Too many numbers and worse word choice: Why readers find data-driven news articles produced with automation harder to understand

Sina Thäsler-Kordonouri◉**, Neil Thurman**◉**,
Ulrike Schwertberger and Florian Stalph**◉
LMU Munich, Germany

Abstract

The production of data-driven journalism is becoming increasingly automated, impacting its composition and comprehensibility. Given the importance of data-driven reporting for democratic participation, this study investigates, firstly, how readers evaluate the composition of data-driven articles produced with and without the help of automation and, secondly, how these evaluations affect readers' perceptions of the articles' comprehensibility. In an online survey experiment, 3135 online news consumers evaluated 24 articles produced with or without automation using criteria developed in prior research. Our factor analysis reduced those criteria to five categories that matter in readers' evaluations of the articles' composition: numeric features, writing style, sentence and paragraph length, descriptive language, and word choice. Our results show, firstly, that although the perception of news stories produced with automation did not differ significantly from that of news stories produced without automation regarding sentence and paragraph length and writing style, the stories produced with automation were evaluated as significantly less comprehensible; and, secondly, that this can be explained partly by readers' perceptions of some of the other article composition categories, which were rated significantly worse in automated articles. Our findings suggest that using automation to produce data-driven news articles changes their perceived composition in ways that negatively impact comprehensibility. However, this study also suggests how such articles could be made more comprehensible. Specifically, when 'post-editing' automated articles, journalists should aim to further reduce the quantity of numbers, better explain words that readers are unlikely to understand and change inappropriate wording.

Corresponding author:
Sina Thäsler-Kordonouri, Department of Media and Communication, LMU Munich, Oettingenstr 67, Munich 80538, Germany.
Email: sina.thaesler-kordonouri@ifkw.lmu.de

Keywords

Automated journalism, data-driven journalism, comprehensibility, online survey experiment, post-editing, artificial intelligence, news perception, Audience Studies, AI journalism, news writing

Introduction

The algorithmic scaling of data-driven news production has become a trend in news organisations internationally—a development gaining momentum due to the increasing technological capabilities of generative AI systems (Beckett and Yaseen, 2023; Newman, 2024). These automation systems are transforming the production of data-driven news articles by impacting how journalists approach news production and, thus, how news texts are written: with rule-based automated journalism, the news articles produced may contain traces of the rigid templates used to create them (Diakopoulos, 2019), fewer of the key editorial features of journalistic writing (Caswell, 2019), and more numerical information (Thäsler-Kordonouri, 2024). Some journalists believe that such automated news production can reduce the attractiveness and comprehensibility of news articles for readers, and they therefore post-edit the automatically generated articles before publication (Thäsler-Kordonouri and Barling, 2023).

Because the composition of texts is key to their comprehensibility (Göpferich, 2009), the impact of automation systems on the composition of data-driven news texts should be critically examined in light of today's increasingly data-driven information environments and the importance of comprehensible data-driven reporting for citizens' democratic participation (Nguyen and Lugo-Ocando, 2016). Important elements in the composition of data-driven journalism include the presentation of numerical information (Waite, 2020), word choice, and linguistic clarity (Meyer, 1975). Therefore, in light of 1) the increasing automation of data-driven journalism and 2) the relevance of article composition to readers' understanding of numbers in the news, this study examines how readers evaluate the composition of data-driven articles produced with and without the help of automation (RQ1), and how these evaluations affect readers' perceptions of the articles' comprehensibility (RQ2).

We conducted a between-subjects online survey experiment ($N = 3135$) with a sample of UK online news consumers, in which we presented respondents with news articles ($N = 24$) produced either with or without the help of rule-based automation and asked them about their perceptions of the articles using criteria developed in a qualitative pre-study (Stalph et al., 2023).

Factor analysis reduced the criteria investigated to five composition factors: *numeric features*, *writing style*, *sentence and paragraph length*, *descriptive language*, and *word choice*. Although the perception of news stories produced with automation did not differ significantly from those produced without regarding *writing style* and *sentence and paragraph length,* the stories were evaluated as significantly less comprehensible. Mediation modelling showed this last result can be explained partly by readers' perceptions of other factors—*numeric features* and *word choice*—which were rated

significantly worse in articles produced using automation than in those that were fully manually written.

Our results suggest that the use of rule-based automation in the production of data-driven news articles changes their perceived compositional makeup in ways that negatively impact comprehensibility. However, our study also suggests ways in which such articles could be made more comprehensible. Specifically, when initially templating or subsequently post-editing such automatically generated articles, journalists should aim to further reduce the quantity of numbers, better explain words that readers are unlikely to understand and change inappropriate wording.

The composition of data-driven news stories and comprehensibility

Comprehensibility has been defined as the reader's subjective assessment of how easily understandable a text is (Isaacs and Trofimovich, 2012). Using a cross-discipline communication framework—drawing from cognitive science, educational psychology, linguistics, communication theory, and semiotics—Göpferich (2009) argues that compositional features "determine a text's comprehensibility" (33). These features include *(linguistic) simplicity*, that is "words and sentence constructions regarded as simple" (46); *arrangement/structure*, that is the "content structure" of a text (44); *correctness*, that is "[lack of] linguistic mistakes in the text itself" (42); and *perceptibility*, that is "layout and design characteristics, the fonts used and other paraverbal features" (48).

Data-driven journalism is the journalistic practice of applying "quantitative techniques" (McCombs et al., 1981: 22) to journalistic sense-making and role performance (Parasie and Dagiral, 2013). It involves investigating and presenting news based on abstract quantified information categories, that is data (Lowrey and Hou, 2021). The numbers-heavy journalistic routine in data-driven journalism generally affects how compositional features are implemented in news articles, with there being a higher prevalence of numerical and statistical information than in non-data-driven reporting. This prevalence has prompted several studies to investigate how readers process numbers and statistics in the news, often hypothesising that readers are usually not very well equipped to do so (Callison et al., 2009; Zillmann et al., 2009).

The presentation of numbers plays a key role in this processing, e.g., Nguyen et al. (2022) found that the use of rounded numbers compared to exact numbers improved readers' ability to remember these numbers and estimate their exact values. Hullman et al. (2018) showed how the inclusion of numerical analogies (e.g. using "the size of a football field" instead of 7527.2 m^2) meant news articles were perceived as more 'helpful' by readers. Indeed, several cognitive psychology studies show that statistical information, when presented engagingly, can be more persuasive than anecdotal information (Hornikx, 2007).

