Visual Indicators to Increase the Detection of Linguistic Media Bias

Smi Hinterreiter, Anna Chelsea Bahß, Ann-Christin Gah, Timo Spinde, Isao Echizen, Marc Erich Latoschik

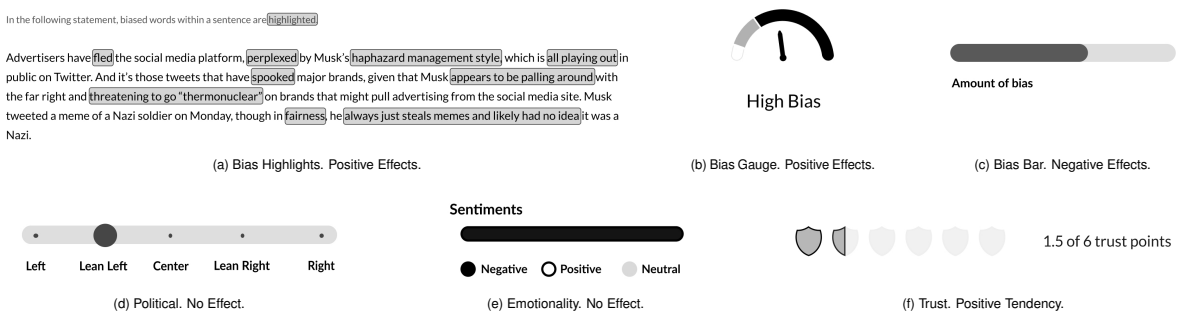

Figure 1: The six tested indicators on one example statement from the study. Top row from left to right: *Bias Highlights*, *Bias Gauge*, and *Bias Bar*. Bottom row from left to right: *Political*, *Emotionality*, and *Trust*.

Abstract—The influence of linguistic bias in online news articles is a growing concern, particularly in the context of shaping public opinion and rising political polarization. While there is a growing body of literature on indicators for misinformation, none have been sufficiently tested to counteract the influence of media bias. Hence, we design six indicators (Bias Bar, Bias Gauge, Bias Highlights, Political Scale, Sentiment Scale, and Trust Score) and test their impact on linguistic bias detection and perception in a two-phased experiment (n = 214). First, we expose participants to short, social-media-like statements along with one indicator and query bias perception. Second, we evaluate bias detection by removing the indicator and asking participants to mark biased words. In addition, we examine how trust, sharing discernment, and sentiment relate to bias perception and detection. Our results show that highlighting biased phrases and showing total bias with contextual information in a gauge significantly improve bias detection skills. However, the strongest predictor for reduced bias detection was political congruency between the statement and the participant. We conclude with design recommendations for linguistic media bias indicators in online news environments.

Index Terms—Media bias, news bias, media literacy, news literacy, information literacy, online news, partisan bias.

1 Introduction

News articles often contain bias toward political ideologies, groups, or narratives [36]. This media bias can influence public perception and voting behavior [3,36]. Perceived bias depends on readers' backgrounds [3] and can limit exposure to diverse viewpoints [30]. Media bias has multiple dimensions: selection bias [9, 43] determines coverage, framing [52] or spin bias [46] shapes presentation, and linguistic bias [29, 60] signals viewpoints and evokes emotions through word choice [57]. We focus on linguistic bias, which often goes undetected by readers [56] but can be identified by automated methods [29, 68].

Despite its prevalence, research on media bias mitigation remains sparse [57]. First studies on visual indicators, such as graphical cues that communicate bias or credibility at a glance [67], showed promising results by making linguistic bias visible [6, 58]. Yet only three studies specifically tested visual indicators in the context of media bias. Among tested approaches such as warning messages and political classifications, highlighting biased phrases has shown the most consistent results [6,56,58]. Broader work on misinformation and news credibility indicators offers more evidence on which design strategies for visual indicators might help critical evaluation of information [8]. However, prior work is difficult to compare: studies diverge in their metrics and content, and most test only one indicator at a time [6]. Critically, most studies rely solely on self-reported perception, without active detection tasks that could reveal whether indicators genuinely improve readers' ability to identify bias [58]. Further, not all study designs account for political leaning of participants and congruency effects [62].

This study addresses these gaps by identifying visual indicators for linguistic bias, then designing and comparing six indicator types using standardized measures and a detection task. Indicator type serves as the treatment variable, with a controlled first-exposure design testing transfer effects for linguistic bias and related concepts after exposure. We follow these research questions:

- **RQ1:** Which visual indicators increase linguistic bias detection?
- **RQ2:** Which visual indicators increase self-reported bias perception?
- **RQ3:** How do sharing intention, perceived emotionality, and trust relate to perceived bias?
- **RQ4:** How does political congruency moderate bias perception and bias detection?

