer a small decrease during the project.[40]

Four out of five moderators continued to believe that the comments are inevitably flooded by racist, violent, brutal and intolerant speech. While in the moderator "exit" survey, applied to them in June 2016, they agreed that the benefits of moderation outweighed its disadvantages, when they answered the same question in March 2017, two of them disagreed.

[38] All of them agreed that it was possible to separate comments that are too aggressive from those that do not need moderation, that you can single out the hate speech and prejudice and separate it from the comments that can be published, and that intolerant comments can be separated from the rest of the comments.

[39] What we know is that the share of moderated comments out of all comments posted decreased over the experimental period. The number of total comments experienced a small decrease, whle the number of commenters (and pageviews) increased.

[40] In July 2016, none of them thought either the comment or commenter numbers had gone down, but by March 2017 one had come to believe that there was a drop in the number of commenters.

GOING TO THE NEXT LEVEL

Both from their testimonies and from their answers to various surveys, it appears that by the end of the project, the moderators also saw a need for the newsroom to think not just about moderation, but about how the comments are shaped by the news environment and how the users can be engaged and influenced by the journalists. In the last moderator survey, the entire team agreed that the tone and content of articles influence the tone and content of the comments. Likewise, the majority of them also believed that it would be better for both readers and the newsroom if the authors actively participated in the comment section. All thought that readers are more civil if a journalist is involved in the discussion. They also thought it was a good idea to highlight good reader contributions (a feature that had been enabled as part of the project) or to adopt a policy for how to communicate with the commenters (for instance to thank them when they report errors or to answer to good comments).

As to whether their colleagues would be on board with such activities, the moderators were not sure. When asked whether they thought the rest of the newsroom had a positive view of the moderation, only two of them tended to agree, two were undecided and one disagreed. In their own written impressions, some showed doubt as to whether most of their peers had ”gotten“ what ”Less Hate, More Speech“ was all about.

NEWSROOM VIEWS AT THE END OF SUPERVISED MODERATION

VIEWS ON MODERATION

Contrary to the perceptions of the moderators, most of their colleagues said they thought moderating comments was a worthwhile activity. From the non-moderator journalists polled at the end of the project, eight out of ten agreed that the benefits of moderation outweighed its disadvantages.[41] Three quarters believed it had let to a decrease in the number of vulgar, aggressive or intolerant comments.

Most of the journalists who did not act as moderators thought the rules were neither excessively harsh nor too lenient. There was greater support for a more aggressive approach to moderation than there was for relaxing of the rules: 11% of the journalists thought too much was moderated, while 39% thought too little was moderated. Although 63% of the journalists thought the commenters had a negative view of the moderation, most of them had not noticed a drop in the number of comments or commenters as a result of the experiment.[42]

Concerns that this process was against free speech or simply impossible to implement systematically had reduced, two years after the first survey. Whereas 22% of the respondents to the first survey agreed moderation was an infringement of freedom of speech, only 6% of the respondents to the last survey agreed. The ranks of those who believed it was not possible to totally separate between hate speech or manifestations of prejudice and other aggressive comments had also shrunk: support for this view went from 58% in 2015 to 25% in 2017. When asked more pointedly whether they believed the moderation could not be applied systematically due to moderator subjectivity, only 19% thought so, while 69% disagreed.