These studies, although focusing on various aspects of readers' cognitive processing of numbers, did not address how comprehensible readers thought the data-driven news texts were. As research has demonstrated that comprehensibility is important for audiences' appreciation of media messages (Van Enschot and Hoeken, 2015), the concept is a

relevant one to explore in times when data-driven journalism is of increasing importance for society (Nguyen and Lugo-Ocando, 2016).

In literature by practitioners and journalism educators, we can find specific recommendations about how data stories might be optimally composed to achieve comprehensibility. In his book *Precision Journalism: A Reporter's Introduction to Social Science Methods,* Philip Meyer (1975) encourages reporters to clearly communicate mathematical concepts to readers and use numbers sparingly. He provides practical advice on presenting numerical information, such as using percentages, ratios, and proportions to provide context and enhance understanding. He asserts that journalists are "in the business of writing with words, not numbers" (232).

Many of these suggestions are echoed in later practical handbooks and guides on data-driven journalism. For instance, in the *Style Handbook of the Associated Press* (Associated Press, 2022), the authors recommend that journalists should "avoid overloading a sentence or paragraph with numbers" (361) and give specific suggestions as to how many numbers should be used per paragraph in a news text. In their handbook for journalists, *Working with Numbers and Statistics*, Livingston and Voakes (2005) suggest that journalists should communicate information clearly to engage readers effectively and be sparing in their use of numbers. In their discussion of how journalistic texts should be constructed, they suggest using percentages, ratios, and proportions to provide context, as these are thought to be more comprehensible than absolute counts. Furthermore, the authors recommend limiting the numerical values in sentences to three to avoid overwhelming readers.

Professor of journalism practice Matt Waite (2020) also suggests avoiding overwhelming readers with excessive numbers. His practical tips for simplifying complex stories include using ratios instead of percentages, rounding numbers when appropriate, and comparing data to relevant benchmarks. However, in certain contexts, such as when reporting COVID-19 cases, Waite (2020) recommends using exact numbers.

To sum up, academic studies have investigated how specific ways of presenting numbers (e.g. rounding and using numeric analogies) in data-driven news articles affect how the numbers are processed cognitively. In addition, practitioners have made recommendations on how to write with numbers, although not all of these recommendations have been empirically tested with readers. The present study thus holistically investigates how article composition—including, but not limited to, the presentation of numbers—affects the perceived comprehensibility of data-driven news articles. In doing so, and as discussed below, we consider how the increasing automation of data-driven news production impacts article composition, with a potential impact on comprehensibility.

News automation and its impact on article composition

News content production is increasingly automated—meaning that algorithms are used to "convert numerical data, images, or text into written or audio-visual news items with various levels of human intervention beyond the initial programming" (Thurman et al., 2024: 3). In the production of news texts, this practice, commonly called *automated*

journalism, often relies on rule-based natural language generation (NLG) systems that use manually created story templates to transform data into the semantic structure of a readable text (Graefe and Bohlken, 2020). These story templates have to work with whatever values their data sources contain. For this reason and others, they must encompass "the potential for every eventuality of the story" (Rogers in Diakopoulos, 2019: 131). We argue, therefore, that when using template-based automation, journalists have to adapt their approach to news writing to the technological requirements of the NLG system used.

Others have also pointed out that using such NLG systems can affect the composition of data-driven news articles, for example regarding the use of "finegrained editorial structures (description, recitation of events, background, anecdote, direct and indirect quotation, etc.) that are essential components of the human craft of journalism but which remain relatively poorly documented, presenting a barrier to their computational representation" (Caswell, 2019: 1149). In sum, data-driven news articles created using automated journalism may feature fewer of these editorial structures and be written differently than those created without automation.

Mending compositional shortcomings of automated articles through post-editing

As automated journalism's technological constraints can affect news stories' composition, journalists have stated that further editorial input is sometimes needed to improve their composition (Diakopoulos, 2020). Journalists who work with automatically generated news texts deploy copy-editing techniques similar to those used to increase the readability of press releases. When working with press releases, journalists create "shorter and less complex sentences, [use] everyday words, [replace] numbers and symbols by words, and insert short bits of background information" (Maat, 2008: 87). Brooks and Pinson (2022) describe how journalists edit wire service news reports by reformatting and re-editing those reports' style, spelling, and grammar to match their in-house stylebooks. However, in the case of the post-editing of automated journalism, journalistic editing is also aimed at mending the specific textual shortcomings of data-driven automated text generation and is, thus, specific to the textual particularities of this news production type.

Research has shown that journalists' evaluations of automatically produced news stories are strongly driven by their perceptions of readers' expectations and of how readers will assess the comprehensibility of the writing (Thäsler-Kordonouri and Barling, 2023). In post-editing, journalists claim to make alterations to automatically produced article texts to increase their readability and comprehensibility (Thäsler-Kordonouri, 2024). Empirical investigation has shown that automated post-edited articles feature significantly fewer numbers than their fully automated progenitors—a post-editing step motivated by the journalistic assumption that readers would find texts containing too much data difficult to understand (Thäsler-Kordonouri, 2024).

Bridging the gap: Comprehensibility, article composition, and automation

The previous sections suggest that when investigating how the comprehensibility of data-driven reporting is related to article composition, it is relevant to consider certain production-specific factors that can have an impact on how the article is composed (see, e.g., Göpferich, 2009). The numbers-heavy journalistic routine of data-driven journalism already leads to an increased prevalence of numerical information. As readers are normally rather ill-prepared to cognitively process numbers in news stories (Callison et al., 2009; Zillmann et al., 2009), how a story is composed is, therefore, particularly relevant in this reporting context so as not to reduce the story's comprehensibility.

These days, the numbers-heavy journalistic routine of data-driven journalism just described often takes place within the technological affordances of template-based automation systems, a production process that can also affect article composition. Specifically, such automation may create articles that contain traces of the rigid templates with which they can be produced (Diakopoulos, 2019), a potentially lower incidence of certain key editorial features of journalistic writing (Caswell, 2019), and an increased incidence of numerical information (Thäsler-Kordonouri, 2024).

These changes to article composition may be to the detriment of comprehensibility. Given the uneasy relationship that readers have with numbers in the news (Callison et al., 2009; Zillmann et al., 2009), an increased occurrence of numbers in articles created with automation may negatively impact the perceived comprehensibility of the articles. Furthermore, as research suggests that readers' perceptions of the accessibility of a news story are positively affected by editorial features that "describe the experience of an individual somehow involved in the issue" (Gibson and Callison, 2017: 164), including quotes, a more limited occurrence of such features (see Caswell, 2019) in automated articles may also negatively impact comprehensibility perceptions.