To create the indicators, we first review related work from media bias, misinformation, and information credibility research (Section 2.1) to

- *Smi Hinterreiter is with University of Würzburg, Germany. E-mail: smi.hinterreiter@uni-wuerzburg.de.*
- *Anna Chelsea Bahß is with University of Konstanz, Germany. E-mail: a.bahss@online.de.*
- *Ann-Christin Gah is with Technical University of Munich, Germany. E-mail: ann-christin.gah@tum.de.*
- *Timo Spinde is with University of Göttingen and National Institute of Informatics, Tokyo, Japan. E-mail: t.spinde@media-bias-research.org.*
- *Isao Echizen is with National Institute of Informatics, Tokyo, Japan. E-mail: iechizen@nii.ac.jp.*
- *Marc Erich Latoschik is with University of Würzburg, Germany. E-mail: marc.latoschik@uni-wuerzburg.de.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

create a design framework (Section 3.1). We then design six indicators, shown in Figure 1, covering bias amount (*Bias Bar*, *Bias Gauge*, *Bias Highlights*), political leaning (*Political*), sentiment (*Emotionality*), and trustworthiness (*Trust*), and compare them against a control group in an online study (n = 214). The study focuses on short, social-media-like statements and includes both a bias perception measure and a word-level annotation test for evaluation beyond self-report (Section 3.2). In two phases, participants first read statements with their group's indicator before moving to a testing phase where they complete a bias detection test without their group's indicator (Figure 2).

Our results show that *Bias Highlights* and, partly, *Bias Gauge* significantly improve detection accuracy (Section 4). However, conservative or liberal political leaning and incongruence with a statement's slant are strong predictors of both detection and perception across all groups. We also find a strong correlation between perceived emotionality and perceived bias. We contribute an empirical comparison of visual representation strategies for linguistic bias indicators, evidence that phrase-level and contextualized aggregate encodings outperform abstract summary cues for transfer-based bias detection, and design implications about interpretability, reference frames, and calibration for reflective bias judgment tools under political congruence (Section 5).[1]

2 Related Work

Prior work suggests that making bias visible can raise awareness and mitigate negative effects [1, 6]. Among tested approaches for media bias are highlighting facts in the text, highlighting instances of framing, highlighting biased language [6, 56], forewarning messages, and political maps [1] or scales [58]. So far, the highlighting of biased phrases (see Appendix Figure 2) emerged as the most promising approach and has been implemented in online applications [27]. However, studies on bias mitigation remain limited, and only a small number directly test visual indicators for media bias perception [6, 56, 58], with more evidence coming from related work on news indicators for misinformation mitigation and credibility assessment.

2.1 Related Indicator Designs

To identify indicator designs for making linguistic bias visible, we reviewed 33 publications on news-related indicators, also including commonly used commercial systems such as NewsGuard [4], The Factual, and AllSides.[2]. The full table of the literature review is available in the supplementary material (Section 6). Across included work, indicators differ in the strategy they use to indicate credibility or bias, timing of the indicator, and the data type displayed. Regarding timing, some intervene before reading (e.g., forewarning messages [65]), while others provide information during reading through labels, highlights, scales, or ratings (e.g., bias highlights [58]). Other interventions use debunking strategies [51] after reading or provide alternative sources [56]. We further classify interventions into six categories based on the strategy of the intervention (see Appendix, Figure 1 and 2): Manipulation of news presentation [35, 56]; Warning messages [45, 65]; Educational prompts or messages [22, 45]; Social cues such as likes or endorsements [39]; Recommendations [4, 7, 33, 34]; Information about article or source [1, 56, 58]. Of the six strategies, information and recommendation indicators are used most frequently, especially by commercial applications.

Recommendation Indicators designed to assist users to evaluate the credibility of news can potentially be more effective in enhancing perceived credibility than fact-checking indicators [7]. Examples include The Factual's credibility score and NewsGuard's red or green shield icons, although the latter did not demonstrate significant improvements in news consumption [4]. Other work [33, 34] indicates that source ratings can impact the perceived credibility of an article, with article ratings exerting a stronger influence on believability than source ratings. Thus, recommendations present a conflicting research landscape, highlighting a need for further investigation.

Information Indicators present content-related cues outside the content itself to help readers assess its credibility and bias. For example, the highlighting of instances of framing bias [6] and biased words and phrases [58] increases media bias awareness and reduces ideological polarization. Platforms like AllSides often use political classifications to encourage readers to consider diverse perspectives. However, ideological left-right maps did not significantly affect bias perception in the study of Spinde et al. [58] while other studies describe positive effects [49].