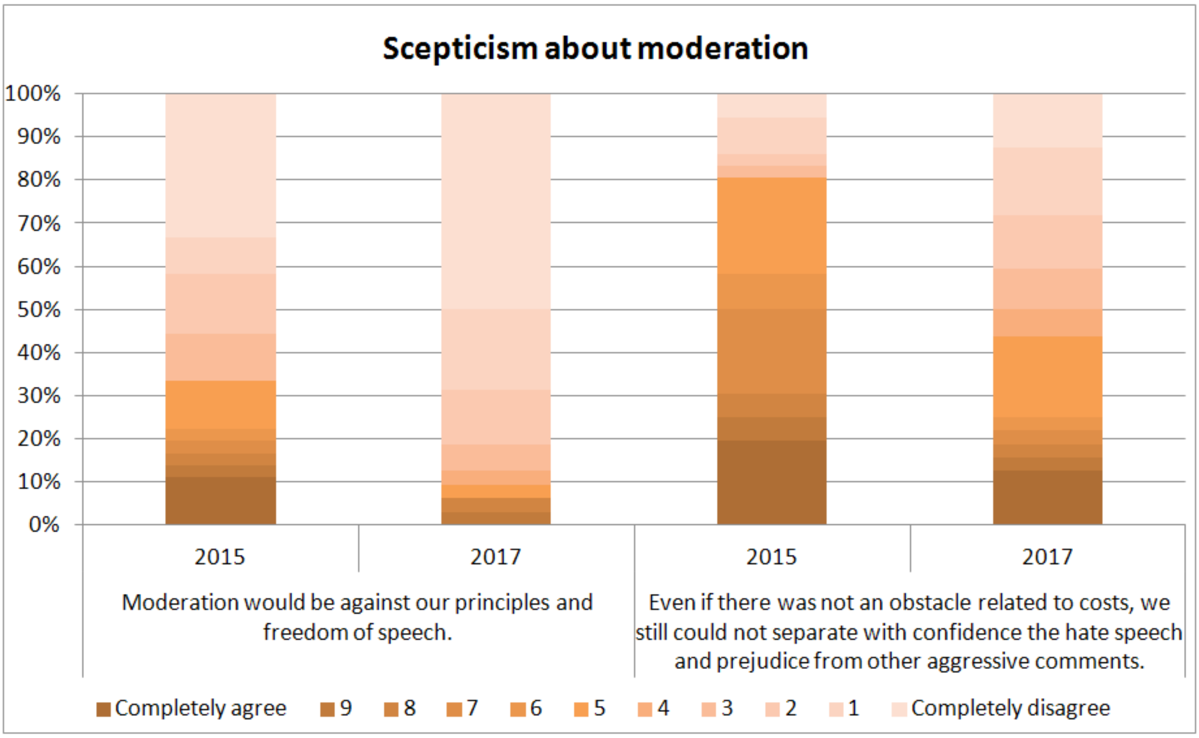

Fig. 6. Newsroom views in moderation from 2015 and 2017 (all journalist respondents, including moderators). First newsroom survey (December 2014-January 2015) and third newsroom survey (February-March 2017).

[41] This was the third and last newsroom survey, taking place in February-March 2017. It had an 85% response rate, with 34 respondents out of 40 contacted. Some of the people who took the first or second survey did not answer this one, and there were some new respondents, too (who were not in the newsroom in 2015).

[42] There was, in fact, a small drop in the number of comments during the moderation experiment, but the number of commenters and pageviews actually increased over the same period.

OUTLOOK ON COMMENTS

Still, it cannot be said that the newsroom started reading the comments more readily or appreciating them more for what they have to offer. The number of those reporting that they read the comments for most or all of their articles actually went down, to a share of about 67% of the respondents, a small trend that is confirmed even if we look only at the people who answered both the first survey and the last one and ignore the others.[43] The share of journalists who said they read the posts on other GSP articles every day or multiple times a day stayed around 46%.[44] The number of people who stayed away from the comments also remained the same.

Over time, newsroom opposition to moderation on free speech grounds, which admittedly started out at a relatively low level, reduced. In 2017, only 6.3% of respondents tended to agree with this view, compared to 19.4% in 2015. Of the people who answered this question both in 2015 and in 2017, 56% reduced their agreement that moderation is against free speech (11% agreed more). The mean opinion change that they experienced was -2.07 on an 11-point scale and statistically significant.[45]

While in 2015 the men were more opposed to moderation on free speech grounds than the women, by 2017 this difference had disappeared, and in fact it is the men who experienced a significant shift in opinion between the two points, whereas the views of the (admittedly much fewer) women in the newsroom barely moved at all during this period.[46, 47]

Skepticism about the ability to separate hate speech and intolerant discourse from other aggressive content also reduced significantly over time. While in the first newsroom survey, 58.3% believed that one could not properly separate these things, in the last newsroom survey, only 25% shared this view. Of those who took both surveys, 55.6% moved towards less skepticism (22.2% actually became more skeptical), and mean agreement with this statement reduced by 2.1 points on an 11-point scale.[48]

Overall pessimism about incivility did not actually reduce following the moderation experiment. Most of the journalists held on to their belief that the comment section is inevitably flooded by violent, racist or intolerant language (62%).