Research investigating how journalists post-edit automatically generated news stories has already touched on these considerations, finding that journalists believe it may be necessary to editorially adjust automatically generated stories before publication to improve their comprehensibility (Thäsler-Kordonouri and Barling, 2023). However, to date there is no research of which we are aware on how such adjustments affect comprehensibility.

To fill this research gap, this study explores how readers' perceptions of the article composition of data-driven reporting produced with and without the help of automation affect their assessments of such reporting's comprehensibility. In conducting this investigation, we take a holistic approach, addressing article composition features specifically related to the use of numbers as well as more general ones.

Taking the readers' perspective into account

Investigations of a text's comprehensibility are most accurate and reliable when reader-focused, "because [comprehensibility] is a relative text quality which depends on the audience, whose comprehension and comprehension problems are central for its

evaluation" (Göpferich, 2009: 32). However, researchers appear not to have yet investigated the relationship between readers' perceptions of the composition of data-driven news stories—including those produced with the help of automation—and the stories' comprehensibility. Furthermore, most perception studies of automated journalism have not considered the practice of journalistic post-editing (the only partial exceptions are: Waddell, 2019; Wölker and Powell, 2021).

Research into the perception of automated and manually written data-driven reporting has reported differences in readers' evaluations of texts' 'readability' and 'quality', with a supposed "huge advantage for human-written news" in terms of 'readability' and a "small advantage for human-written news" in terms of 'quality' (Graefe and Bohlken, 2020: 50). However, these studies did not investigate the reasons behind readers' evaluations, for instance by asking them why they perceived the manually written articles as better-written and, thus, more readable than their automated counterparts.

Furthermore, the comparability of these studies' results is limited as scholars followed different strategies to pair automated and manually written news articles to be used as stimulus material, for instance by commissioning professional journalists (Jung et al., 2017; Melin et al., 2018) or journalism students (Van der Lee et al., 2018) to write articles to pair with automated stories. In some instances, automated and manually written stories were paired even though they were written in different journalistic styles (Clerwall, 2014). Some scholars edited the articles before showing them to readers (e.g. Jung et al., 2017; Tewari et al., 2021). And often, only a few sets of stimuli were investigated (e.g. Wölker and Powell, 2021), limiting generalisation about the larger message category of automated journalism, as "any particular message chosen to represent any message category must be assumed to differ from other members of the category in unknown and indefinitely numerous ways" (Jackson and Jacobs, 1983: 171).

The present study attempts to overcome these methodological shortcomings to better evaluate readers' relative evaluations of news articles produced with the help of automation and without. To reflect how the automated articles we used in this study are actually incorporated into journalistic workflows, we only included automated news articles that had been post-edited—further developed by journalists and published by news outlets.

Based on the previous sections' insights—and findings from a qualitative pre-study (Stalph et al., 2023) in which we identified relevant perception criteria related to article composition—we ask the following questions:

> RQ1. How do readers evaluate the composition of data-driven news articles that were produced with and without the help of automation?
>
> RQ2. How do those evaluations of article composition affect readers' perceptions of the comprehensibility of data-driven articles produced with and without the help of automation?

Methodology

The data used in this study comes from a large-scale between-subjects online survey experiment using a sample representative of UK online news consumers by age and

gender. To accommodate the focus of this paper, cases related solely to readers' evaluations of manually written and automated post-edited data-driven articles, which covered a range of 12 topics, were included in the analysis ($N = 3{,}135$, $n_{manually\ written} = 1{,}542$, $n_{automated\ post\text{-}edited} = 1{,}593$). To try to ensure the evaluations of the articles had high external validity, each article was read by an average of 130 respondents. The consequent large overall sample size meant that we were able to detect small effects, which is important because the perception of journalistic texts is influenced by numerous parameters, a selection of which were included in this study. The perception criteria measured in the survey were developed in a qualitative pre-study (Stalph et al., 2023) and the survey instrument was pre-tested using developmental expert reviewing and cognitive interviews with respondents (Willis, 2016). The study was approved by the Research Ethics Committee of the Faculty of Social Sciences at Ludwig-Maximilians-University Munich (LMU) for compliance with ethical guidelines.

Experimental stimuli

The experimental stimuli (see Table A in Supplemental Material) were published data-driven news articles from local, regional, and national British online news websites. The websites represent various news outlets and publishers with different geographical foci, funding models, target audiences, and ownership structures. The articles cover various topics including public health, crime, sport, transport, and social affairs. To eliminate potentially confounding variables, we stripped the articles of bylines, photographs, advertising, and the publishers' logos and branding and showed respondents only the articles' text in basic HTML formatting. The automated post-edited articles ($N = 12$) had been developed directly from automated stories generated by the RADAR news agency, which is partly owned by the UK's national news agency, PA Media, and distributes automatically generated news stories to local and regional newsrooms in the UK. Typical post-editing steps applied to RADAR stories include adding quotes from local spokespeople or deleting content deemed irrelevant to the target audience. The manually written articles ($N = 12$)—whose provenance was confirmed in personal correspondence with the articles' authors—drew on the exact same data used in the automated post-edited versions. The articles were found via extensive online research. To minimise confounding variables, the articles in each of the 12 story sets (each set consisted of one manually written and one automated post-edited article) were not only based on the same data source(s) but also featured the same story angle and covered the same locality.

Interview-based pre-study

To investigate the criteria with which readers evaluate data-driven news reporting—including the fully manually written and automated post-edited kind—we conducted eight online group interviews with 31 UK news consumers from demographically diverse backgrounds. In the interviews, we showed participants 21 sets of English-language, data-driven news articles (56 articles in total), produced with and without the help of automation, covering a range of topics, and published by UK and international news

organisations. Based on these group interviews, we extracted 28 perception criteria, which we grouped into four dimensions (Stalph et al., 2023). As explained below, some of these criteria were the basis for the measures used in this study.

Perception criteria and measures

We asked respondents about the *overall comprehensibility* of the articles they read and how well they *comprehended the numbers* in the article. The interview-based pre-study had shown that both criteria mattered to readers (Stalph et al., 2023). These variables were measured on a continuous scale from 0 to 60, modelled after the affective slider (Betella and Verschure, 2016). Perceptions of article composition were measured using various criteria and dimensions that the previous qualitative study also showed mattered to the reader (Stalph et al., 2023). These included *tone*, *word choice*, *language repetitiveness*, use of *descriptive language*, *sentence length*, *paragraph length*, *narrative flow*, *narrative structure*, *volume of numbers*, and the use of *rounded numbers*, *percentages*, *numeric analogies*, and *absolute numbers*.[1] The wording of all questions is given in the Supplemental Material.