2.2 Visual In-Text and Summary Encodings for Media Bias

Highlighting biased phrases shows the most consistent effects on bias awareness [6, 56, 58]. Text highlights act as in-text encodings that increase perceptual salience rather than summarizing content. They function as pre-attentive popout cues whose effectiveness depends on density, contrast, and whether they encode categorical or continuous distinctions [61, 67]. Like overlays, highlights guide attention within text, unlike summary indicators such as bars, scales, or icons that present aggregated judgments. In media bias contexts, highlights support cue localization and learning-by-example [24], which may enable transfer, while summary indicators support holistic assessment via heuristics [14, 63]. Their effects on linguistic bias perception and detection have not been systematically compared.

Summary indicators aggregate bias signals into a single display outside the text. Continuous bars provide accurate magnitude estimates [12, 25], but without reference points, values are hard to interpret [53]. Categorical segmentation adds labeled thresholds that reduce cognitive load in mapping values to judgments [14, 64]. Ordinal labels encode rank but not distance [40, 48], so category steps lack quantitative meaning. Composition displays (e.g., sentiment bars) increase complexity; accuracy declines with more components, and readers anchor on salient categories [54]. Icon-based displays (e.g., stars, shields) act as heuristic cues that shape credibility judgments, often independent of the encoded score [33, 38, 63].

2.3 Gaps in Indicator Studies

Existing indicator studies have four main limitations: (1) most compare a single indicator to a control group, which makes results hard to compare across studies [62], (2) they often rely only on self-reported bias perception, which does not necessarily reflect actual detection ability [59], (3) they frequently omit personal factors such as political leaning, age, and education despite their known influence on news judgments [58, 62], and (4) differences in materials and measures, such as content formats, items, and scales, limit comparability across misinformation and media bias research [62]. Misinformation studies typically use headlines and ask for binary veracity judgments or sharing discernment [8], whereas media bias studies use longer texts and measure bias perception or topic opinions on rating scales [6, 58].

We address the gaps with three contributions: (1) the first empirical comparison of six linguistic-bias indicator types that reflect commonly deployed strategies, including two indicators previously tested for media bias (*Bias Highlights*, *Political*), two novel summary designs (*Bias Bar*, *Bias Gauge*), and two adapted from credibility tools (*Emotionality*, *Trust*), (2) a transfer-based bias detection task as a behavioral measure beyond self-reported bias perception, and (3) models that include demographic and political factors. Using a transfer design, we first expose participants to an indicator and then remove it to test unaided bias detection on new short, social-media-like statements, capturing learning-like transfer rather than in-context assistance [37]. The treatment variable is indicator type rather than encoding parameters. As exploring all possible design combinations is infeasible, we adopt a top-down approach, selecting indicator types that cover commonly deployed indicator strategies displayed during reading which aim to inform or recommend, rather than exploring individual design details or exhausting the full design space. Findings should be interpreted as evidence about broader indicator strategies under first-exposure short-form conditions rather than to specific implementation choices.

[1] Code and data available at doi.org/10.5281/zenodo.19347505

[2] web.archive.org/web/20240813080330/https://www.thefactual.com/how-it-works/, www.allsides.com/

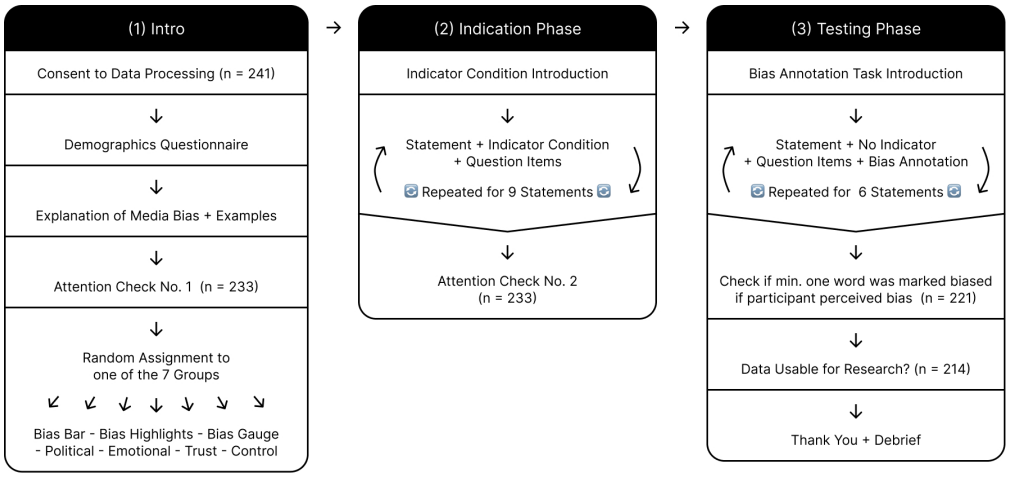

Figure 2: Study procedure with the three phases (1) Intro, (2) Indication Phase, and (3) Testing Phase, each with measures, steps, and included participants.