[43] There are 28 journalists who answered this question both times around, and the number of those who said they read the comments for the majority of their articles, for all of their articles, or for all of their articles multiple times a day went from 21 to 18.

[44] The number of people who answered like this, among those who responded to both surveys, also stayed the same.

[45] Paired samples t-test, t(26) = -3.102, p = .005.

[46] Men moved towards less opposition to moderation on free speech grounds between wave 1 and wave 3 (mean difference -2.47), whereas women moved (just) slightly towards more opposition (mean difference = 0.25). t(25)= - 2.434, p=0.02

[47] On the other hand, in 2017, a surprising age pattern emerged, in the sense that the members of the newsroom below 33 years old who filled out the survey tended to agree more than those 33 and above that moderation is against freedom of speech (although agreement with the statement was low among both groups). This appears to be a result of a change in composition in the respondent group, as no such difference was seen in the first wave of the newsroom survey, nor did the younger members change their views more over time than the older members did. Nevertheless, since the difference between the two age groups was barely statistically significant, it is too early to draw any conclusions. Under-33 mean = 1.86, on a scale from 0 to 10 where 10=completely agree, compared to 0.63 for those 33 years old and above. t(28)=1.791, p = .092

[48] paired samples t-test, t(26)= -2.874, p= .008

People also changed their mind about the importance of certain „journalistic achievements". On the one hand, there was a change in their response to the idea of having many positive comments on an article: respondents to both the first and the last survey rated this as less important in 2017 than in 2015.[49] Much - but not all - of the change in this appears to be a result of the change in the views of the younger members of the newsroom. In the first newsroom survey, the younger respondents (below the age of 33) saw having an article with positive comments as a more important achievement than their older colleagues. Yet between 2015 and 2017, the younger journalists who answered both surveys appeared to experience a much larger and statistically significant opinion change, such that in 2017, they awarded this less importance than they had before.[50] The below-33 and 33-and-above groups were not significantly different in opinions on this matter in 2017.

Having many clicks and likes appeared to become more important to the journalists in time.[51] Furthermore, having an article make the front page also was rated as more important in the last survey compared to the first one, while the importance rating of receiving professional awards did not change.[52]