Instrument pre-test

The questionnaire was pre-tested in October 2022 after its initial development. We used developmental expert reviewing and cognitive interviews ($N = 10$) with respondents, involving think-aloud complemented with verbal probing procedures (Willis, 2016). Our goal was to identify and repair problematic measures, questions, and concepts; and issues with the usability of the survey. To generate feedback, we administered a survey prototype containing nine different articles. The articles covered different topics to test variations in the responses related to individual articles and story topics. Each article was sourced from Birmingham news outlets, and we only recruited pre-test candidates who lived in the Birmingham area. By doing so, we ensured that our respondents were only exposed to stories relevant to their geographic interest. We modified questions and repaired survey defects based on the feedback collected from the pre-test interviews (Willis, 2016). Before the survey's broader distribution, we conducted a soft launch with approximately 100 respondents.

Survey administration and data collection

YouGov fielded the survey to their proprietary online panel between 26 January and 1 March 2023. Participants had to be at least 18 years of age and use online news media at least once per month to be eligible for participation. Respondents only read articles that were about where they lived. Thereby, we ensured that respondents were only exposed to articles that were relevant geographically. To make sure that the recruited participants read the stimulus article, only respondents who passed an attention check were included in the final sample (Ruble, 2017). Each of the 24 experimental treatment groups analysed in this study comprised at least 100 participants, with common quotas set on age and gender for the recruitment of each group. As shown in Table 1, there were no statistically significant differences between the respondents who read the fully

Table 1. Gender and age distribution of the participants, broken down by article type read.

Participants in the experimental groups that were exposed to	Gender (%)		Age bands (%)		
	Male	*Female*	*18–34*	*35–55*	*56+*
Fully manually-written articles	43.8	56.2	16.9	42.7	40.5
Automated post-edited articles	45.1	54.9	16.9	41.2	41.8

Note: Differences between respondents in the two article groups are not statistically significant in terms of distributions of gender or age.

manually written and the automated post-edited articles in terms of distributions of gender or age.

Results

Recoding of non-linear variables

To measure the valence of readers' judgements about the compositional features of the articles in the survey experiment, we opted for non-linear measures of some variables. To prepare these variables for the linear statistical models used in our analysis, we followed the recoding procedure recommended by Preston and Colman (2000). Every variable was recoded to a 0 to 100 quasi-metric scale.

Factor analysis

To detect any underlying structure of the article composition features that matter in readers' evaluations of data-driven journalistic articles, we performed a factor analysis (Promax rotation; KMO = 0.797, Bartlett test: X^2 = 8205.28, df = 91, p < .001). All items with loadings below r = 0.40 were deleted from the factor solution. To develop a stable factor prediction, we additionally controlled for cross-loadings and large deltas between items loading on the same factor. Based on theoretical considerations, we made classification decisions for the three items: 'word choice', 'descriptive language', and 'volume of numbers'.

Eigenvalues and scree plot predicted a solution with three distinct multi-item factors (see Table 2). The first factor comprises all queried *numeric features* of the news articles (M = 78.41, SD = 22.28, α = 0.635). For theoretical reasons, we added 'volume of numbers' to this factor. The other two multi-item factors describe the article's *writing style* (M = 60.94, SD = 20.12, α = 0.674) and *sentence and paragraph length* (M = 88.76, SD = 18.65, α = 0.611).

Two items did not fit the three-factor solution: *word choice* had a relatively low loading on the *sentence and paragraph length* factor. To reach a solution that was as stable as possible given the exploratory nature of this paper, we opted to drop the *word choice* item from the factor solution and integrate it as a single-item factor in the analyses (M = 92.53, SD = 16.53). The *descriptive language* (M = 77.58, SD = 33.84) item loaded on none of the factors in our final solution. So, again, we opted to make it a single-item factor in our analyses.

Table 2. Factor analysis of the perceptions of article composition features of data-driven news articles ($N = 24$) by UK online news consumers ($N = 3135$).

Original features	Post analysis factors		
	Numeric features	Writing style	Sentence and paragraph length
Use of absolute numbers	.762		
Use of rounded numbers	.691		
Use of percentages	.606		
Use of numeric analogies	.583		
Volume of numbers*	.524	.433	
Narrative structure		.855	
Narrative flow		.842	
Repetitive language*		.673	.433
Tone		.502	
Sentence length			.806
Paragraph length			.775
Word choice‡			.483
Descriptive language‡			
Article length†			

Note: Items with cross-loadings (marked with *) were grouped with theoretically similar factors. Items with low factor loadings but high explanatory power (marked with ‡) were selected as single-item factors. Items with non-loadings with theoretically similar factors and low explanatory power (marked with †) were excluded from further analyses.

Group differences

To answer RQ1, we conducted t-tests to detect non-random differences in reader perceptions of the two article types (automated post-edited and manually written) regarding the composition factors, the articles' overall comprehensibility, and comprehension of the numbers. These tests showed several significant differences (see Table 3).

Both comprehension of numbers (t (3131.34) = 5.77, $p < .001$, $d = 0.21$) and overall comprehensibility (t (3131.54) = 4.67, $p < .001$, $d = 0.17$) were rated significantly better in the manually written stories than in the automated post-edited stories. Similar patterns were found regarding some of the articles' composition factors. Specifically, readers evaluated the manually written stories better than the automated post-edited ones regarding their use of *numeric features* (t (3131.76) = 4.02, $p < .001$, $d = 0.14$), *descriptive language* (t (3132.98) = 1.99, $p < .05$, $d = 0.07$), and *word choice* (t (3102.03) = 3.08, $p < .01$, $d = 0.11$). However, reader evaluations of the multi-item factors *writing style* and *sentence and paragraph length* showed no significant group differences. We must note here that where there *were* significant differences between groups, those differences were rather small on the large scales developed in our recoding process, as indicated by the small effect sizes (<0.2) (Cohen, 1988). This observation aligns with prior research that compared readers' evaluations of news texts produced with and without the help of

Table 3. Comparisons of readers' perceptions of article composition factors (post factor analysis) and comprehensibility in manually written and automated post-edited news articles (recoded scales).