3 Method and Materials

We conducted a 7x3 mixed-design study with repeated measures, comparing six visual indicators with a control condition. The study had one between-subjects factor *Indicator Group* (*Bias Bar*, *Bias Gauge*, *Bias Highlights*, *Political*, *Emotionality*, *Trust*, *Control*) and the within-subject factor *Political Slant* (left, center, right). The procedure consisted of three phases displayed in Figure 2: (1) introduction to linguistic bias and attention check, the (2) indication phase, in which participants read statements with one assigned indicator, and the (3) testing phase, in which participants identified biased words without indicator support (Section 3.4).

The study combined subjective perception measures and a more objective test (Section 3.2). We measured perceived bias, perceived emotionality, trust, and sharing intention after each statement, and we measured bias detection through a word-level annotation task in the testing phase. We designed the testing phase as a transfer test: after exposure to one representation type during reading, participants completed bias detection on new statements without visual support. The study and models also included political leaning, political congruence, age, gender, education, topic opinion, and topic relevance in the models to examine how personal background shaped indicator effects.

3.1 Indicator Design

Following visual encoding theory [12, 25, 40], we used Munzner's nested model [47] as a design framework, extending it by our indicator review. We defined five design dimensions: (1) indicator strategy and goal (increasing bias detection and perception by showing information or giving recommendations), (2) indicator context, (3) data type abstraction (e.g., bias amount, biased words, sentiments), (4) visual concept and mental model fitting the data abstraction, and (5) visual design details and platform implementation. Our study holds three dimensions constant: (1) indicator strategy and goal are use information indicators and a single indicator that extends on information by adding an element of recommendation, (2) indicator context is set do during reading, and (5) visual design details like size, placement, and grayscale are standardized. The varying factors are (3) data type abstraction and (4) visual concept and mental models.

Using this framework, we selected six indicators (Table 1, Figure 1c–Figure 1f), four representing common designs in prior work (*Bias Highlights*, *Political*) or deployed systems (*Trust*, *Emotionality*) and two new summary indicator designs for bias (*Bias Bar*, *Bias Gauge*). Both bias summary indicators use bias amount as a continuous quantitative value that maps to filled length along a common scale, one of the most accurately perceived visual encodings [12, 25]. *Bias Gauge* adds categorical segmentation, providing a reference frame that reduces the cognitive demand of interpreting raw magnitudes [14, 64, 67]. The six indicators differ along data type (bias amount, political slant, sentiment, credibility) which constrains feasible visual forms and encoding types (in-text localized highlights vs. summary aggregate as ratios above the text as scales or gauges or credibility values encoded in icons like shields or stars), and cognitive role (example-based learning vs. heuristic cue vs. reference frames), making the comparison more complex than a pure encoding study and grounded in real-world deployment practices. Visual design details and implementation such as size, placement, and color were standardized across indicators to ensure comparability, while acknowledging that some characteristics are inherently tied to the indicator concept.

Three indicators focused directly on linguistic bias. *Bias Bar* and *Bias Gauge* encode the proportion of biased words in a statement as a summary cue, while *Bias Highlights* marked biased words in the text as a local in-text cue. Grey highlighting was chosen over color to minimize evaluative connotations [67]. Unlike the remaining indicators, which summarize information externally through scales, scores, or labels, *Bias Highlights* localize evidence at the word level and may support learning through repeated exposure to examples of biased language [61, 67]. Three additional indicators presented summary cues commonly used in practice: *Political* showed the political slant of the source on a five-level scale, *Emotionality* showed the proportion of positive, negative, and neutral sentiment in the statement as a pragmatic approximation for emotional intensity through sentiment polarity by NLP tools, and *Trust* summarized content credibility as a shield-based score. The trust score is calculated and shown as shields from the average score of political rating on AllSides (-2 for far right/far left, -1 for right/left, 0 for center) and amount of biased words (0: <10%, -1: <30%, -2: >30%) as a symmetric baseline in the absence of validated weights.[4]

We selected these six indicators to achieve coverage across the most common data types used in deployed bias and credibility tools (see Section 2.1 and supplementary review table), rather than to construct a fully factorial design. The comparison allowed us to test whether signals and interventions commonly used for credibility or veracity discernment also affect linguistic bias perception and detection. We further ensured that the indicator data types could be automatically computed with classifiers for later implementation. Each indicator represents a distinct data type: bias amount (*Bias Bar*, *Bias Gauge*), biased word locations (*Bias Highlights*), political slant (*Political*), sentiment composition (*Emotionality*), and credibility (*Trust*). Including two bias-amount indicators (Bias Bar and Bias Gauge) was deliberate

[4] Sensitivity checks with 25/75 and 75/25 splits correlated with the 50/50 baseline at r = .90 and r = .85 (Spearman), indicating the trust score was largely stable across weightings, though the 75/25 split produced somewhat greater divergence, suggesting that statements from outlets with high political extremeness are most sensitive to this choice.