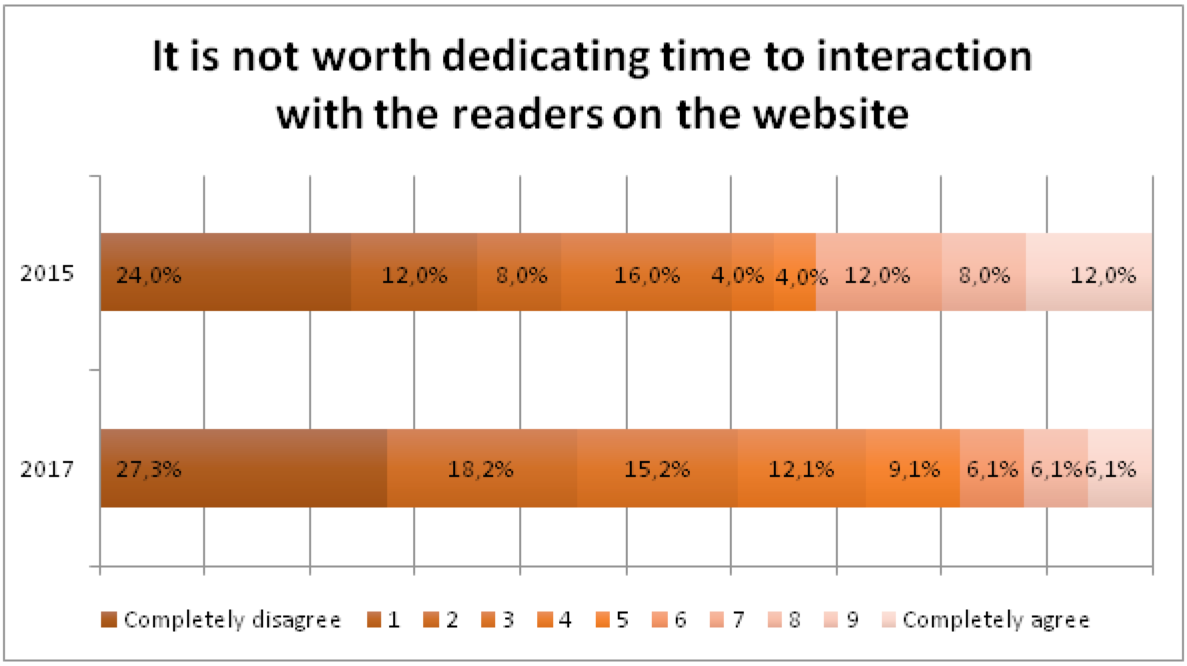

Fig. 7. Newsroom views on engaging with the online readership in 2015 and 2017 (all journalist respondents, including moderators). First newsroom survey (December 2014-January 2015) and third newsroom survey (February-March 2017).

[49] Mean opinion change was 0.81 and statistically significant --meaning comments are rated as less important in wave 3 than in wave 1. Paired samples t-test, t(26) = 1.923, p = .066

[50] Those below 33 moved towards awarding this achievement less importance between wave 1 and wave 3 (mean change = 2.2, on a scale from 1 to 5, where 1 means top priority, whereas those 33 and above moved only 0.133), t(23)=2.653, p=.015.

[51] Mean opinion change - 0.66 (meaning clicks and likes became more important to them), statistically significant. Paired samples t-test, t(26)= - 2.111, p= .045

[52] Mean opinion change -0.925, statistically significant - meaning making the front page becomes more important. Paired samples t-test t(26) = -2.222, p = .035

Although views on moderation did change among those who answered our surveys two years apart, we did not see the same evolution of views on reader-engagement. Looking at the responses to both surveys, we do see that a greater share believed interacting with the reader was worthwhile in 2017, compared to 2015. While in 2015, 32% of the newsroom thought it was not worth trying to interact with the audience on the website, by the time of the last survey, the share of the respondents who held this opinion had gone down to 18.2%. The percentage of those fully convinced interaction with the readers was not worth it halved. However, this appears driven by the fact that some people responded to one survey and not the other, as respondents to both surveys did not experience a statistically significant shift in opinion.

In terms of what the newspaper could do to further encourage good comments and discourage intolerant or aggressive discourse, it is interesting to note that the other journalists seemed pretty supportive of ideas the moderators had embraced too. Of the journalists who did not act as moderators, 71% agreed that it would be better for both the newspaper and the readers if the paper had a policy on how to approach commenters, like thanking them for signalling mistakes or responding to good comments and 82% thought it was a good idea to highlight quality comments. They, too, thought banning would be a good feature (90% thought so). By contrast, not all of them agreed with the moderators' assessment that the tone and content of articles influences the tone and content of the comments; only 44% shared this view.

CONCLUSION

What we set out to do with the moderation was find out if changing the rules of a comment section, enforcing them systematically, as well as finding new ways to communicate with the commenters, would alter their behaviour without driving them away. We were not anticipating dramatic changes, but we did find that the share of comments that required moderation slowly crept down, with only a slight accompanying decrease in the overall number of comments.

An initially unforeseen result of the GSP-MRC partnership was the richness of insight we gained from the unexpectedly lengthy direct collaboration between the journalist-moderators and the researchers. We could not have predicted both the intensity of the moderators' engagement with the work and the debates that resulted from online comments, nor their reports on how this deliberative process influenced their thinking.