		M	SD	T	df	p	Cohen's d
Overall compre-hensibility	Manually written	76.79	21.51	4.67	3131.54	.001***	.167
	Automated post-edited	73.10	22.71				
Numeric compre-hensibility	Manually written	74.07	22.12	5.77	3131.34	.001***	.206
	Automated post-edited	69.38	23.38				
Numeric features	Manually written	80.03	21.63	4.02	3131.76	.001***	.143
	Automated post-edited	76.84	22.79				
Writing style	Manually written	60.89	20.26	−0.10	3133	.920	−.004
	Automated post-edited	60.97	19.98				
Sentence and paragraph length	Manually written	89.12	19.03	1.06	3133	.288	.038
	Automated post-edited	88.41	18.27				
Word choice	Manually written	93.45	15.37	3.08	3102.03	.01**	.110
	Automated post-edited	91.64	17.55				
Descriptive language	Manually written	78.79	33.31	1.99	3132.98	.05*	.071
	Automated post-edited	76.40	34.32				

Note: All variables in this analysis were recoded to scales from 0 (negative evaluation) to 100 (positive evaluation).
*p < .05, **p < .01, ***p < .001.

automation. Those studies have mostly shown that, where differences in perception parameters such as credibility (Tandoc et al., 2020; Wölker and Powell, 2021) and perceived bias (Wu, 2020) were detected, they were small.

Additionally, we present the identified group differences in Table B in the Supplemental Material using the original scales. As, in the original scales, respondents were asked whether they would have liked more or fewer of the respective composition features in the articles, we can directly derive recommendations for journalistic practice from this analytical step, which we present in the discussion. These results are consistent with those of the t-tests presented in Table 3.

Mediation models

To answer RQ2 and investigate the relationship between article type, article composition factors, and comprehensibility, we performed two (multiple) mediation models with PROCESS (Model 4, 5000 bootstrap samples; Hayes, 2022). Each model describes the

relationship between the article type (manually written or automated post-edited) and comprehensibility. As readers' assessments of the two article types differed significantly regarding *numeric features*, *descriptive language*, and *word choice*, these three composition factors were used as mediators to investigate how they might have contributed to the perceived comprehensibility differences. All factors were regressed on both the numeric and overall comprehensibility of the articles as, firstly, these factors co-exist in non-experimental environments and thus impact real-life understanding of news articles; and, secondly, there may be interactions between linguistic and numeric features.

In the first model, we investigated the influence of the article type (automated post-edited or manually written) on the perceived comprehensibility of numbers and statistics in the reporting via the perception of the composition factors (see Figure 1). The total effect model was significant ($b = -4.69$, $p < .001$). Overall, the factors tested explained 14.48% of the variance of numeric comprehensibility. There was a significant direct effect of article type on numeric comprehensibility ($b = -3.29$, LLCI $= -4.78$, ULCI $= -1.80$). All three tested mediators had a direct impact on numeric comprehensibility. In the case of *word choice* and *numeric features*, the direct effect of article type on numeric comprehensibility was reinforced through mediation: we found a negative indirect effect through *word choice* ($b = -0.47$, LLCI $= -0.78$, ULCI $= -0.18$), and *numeric features* ($b = -0.75$, LLCI $= -1.18$, ULCI $= -0.38$). However, there was no significant indirect effect via *descriptive language* ($b = -0.18$, LLCI $= -0.37$, ULCI $= 0.00$).

The second model (Figure 2) replicated these associations. Results were similar to our first model: the total effect model was significant ($b = -3.69$, $p < .001$); the tested factors in combination explained 15.36% of the variance of overall comprehensibility. Again, there was a significant direct effect of article type on overall comprehensibility
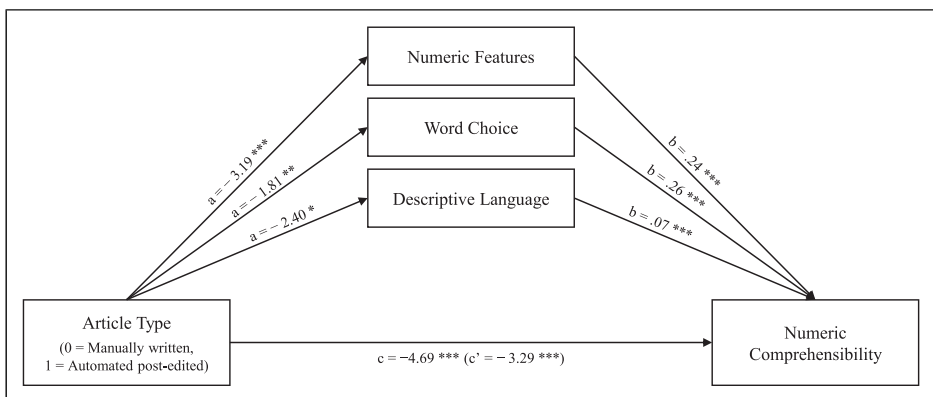

Figure 1. Mediation model with significant article composition factors and numeric comprehensibility. * $p < .05$, ** $p < .01$, *** $p < .001$.
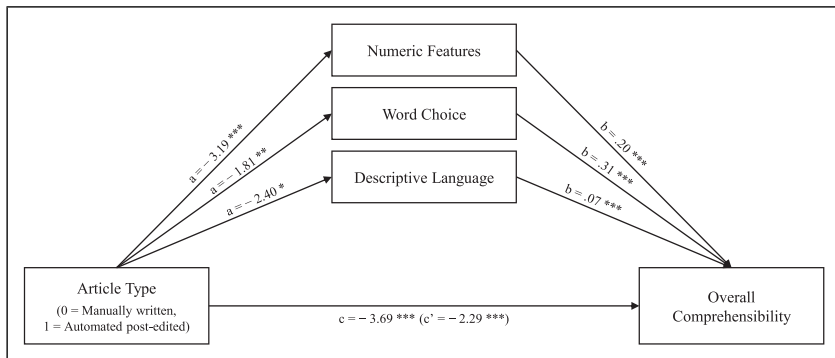
Figure 2. Mediation model with significant article composition factors and overall comprehensibility. * *p* < .05, ** *p* < .01, *** *p* < .001.

(*b* = −3.32, *p* < .001) that was reinforced through mediation. We thus found negative mediations via *word choice* (*b* = −0.36, LLCI = −0.65, ULCI = −0.11), and *numeric features* (*b* = −0.67, LLCI = −1.04, ULCI = −0.34), However, there was no significant indirect effect via *descriptive language* (*b* = −0.18, LLCI = −0.38, ULCI = −0.00).