Table 1: Summary of the six tested indicators.

Indicator, Data Type, Encoding, Mechanism & Expected Cognitive Role
Bias Bar: Continuously displays amount of biased words. (Figure 1c) **Data Type:** Bias Amount (Percentage of biased words; text-level). **Encoding:** Following Harris et al. [23] and Gao et al. [18], the *Bias Bar* shows the percentage of biased words in the statement (see Section 3.3) through a dark fill, excluding stop words. **Mechanism:** Raw quantitative summary. **Cog Role:** Compact magnitude cue.
Bias Gauge: Continuously displays percentage of biased words and categorizes the level of bias in low/medium/high. (Figure 1b) **Data Type:** Bias Amount + Frame (Percentage of biased words; text-level). **Encoding:** Drawing inspiration from Kherwa et al. [32], the *Bias Gauge*, similar to the *Bias Bar*, fills up continuously to represent bias amount. It is divided into three sections ('low bias' for bias percentages below 10%, 'medium bias' between 10% and 30%, and 'high bias' for over 30%) with corresponding text below. The thresholds were established after analyzing the percentage values in statements categorized as low, medium, and high bias during material collection (Section 3.3). The *Bias Gauge* aims to provide an intuitive understanding of the bias level, akin to recommendation labels on The Factual and NewsGuard. **Mechanism:** Reference-framed summary. **Cog Role:** Calibration and interpretable magnitude judgment.
Bias Highlights: Visually shows which words are biased. (Figure 1a) **Data Type:** Biased Words (in-text / local). **Encoding:** *Bias Highlights* marks biased words in place, using an in-text highlight that uses perceptual salience to direct attention to the specific linguistic cues that constitute bias. We create the highlights based on the expert annotations (see supplementary material). The design is included due to its promising results in previous studies [6, 58, 61]. **Mechanism:** Example-based learning. **Cog Role:** Direct localization of bias cues.
Political: Displays the outlet's political slant. (Figure 1d) **Data Type:** Slant of Outlet (1=left, 5 = right; source-level). **Encoding:** Adapted from An et al. [1] and Spinde et al. [58], *Political* employs a five-level political scale (left, lean left, center, lean right, right). The classification is based on the AllSides Media Bias Chart.[3] Despite conflicting research [1, 58], political scales remain common on platforms like AllSides or The Factual. Inclusion choice prioritized ecological relevance over theoretical completeness. **Mechanism:** Context cue. **Cog Role:** External ideological framing.
Emotionality: Displays what sentiments are in the statement. (Figure 1e) **Data Type:** % of positive/negative/neutral sentiments (text-level). **Encoding:** The indicator explores the influence of emotional language on the perception and detection of linguistic bias. While emotionality is not synonymous with bias, there may be a strong correlation between the two, making it a crucial aspect to consider. The *Emotionality* indicator visualizes the percentage of positive (white), negative (black), and neutral (grey) sentiments in the statement as a pragmatic operationalization of affective tone [23]. To derive sentiments, we used Google's Natural Language API [20]. **Mechanism:** Affect cue. **Cog Role:** Emotional salience.
Trust: Displays credibility of statement with single score/icon. (Figure 1f) **Data Type:** Aggregated score from extremeness of political slant and percentage of biased words (text- + source-level). **Encoding:** The *Trust* indicator shows a visual scale with six shields, resembling a trust-based star rating system and consistent with research on icon-based trust cues [4, 31, 33]. It combines elements of both recommendation and information. As a mental shortcut, *Trust* aims to provide a quick, intuitive understanding of the credibility of a news source or statement. The number of filled-out shields indicates the level of trustworthiness (zero = low, six = high). The score combines extremeness of the political rating of the source and the percentage of biased words in the statement 50/50. **Mechanism:** Heuristic credibility cue. **Cog Role:** Fast credibility judgment.

as they share the same underlying data but differ through the presence of a categorical reference frame established in Section 3.3, allowing a direct test of whether reference framing affects detection beyond raw magnitude encoding. To improve comparability across conditions, all indicators used a grayscale design [38] and were placed above the statement in similar size, except for *Bias Highlights*, which appeared within the text. This is not a fully controlled factorial manipulation, covers a limited amount of indicators, and the conditions differ along multiple dimensions. We therefore treat this as a coverage study of deployed indicator types for linguistic bias rather than an extensive decomposition of individual encoding variables, and interpret effects at the indicator level rather than attributing them to specific encoding features.

3.2 Measures

Our main dependent variables were bias detection and bias perception.