The moderators came to have more confidence in the idea of moderation itself and the fact that it can be done systematically, although they did not unquestioningly embrace all aspects of the moderation rules and procedures we implemented. They continued to believe banning should have been applied, and some thought the commenting rules could have been more harsh, while in other cases they believed we went too far. While noting the positives of moderation and an overall decrease in incivility among at least a portion of the commenters, they held on to their view that comment sections are inevitably flooded by incivility and intolerance and note that some commenters remain irrepressibly offensive and aggressive. In other words, they came to see some people could change their behaviour, but they still did not believe that all or even most of them would.

> "The main idea of this project was to clean up vulgar, offensive comments, violence and racism and to encourage messages that really say something. Less hate, more speech.
>
> And this was successful, up to a point. Up to the endless Dinamo-Steaua wars, up to the endless accusations against Simona Halep after a defeat or the almost daily racism when fans of Rapid are brought up.
>
> But I say it was successful because people with good sense understood or started to understand the message, they understood that they can write in a civil manner too. [...] There are the same users you can count with both hands [...] who incite and all they do is offend. [...] But they have more comments and get the others involved oftentimes."
>
> \- Moderator

The surprising change they did report, however, was that all of the work of reading, thinking about and debating other people's discourse caused them to sometimes find themselves thinking differently about their own speech or writing.

As for the rest of the newsroom, since the project's experimental focus was reader behaviour rather than journalists' views, we did not really expect any shifts in their outlook on comments or engagement. Nevertheless, likely due not just to the project but also to an evolving media environment, in the newsroom we saw after two years greater support for comment moderation and confidence that it can work in practice, as well as more interest in communicating with the online readers.

The ”Less Hate, More Speech“ story is not about dramatic shifts, but about incremental change, in limited areas. It is, both on the commenter and, it seems, also on the gate-keeper side, about how, if the conditions are right, some people can change the way they think or behave in small but potentially significant ways.

> ”Until I moderated comments I didn’t realise that, unintentionally, I sometimes discriminated against certain categories. Let’s say now I think three times before I say something, as I am twice as aware that I could hurt someone with a joke. I’m much more aware of any discriminatory situation, and it bothers me when people treat the matter superficially.“
>
> \- Moderator

ENDNOTES

i For the newsroom survey, we had the following list of stereotype items:
a) Regarding the Roma: The Roma do not break the law more often than other people; When they start a job, the Roma do not work as hard as other people.
b) About Hungarians: Hungarians from Romania are more loyal/faithful to Romania than to Hungary; Hungarians from Romania are not willing to speak Romanian even if they know it; Hungarians from Romania do not want Transylvania to become part of Hungary.
c) Regarding Jewish people: Jewish people are more preoccupied with money than other people; Jewish people do not use persecutions they faced in the past to obtain (certain) benefits; Jewish people control politics and finance at the international level.
d) About Gay people: Homosexuality is abnormal; Homosexuals ask for certain privileges and are not happy to be treated the same as everybody else.
e) About Poor people: Poor people stay poor because they did not try hard enough to get out of poverty; When they have some money, poor people spend them less responsibly than others; Talented people do not stay poor, no matter the circumstances they are in.
All items were on a scale of 1 to 7, where 1 meant "totally agree". As you can see, some of the items are phrased in such a manner that they are the opposite of a stereotype. To compute the stereotype scores, we first made sure all items went in the same direction, where 1 meant agreement with the negative stereotype, and then reversed the values such that a larger number indicated greater support for a negative stereotype (7 = full embrace of the negative stereotype). We then computed an average agreement score for each group (e.g. for the Roma we calculated the mean agreement with the two stereotype items). Then we calculated the mean stereotype score regarding each vulnerable group across the entire newsroom, within the newsroom without the moderators, and among the moderators alone.