Discussion

Scholars have argued that the composition of texts—including journalistic ones—determines their comprehensibility (Göpferich, 2009). The more relevant that data-driven reporting becomes for democratic participation (Nguyen and Lugo-Ocando, 2016), the more important it is to investigate how the composition of news articles influences their comprehensibility. Such investigation makes it possible to deduce how data-based information can better be communicated to readers. Journalism educators have recommended that data-driven news texts should use numbers sparingly and present data in particular ways, for example using rounded numbers and numeric analogies (see, e.g., Waite, 2020). Some of these recommendations are supported by empirical evidence in terms of their effects on readers' recall (Nguyen et al., 2022) and how 'helpful' readers considered articles to be (Hullman et al., 2018).

Since some of these recommendations were made, data-driven journalism has become increasingly automated. This technological development directly affects the composition of data-driven news articles (Caswell, 2019; Diakopoulos, 2019) and is, therefore, another reason for investigating comprehensibility perceptions. From a journalistic perspective, automated news production can decrease the appeal and comprehensibility of news articles (Thurman et al., 2017); therefore, journalists often decide to post-edit automatically generated stories before publishing them (Thäsler-Kordonouri and Barling, 2023).

Given journalism educators' and researchers' recommendations about writing with numbers, the influence of news automation on article composition (Caswell, 2019), and

journalists' motivations for post-editing (Thäsler-Kordonouri and Barling, 2023), this study investigated how readers evaluated the composition of data-driven articles produced with and without the help of automation (RQ1) and how those evaluations affected readers' perceptions of the articles' comprehensibility (RQ2).

Our analysis showed that the investigated article composition features can be grouped into three multi-item factors: *numeric features*, *writing style*, and *sentence and paragraph length*. The multi-item factor *numeric features* includes composition features that relate to the presentation of numbers and statistics, including the use and volume of absolute numbers, percentages, rounded numbers, and numeric analogies. The multi-item factor *writing style* includes the article's perceived tone, narrative structure and flow, and repetitiveness. Furthermore, we defined two single-item factors: *descriptive language*, that is language that helps readers picture what the story is about, and *word choice*, that is the presence of words, phrases, or abbreviations that readers are unlikely to understand or that they perceive as inappropriate. These factors cover a range of journalistic stylistic devices, including those specific to data-driven journalism, such as the presentation of numbers and statistics.

Unsurprisingly, as this is an investigation into the perception of news texts, the discriminatory power of the five factors identified here has limitations because some items had double-loadings. For example, the 'volume of numbers' item double-loaded on *numeric features* and *writing style*. This makes sense as the perceived volume of numbers in a news text is an aspect that plays into the perceived presentation of numbers in the text and also into perceptions of its overall style. We decided to group double-loading items with theoretically similar items in such cases. For instance, we grouped 'volume of numbers' with the *numeric features* factor. With regard to the single-item factors, no factor loading was observed.

Given these limitations, it would be fruitful to test our results in follow-up research to improve our knowledge of the factors determining reader perceptions of article composition. Furthermore, future research could investigate if the identified composition factors and their respective items apply to other journalistic genres besides data-driven journalism. Nevertheless, these five factors give an unprecedented insight into the perception of article composition of data-driven reporting.

In the comparison of automated post-edited and manually written articles (RQ1), reader evaluations differed regarding *descriptive language*, *word choice*, and *numeric features* but not regarding *writing style* and *sentence and paragraph length*. Readers were slightly but significantly more satisfied with the *numeric features*, amount of *descriptive language*, and *word choice* in the manually written stories than in the automated post-edited ones. These results suggest that with regard to *writing style* and *sentence and paragraph length*, the automated post-edited articles were of similar quality—in the eyes of the readership—to the fully human-written reporting.

Diakopoulos (2020) has suggested that the template-based automation approach of RADAR—the original source of the automated post-edited articles used in our experiment—generates output of high editorial quality. Our results confirm this assessment, as RADAR articles, with additional journalistic post-editing, were evaluated similarly to articles written manually in terms of *writing style* and *sentence and paragraph length*. Further empirical research could analyse to what extent the perceived similarity of

the automated post-edited stories to the manually written ones is due to a well-adapted automation process and to what extent it is due to the additional editorial effort in post-editing. Furthermore, it can be assumed that with the increasing computational capabilities of automation systems that use generative AI instead of—or in addition to—templates, the subsequent human input that may be required to improve news stories in these regards could be reduced. Therefore, future research could investigate how the post-editing of automated journalism that uses generative AI differs from that which uses rule-based algorithms and manually written templates.

Although all *numeric features* items were perceived as being too prevalent in both manually written and automated post-edited articles, participants who read the manually written stories were slightly but significantly more satisfied with the overall quantity of numbers and the use of absolute numbers, rounded numbers, and numeric analogies in the reporting (see Table 3 and Table B in the Supplemental Material). A general dissatisfaction with the quantity of all types of numerical information might be explained by the methodological design of this study, which focused on data-driven journalism—a journalistic genre that relies heavily on numbers. Still, the results showed that readers of the automated post-edited articles were significantly less satisfied in this regard. Therefore, when initially templating or subsequently post-editing automatically generated articles, journalists should aim to further reduce the overall quantity of all number types.

This preference for the manually written over the automated post-edited articles also manifested in readers' evaluations of the *word choice* and *descriptive language* used in the articles. Specifically, although the mean values indicate that the manually written reporting did not have perfect *word choice* in the eyes of the readership, respondents who read the automated post-edited stories were slightly but significantly less satisfied in that regard. As *word choice* addresses readers' evaluations of the appropriateness and accessibility of the language used, these findings suggest that the news articles produced with the help of automation contained too many incomprehensible or inappropriate terms, which were not removed by post-editing. Therefore, when initially templating or subsequently post-editing automatically generated articles, journalists should pay even more attention to explaining any technical terms, phrases, or abbreviations that readers are unlikely to understand and to the deletion of inappropriate wording.

Similarly, readers were slightly but significantly more satisfied with the amount of language that helped them picture what the story is about in manually written articles than in the automated post-edited ones. This might be because automatically generated texts rely more on pragmatic phrasing than manually written texts do (Diakopoulos, 2019). Therefore, when initially templating or subsequently post-editing automatically generated stories, journalists should invest even more time into making the reporting language descriptive to help the reader picture what the story is about.