For detection, participants completed a word-level annotation task in which their selections were compared with the expert standard. We computed both F1 and d' scores. F1 captures the balance between precision and recall and is widely used in computer science studies on bias detection [26, 50, 57]. d' is commonly used in psychology research and captures sensitivity in distinguishing biased from unbiased words, helping to separate discrimination ability from response bias [5].

For perception, we measured perceived bias after each statement using a standard bias item [59], perceived emotionality [42], trust in the statement [2, 34], and sharing intention using scales commonly used in misinformation research [11, 13, 42]. All questions are listed in the Appendix Table 1.

In addition, we recorded gender, age, education, political leaning, news consumption, topic opinion, and topic relevance as covariates. Especially controlling for political leaning is important as articles aligning with readers' political orientation are perceived as less biased [58], known as the congruency effect. Therefore, we calculate political congruence for each statement. Political leaning is measured on a scale from -10 (very liberal) to 10 (very conservative). Further, we assess the personal relevance of the topic and corresponding opinion, following Moravec et al. [45].

3.3 Materials

We created the study materials in four steps. First, we selected two politically contested topics prominent during data collection: Elon Musk's Twitter/X acquisition and abortion. We collected 54 news articles and assigned each source a political ranking using the AllSides five-category system. Second, we extracted 40–70 word statements that remained understandable without article context. Statements were preliminarily labeled as low, medium, or high bias with a political leaning. We used short statements to match short-form news exposure on social platforms and feeds. Third, three researchers with prior experience in media bias annotation independently labeled biased words in the statements ($n = 1,134$), following the annotation guidelines of Spinde et al. [60] and definitions of linguistic bias [57, Table 3]. A word was considered biased if all three annotators agreed or, when two marked it, after discussion and consensus, following a procedure similar to Fuhr et al. [15].

The inter-annotator agreement was Krippendorff's $\alpha = 0.47$, reflecting task subjectivity. Still, this exceeds prior word-level agreement and the sentence-level agreement reported by Spinde et al. [60]. The two most experienced annotators achieved α of 0.61, indicating that experience improves agreement.

In total, 119 words were labeled as biased by all three annotators, while 97 additional words were labeled by two annotators and later resolved through discussion. For example, in the statement: *"But now members of the anti-abortion movement are grappling with whether their greatest legal triumph led to political defeat in this week's midterms. With historical trends and a curdling economy suggesting that Democrats could face historic losses, Team Blue made a risky bet on abortion rights."*, all annotators identified "grappling" and "Team Blue made a risky bet" as biased. Annotator 1 also labeled "on abortion rights," whereas Annotators 2 and 3 labeled "greatest legal triumph" and "curdling." After discussion, they agreed on "greatest legal triumph" and "curdling" as biased.

Fourth, we calculated the share of biased words in each statement, excluding stop words, and used these shares to define low-, medium-, and high-bias categories for *Bias Gauge*. Statements with 0–10% biased words were labeled low bias, 10–30% medium bias, and above

30% high bias.[5] The final stimulus set contained 15 statements: nine for the indication phase and six for the testing phase, balanced across bias level and political slants.

3.4 Procedure

The study was implemented as a custom web application and followed three phases. In the *intro* phase, participants were informed of the study's content and data processing and gave informed consent. Participants then completed demographics, and received a short explanation of linguistic bias with examples. A multiple-choice attention check ensured basic understanding of the concept of linguistic bias.

In the *indication* phase, participants were randomly assigned to one of the six indicator conditions or the control condition. They then read nine short news statements on the Twitter/X topic. After each statement, they answered six questions on perceived bias, sentiment, trust, and sharing intention. We added a directed-response item as a second attention check to identify automated or inattentive responses.

In the *testing* phase, we removed all indicators and presented six abortion-related statements. Participants were tasked to identify biased words by clicking on them and then answered the same perception questions as before. This study design estimates whether representation effects transfer to new content and persist after cue removal, rather than only measuring immediate in-situ assistance.

At the end of the study, we asked participants if their input is usable for research to ensure data quality, with assurance of full pay independent of their answer. Before the main study, a pre-test with identical setup and pay was conducted with 14 participants to identify and resolve any potential issues. As an anonymous online survey via Prolific, formal ethics board review was not required under our institution's guidelines. The consent form all participants filled out prior to participation is available in the supplementary materials (Section 6).

3.5 Statistical Analysis

We analyzed the effects of indicators on detection and perception using three linear mixed models in R[6] for F1, d' (RQ1), and perceived bias (RQ2). The fixed effects included indicator group, perceived emotionality, political leaning, political congruence (for liberals, moderates, and conservatives), age, gender, education, topic opinion, and topic relevance. We also modeled the interaction between political leaning and congruence (RQ4). Ground-truth statement bias and statement sentiment were also modeled as fixed effects to account for material differences. Participant ID was included as a random intercept.