ii The national survey was implemented in December 2015-April 2016. More details about it can be found on the project website lesshate.openpolitics.ro. Most items in the newsroom survey appeared in the national survey, but in the latter, they were randomized such that half of the respondents got a stereotype in the affirmative, the other half in the negative. In the newsroom case we simply picked some items to ask in the affirmative and others in the negative, eliminating the randomization. Moreover, some items in the national survey did not appear in the newsroom survey, due to time constraints. Finally, each Jewish people-related item in the national survey was only presented to a third of the respondents. When we created the indexes, we used only the items that appeared in both surveys and that exhibited good reliability in the newsroom survey. The index scores for the newsroom and the national survey were then compared with one sample t-tests. For the Jewish-related stereotypes, we separately compared agreement with each stereotype within the newsroom and within the group of respondents to the LHMS survey.

iii The eight items we used are the following: I try to act in non-prejudiced ways towards Roma people because it is personally important for me; I don't want to appear racist, not even to myself; I feel guilty when I have negative thoughts about Roma people; When talking to Roma people, it is important for me that they think I am not prejudiced; I aim to be non-prejudiced towards Roma people due to my own convictions; I get angry with myself when I have a prejudiced thought; In today's society, it is important not to be prejudiced; It is important for me that other people think I am not prejudiced. They were all measured on a 1-7 scale, where, in the first newsroom survey, 7 meant complete agreement, and in the last survey, two years later, 7 meant total disagreement. To calculate the scores we put all items on the same scale, where 7 means the most motivation, and we computed the average agreement across all eight items for each person, and then looked at the mean, median and standard deviation across the whole newsroom.

iv The same results in terms of motivation to control prejudice came out from both the moderators' responses to the first newsroom survey and their responses to the first moderator-only questionnaire.

v As for the moderators, both in the first newsroom survey and in the first moderator-only survey, one out of the five tended to agree that you can trust most people, while the others tended to agree that it is better to be careful when dealing with others.

REFERENCES

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2013). The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication*, 19(3), 373-387.

Blinder, S., Ford, R., & Ivarsflaten, E. (2013). The better angels of our nature: How the antiprejudice norm affects policy and party preferences in Great Britain and Germany. *American Journal of Political Science*, 57(4), 841-857.

Council of Europe. (2012). *Descriptive Glossary of terms relating to Roma issues*. Retrieved from http://bit.ly/LaszYh.

Dovidio, J.F., Gaertner, S.L., Kawakami, K. (2010). Racism. In J.F. Dovidio, M. Hewstone, P. Glick & V.M. Esses (eds.), *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* (312-327). London: SAGE Publications.

Dunlap, D.W. (2015). 1970s | Reining In a Racial Slur in The Times. *The New York Times*. Retrieved from https://www.nytimes.com/times-insider/2015/07/02/1970s-reining-in-a-racial-slur-in-the-times/?_r=3.

Kinder, D. R., & Mendelberg, T. (2000). Individualism reconsidered: Principles and prejudice in contemporary American opinion. In D.O. Sears, J. Sidanius & L. Bobo (Eds.), *Racialized politics: The debate about racism in America* (44-74). Chicago: University of Chicago Press.

Masullo Chen, G., & Pain, P. (2016). *Journalists and Online Comments*. Retrieved from https://engagingnewsproject.org/wp-content/uploads/2016/08/ENP-Journalists-and-Online-Comments.pdf.

Mendelberg, T. (2001). The Race Card: Campaign Strategies, Implicit Messages, and the Norm of Equality. Princeton: Princeton University Press.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology*, 67(4), 741.

Taub, A. (2016). The rise of American authoritarianism. *Vox.com*. Retrieved from http://www.vox.com/2016/3/1/11127424/trump-authoritarianism#whatis.

This report is part of the 11 SEE/30.06.2014 - "Less Hate, More Speech: An Experimental and Comparative Study in Media and Political Elites' Ability to Nurture Civil, Tolerant, Pro-Democratic Citizens" project. It is financed under the "Research within Priority Sectors" Programme funded by the EEA Financial Mechanism 2009-2014.