Generally, readers perceived the manually written stories as significantly more comprehensible overall and in terms of numbers than the automated post-edited stories. Mediation analyses showed that reader perceptions of *numeric features* and *word choice*, significantly impacted these comprehensibility evaluations (RQ2). Taken together, all compositional factors explained more than 14% of the variance of perceived comprehensibility overall and perceived comprehensibility of numbers and statistics. These

empirical findings support scholarly arguments regarding the impact of a text's composition on its perceived comprehensibility (Göpferich, 2009).

In all mediation analyses, we observed an indirect effect through *numeric features* and *word choice* on the comprehensibility variables. Thus, the way data-driven stories are produced (with or without automation) influences perceptions of how numbers are presented in the article and what wording is used, with these perceptions, in turn, affecting perceived comprehensibility. In this case, the use of automation for news production negatively impacted reader perceptions of the *numeric features* and *word choice* in the reporting, which, in turn, negatively impacted perceived comprehensibility. Additionally, we observed direct effects on comprehensibility via *descriptive language*.

The mediation models emphasise the relevance of *numeric features*, as they influence the perceived comprehensibility not only of numbers and statistics but of the article as a whole. It is therefore crucial to consider *numeric features* in data-driven reporting. These findings support recommendations made in data journalism handbooks that suggest journalists pay special attention to the communication of numbers and statistics to ensure comprehensibility (e.g. Livingston and Voakes, 2005).

Overall, this study showed that although news stories produced with automation were perceived equally to those produced without regarding *sentence and paragraph length* and *writing style*, they were evaluated as less comprehensible overall and with regard to the presentation of numbers and statistics. These results can be partly explained by readers' perceptions of certain composition factors, including *numeric features* and *word choice*, which were rated significantly worse in articles produced using automation than in those written manually. The results suggest that news production automation changes news articles' perceived composition in ways that negatively impact comprehensibility. Even subsequent editorial input could not compensate for this effect.

Conclusion

This perception study provides insights into readers' evaluations of the comprehensibility of data-driven reporting by examining how these evaluations are related to article composition. Results show how the automation of data-driven news production, despite subsequent editorial investment by journalists, changes the perceived compositional makeup of data-driven news stories in ways that negatively impact their perceived comprehensibility.

These results indicate the importance not only of maintaining human involvement in the production of data-driven news content, but of refining it. Our findings suggest that when initially templating or subsequently post-editing automatically generated stories, journalists should aim to further reduce the quantity of numbers, better explain words that readers are unlikely to understand, change inappropriate wording, and increase the amount of language that helps the reader picture what the story is about. With the automation of data-driven news production increasing (Diakopoulos, 2019), journalistic practice can be supported by research into how journalists should work with automatically generated news. We hope this study has made a contribution in this regard.

ORCID iDs

Sina Thäsler-Kordonouri ⬤ https://orcid.org/0000-0003-3455-9315
Neil Thurman ⬤ https://orcid.org/0000-0003-3909-9565
Florian Stalph ⬤ https://orcid.org/0000-0003-4018-0918

Supplemental Material

Supplemental material for this article is available online.

Note

1. The variables included in the present study used a variety of response scales. For data analysis purposes, these scales were standardised using Preston and Colman's (2000) method.

References

Associated Press (2022) *The Associated Press Style Handbook*. 56th edition. New York City: Basic Books.

Beckett C and Yaseen M (2023) *Generating change: A global survey of what news organisations are doing with AI*. London: Report for LSE's JournalismAI Project. https://www.journalismai. info/research/2023-generating-change.

Betella A and Verschure PFMJ (2016) The affective slider: a digital self-assessment scale for the measurement of human emotions. *PLoS One* 11(2): e0148037. DOI: 10.1371/journal.pone. 0148037.

Brooks BS and Pinson JL (2022) *The Art of Editing in the Age of Convergence*. 12th edition. New York and London: Routledge.

Callison C, Gibson R and Zillmann D (2009) How to report quantitative information in news stories. *Newspaper Research Journal* 30(2): 43–55. DOI: 10.1177/073953290903000205.

Caswell D (2019) Structured journalism and the semantic units of news. *Digital Journalism* 7(8): 1134–1156. DOI: 10.1080/21670811.2019.1651665.

Clerwall C (2014) Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice* 8(5): 519–531. DOI: 10.1080/17512786.2014.883116.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences 2nd Edition*. New York: Lawrence Erlbaum Asoociates Publishers.

Diakopoulos N (2019) *Automating the News: How Algorithms Are Rewriting the Media*. Cambridge, London: Harvard University Press.

Diakopoulos N (2020) Computational news discovery: towards design considerations for editorial orientation algorithms in journalism. *Digital Journalism* 8(7): 945–967.

Gibson R and Callison C (2017) Effects of statistical information in news reports on individuals' recall and understanding of events and issues: Implications for journalistic practices. In: Nguyen A (ed) *News, Numbers and Public Opinion in a Data-Driven World*. London: Bloomsbury Academic & Professional, pp. 163–176.

Göpferich S (2009) Comprehensibility assessment using the karlsruhe comprehensibility concept. *The Journal of Specialised Translation* 11: 31–52.

Graefe A and Bohlken N (2020) Automated journalism: a meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication* 8(3): 50–59. DOI: 10.17645/mac.v8i3.3019.

Hayes AF (2022) *Introduction to Mediation, Moderation, and Conditional Process Analysis 3rd Edition*. New York: The Guilford Press.

Hornikx J (2007) Is anecdotal evidence more persuasive than statistical evidence? A comment on classic cognitive psychological studies. *Studies in Communication Sciences: Journal of the Swiss Association of Communication and Media Research* 7(2): 151–164. DOI: 10.5169/seals-791083.

Hullman J, Kim Y-S, Nguyen F, et al. (2018) Improving comprehension of measurements using concrete re-expression strategies [conference submission]. In: *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montréal, QC, Canada: Association for Computing Machinery. DOI: 10.1145/3173574.3173608.

Isaacs T and Trofimovich P (2012) Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34: 475–505. DOI: 10.1017/S0272263112000150.

Jackson S and Jacobs S (1983) Generalizing about messages: Suggestions for design and analysis of experiments. *Human Communication Research* 9(2): 169–181. DOI: 10.1111/j.1468-2958.1983.tb00691.x.

Jung J, Song H, Kim Y, et al. (2017) Intrusion of software robots into journalism: the public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior* 71: 291–298. DOI: 10.1016/j.chb.2017.02.022.