To examine relationships among perceived bias, sentiment, trust, and sharing intention (RQ3), we computed Spearman correlations. We evaluated the internal consistency of the trust construct with Cronbach's α and Bartlett's test of sphericity. We conducted all analyses with a 95% confidence level. For an initial power analysis, we used G*Power, targeting an .8 power to detect a medium effect size of $f = .25$, with a standard α error probability of .05, resulting in 225 required participants. We preregistered the study on OSF.[7]

3.6 Participants

We recruited 241 participants from the United States via Prolific and randomly assigned them to the seven study groups. Participants were excluded if they failed the initial attention check twice ($n = 8$), reported that their data should not be used for research ($n = 7$), or did not mark any word as biased despite later indicating that statements were biased ($n = 12$). The final sample comprised 214 participants with an achieved statistical power of .78. The sample was balanced by gender (51% women, 47% men, 2% diverse), had a mean age of 37.05 years ($SD = 13.64$), and reported high English proficiency. Participants were somewhat more liberal than the general U.S. population ($M = -3.39$, $SD = 5.7$, scale -10 to 10) [17]. The educational backgrounds of participants were higher than the average U.S. population[8]: 54% held at least a four-year degree. Average news consumption was about once per day. Participants generally select news sources aligned with their political orientations. Participants identifying as centrist ($M = 2.32$, $SD = 0.80$) or conservative ($M = 2.95$, $SD = 0.84$) were more likely to consume news with a more liberal slant relative to their own political views. A linear regression model predicting news consumption frequency showed age as the only significant predictor ($R^2 = .12$, $F(1,212) = 28.19$, $p < .001$) with an 0.032 points increase for each additional year.

4 Results

4.1 Validation

First, we validated the materials and constructs. Across all three linear mixed models, residual diagnostics did not indicate substantial violations of normality or homoscedasticity. Participants' perceived bias aligned with the expert-rated bias labels and the calculated proportion of biased words ($\beta = 0.43$, 95% CI $[0.37, 0.50]$, $p < .001$; Appendix Table 4), as illustrated in Appendix Figure 3. This result suggests that the selected statements reflected different levels of linguistic bias and that participants were generally able to perceive these differences. The trust construct showed high internal consistency. Cronbach's α ranged from .95 to .98, and Bartlett's test of sphericity was significant ($\chi^2(3) = 617.50$, $p < .001$), supporting the use of the three-item trust measure in the following analyses.

4.2 Bias Detection (RQ1)

To assess bias detection, we analyzed both F1 and d' scores from the word-level annotation task (Table 2). Mean scores were highest in the *Bias Highlights* and *Bias Gauge* conditions and lowest in the *Control* and *Bias Bar* conditions, visible in Figure 3 and Figure 4. *Bias Highlights* and *Bias Gauge* showed positive effects in the F1 model, with only *Bias Highlights* showing positive effects in the d' model.

Across all participants, mean F1 was .474 ($SD = 0.194$), with recall of .32 and precision of .76. Participants therefore tended to avoid false positives more successfully than they detected all biased words. The average d' score was 1.08 ($SD = 0.452$), suggesting that while participants' ability to detect bias is limited, it is still above chance. Participants with high d' scores correctly identified 50-60% of biased words with minimal false positives. Overall, the subjective and fine-grained nature of bias makes it challenging for untrained participants to consistently identify bias [16, 60], expectedly leading to lower scores.

Table 2: Mean F1 and d' scores of all indicator groups. Higher values indicate higher ability to detect and discern bias.

Indicator	F1 (SD)	d' (SD)	Participants
Bias Bar	0.430 (0.203)	0.941 (0.729)	31
Gauge	0.526 (0.193)	1.203 (0.743)	33
Highlights	0.510 (0.186)	1.276 (0.714)	30
Political	0.446 (0.181)	1.001 (0.702)	28
Emotionality	0.474 (0.194)	1.077 (0.740)	33
Trust	0.488 (0.184)	1.168 (0.699)	28
None (Control)	0.427 (0.197)	0.957 (0.679)	31

F1 Model. The F1 bias detection model (Appendix Table 2) explained 36% of the variance overall (conditional $R^2 = .36$, marginal $R^2 = .13$). Which indicator a participant saw explained a significant amount of additional variance in the F1 scores beyond predictors such as age, gender and education ($\chi^2(6) = 12.75$, $p = .047$). The strongest improvements were observed for *Bias Highlights* ($\beta = 0.33$, $t(195) = 2.07$, $p = .040$) and *Bias Gauge* ($\beta = 0.34$, $t(197) = 2.18$, $p = .031$), indicating that direct examples of biased language and

[5] Shifting the low/medium threshold by ±3% changed the category assignment of 1 statement (7% threshold); shifting the medium/high threshold by ±5% changed 3 statements (35% threshold). Across all shifts, 11 of 15 statements retained their original category.