Livingston C and Voakes PS (2005) *Working with Numbers and Statistics: A Handbook for Journalists*. 1st edition. New York and London: Routledge. DOI: 10.4324/9781410612960.

Lowrey W and Hou J (2021) All forest, no trees? Data journalism and the construction of abstract categories. *Journalism* 22(1): 35–51. DOI: 10.1177/1464884918767577.

Maat HP (2008) Editing and genre conflict: How newspaper journalists clarify and neutralize press release copy. *Pragmatics* 18(1): 87–113. DOI: 10.1075/prag.18.1.05pan.

McCombs ME, Cole RR, Stevenson RL, et al. (1981) Precision journalism: an emerging theory and technique of news reporting. *Gazette* 27(1): 21–34. DOI: 10.1177/001654928102700103.

Melin M, Bäck A, Södergård C, et al. (2018) No landslide for the human journalist: An empirical study of computer-generated election news in Finland. *IEEE Access* 6: 43356–43367. DOI: 10. 1109/ACCESS.2018.2861987.

Meyer P (1975) *Precision Journalism: A Reporter's Introduction to Social Science Methods*. Oxford: Rowman & Littlefield.

Newman N (2024) *Journalism, media, and technology trends and predictions 2024*. Oxford: Reuters Institute for the Study of Journalism. DOI: 10.60625/risj-0s9w-z770.

Nguyen A and Lugo-Ocando J (2016) The state of data and statistics in journalism and journalism education: issues and debates. *Journalism* 17(1): 3–17. DOI: 10.1177/1464884915593234.

Nguyen HA, Hofman JM and Goldstein DG (2022) Round numbers can sharpen cognition. [conference submission]. In: *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, USA: Association for Computing Machinery. DOI: 10.1145/3491102.3501852.

Parasie S and Dagiral E (2013) Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New Media and Society* 15(6): 853–871. DOI: 10.1177/1461444812463345.

Preston CC and Colman AM (2000) Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104(1): 1–15. DOI: 10.1016/S0001-6918(99)00050-5.

Ruble RA (2017) Experimental manipulation. In: Allen M (ed). *The Sage Encyclopedia of Communication Research Methods*. London: Sage Publications Ltd, 475–477. DOI: 10.4135/9781483381411.

Stalph F, Thurman N and Thäsler-Kordonouri S (2023) Exploring audience perceptions of, and preferences for, data-driven 'quantitative' journalism. *Journalism* 0(0): 1–21. DOI: 10.1177/14648849231179606.

Tandoc EC, Yao LJ and Wu S (2020) Man vs. Machine? The Impact of Algorithm Authorship on News Credibility. *Digital Journalism* 8(4): 548–562. DOI: 10.1080/21670811.2020.1762102.

Tewari S, Zabounidis R, Kothari A, et al. (2021) Perceptions of human and machine-generated articles. *Digital Threats: Research and Practice* 2(2): 1–16. DOI: 10.1145/3428158.

Thäsler-Kordonouri S (2024) Exploring AI integration in UK newsrooms: an investigation into the use and evaluation of news agency automated journalism. In: *IAMCR Presidential PhD Research Webinar.* [conference submission].

Thäsler-Kordonouri S and Barling K (2023) Automated journalism in UK local newsrooms: attitudes, integration, impact. *Journalism Practice* 0(0): 1–18. DOI: 10.1080/17512786.2023. 2184413.

Thurman N, Dörr K and Kunert J (2017) When Reporters Get Hands-on with Robo-Writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism* 5(10): 1240–1259. DOI: 10.1080/21670811.2017.1289819.

Thurman N, Stares S and Koliska M (2024). Audience evaluations of news videos made with various levels of automation: A population-based survey experiment. *Journalism* 0(0): 1–21. DOI: 10.1177/14648849241243189.

Van der Lee C, Verduijn B, Krahmer E, et al. (2018) Evaluating the text quality, human likeness and tailoring component of PASS: A Dutch data-to-text system for soccer [conference

submission]. *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 962–972.

Van Enschot R and Hoeken H (2015) The occurrence and effects of verbal and visual anchoring of tropes on the perceived comprehensibility and liking of TV commercials. *Journal of Advertising* 44(1): 25–36. DOI: 10.1080/00913367.2014.933688.

Waddell TF (2019) Attribution practices for the man–machine marriage: How perceived human intervention, automation metaphors, and byline location affect the perceived bias and credibility of purportedly automated content. *Journalism Practice* 13(10): 1255–1272. DOI: 10.1080/17512786.2019.1585197.

Waite M (2020) *Data Journalism with R and the Tidyverse*. https://mattwaite.github.io/datajournalism/.

Willis GB (2016) Questionnaire pretesting. In: Wolf C, Joye D, Smith TW, et al. (eds) *The Sage Handbook of Survey Methodology*. London: Sage Publications Ltd, pp. 359–381. DOI: 10.4135/9781473957893.

Wölker A and Powell TE (2021) Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism* 22(1): 86–103. DOI: 10.1177/1464884918757072.

Wu Y (2020) Is automated journalistic writing less biased? An experimental test of auto-written and human-written news stories. *Journalism Practice* 14(8): 1008–1028. DOI: 10.1080/17512786.2019.1682940.

Zillmann D, Callison C and Gibson R (2009) Quantitative media literacy: Individual differences in dealing with numbers in the news. *Media Psychology* 12(4): 394–416. DOI: 10.1080/15213260903287275.

Author biographies

Sina Thäsler-Kordonouri is a doctoral candidate and research assistant in the Department of Media and Communication (IfKW) at LMU Munich. Her research focuses on the use of AI and automation in journalism as well as related forms of computational journalism.

Neil Thurman is a Professor of Communication in the Department of Media and Communication at LMU Munich, Germany. He is also an Honorary Senior Research Fellow in the Department of Journalism at City, University of London. Between 2015 and 2024 he held a Freigeist Fellowship from the Volkswagen Foundation for a project on algorithmic news. This article is one of the outputs of that Fellowship.

Ulrike Schwertberger is a PhD student in the Department of Media and Communication LMU Munich, Germany. Her research interests lie in digital media (impact) research, with a particular focus on radicalisation dynamics and multi-screening processes in entertainment contexts.

Florian Stalph was a post-doctoral researcher in the Department of Media and Communication, LMU Munich, Germany. His research focused on data journalism, computational journalism, and the implications of digital media practice and technology for the journalistic field.