[6] RStudio 2023.06.0+421 with lmerTest, lme4, psych, irr. LMMs fit using the lme4 package with p-values and df calculated via Satterthwaite approximation. Model comparisons: likelihood ratio tests with chi-square distribution.

[7] osf.io/ekd6s/?view_only=e5dd63a6342448a5befd3c53fa3496a0

[8] https://www.census.gov/

contextualized bias amounts supported later detection better than the control condition. The raw means of *Bias Gauge* and control in Table 2 corresponded to a model-adjusted gain of +0.067 F1 unit points (95% CI [0.01, 0.13]), indicating a medium standardized effect ($\beta = 0.34$). *Bias Highlights* produced a comparable adjusted gain of +0.067 F1 unit points (95% CI [0.01, 0.12], $\beta = 0.33$). Both improvements were driven almost entirely by recall: *Bias Gauge* raised recall from 0.25 to 0.38 while precision remained almost unchanged from 0.77 to 0.76. *Bias Highlights* raised both recall from 0.25 to 0.37 and precision from 0.77 to 0.82, indicating the additional hits were accompanied by proportionally less false alarms. This observation points to a sensitivity gain, in line with a significant d' effect below. The confidence intervals for both indicators are relatively wide, suggesting the true difference to the control group may range from small to moderate. Despite *Bias Bar* using the same data type as *Bias Gauge*, it showed no statistically reliable evidence of an effect. Neither did *Emotionality*, *Political*, nor *Trust*. Political factors showed small to large effects, detailed in Section 4.5.

The F1 model showed a positive, small to medium effect between F1 scores and perceived bias. Sentiment in the statement had a small positive effect and bias in the statement had a small negative effect (Appendix Table 2). This means more sentiment in the statement and participants perceiving more bias lead to higher F1 scores. However, more bias in the statement decreased F1 scores. Factors such as education and age did not impact detection accuracy. Gender showed a non-significant, negative effect with men achieving slightly lower F1 scores than other genders (Appendix Table 2).

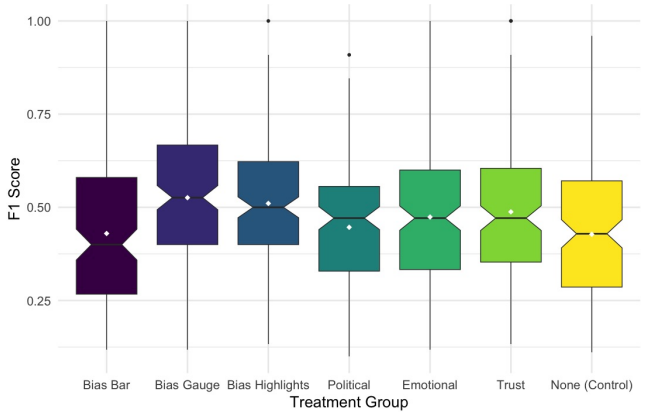

Figure 3: Boxplot with each group's F1 score measuring the degree to which participants detected bias compared to the expert standard. Higher scores indicate higher bias detection accuracy.

d' Model. The pattern for d' was similar. The highest mean d' values occurred in the *Bias Highlights* and *Bias Gauge* groups. The model for d' scores explained 42% of the variance (conditional $R^2 = .42$), with 20% attributable to the fixed effects (marginal $R^2 = .20$). Because d' is expressed in standard-deviation units, the adjusted gains can be read directly as sensitivity improvements: *Bias Highlights* increased d' by 0.25 (95% CI [0.05, 0.46], $\beta = 0.35$), about a quarter of a standard deviation over the control group. This suggests that the indicator group not only improved regarding annotation overlap with the expert standard but also sensitivity in distinguishing biased from unbiased words. Additionally, the *Trust* indicator showed a smaller, non-significant positive contrast of +0.19 d' ($p = .055$) which we interpret as a tendency rather than an effect. Neither the *Bias Gauge*, *Emotionality*, *Bias Bar*, nor *Political* indicators showed significant effects.

Similar to the F1 model, there was a positive relationship with a medium to large effect between perceived bias and d' scores. Again, higher bias in the statement had a strong negative relationship with d' scores, with a small to medium effect (Appendix Table 3). A non-significant, negative relationship was observed between age and d' score, aligning with existing literature [21], with each additional year resulting in a slight decrease in d' scores. Gender, specifically identifying as male, had a non-significant negative effect.

Importantly, the Gauge pattern differed across metrics: it improved F1 but not d'. This suggests that the gauge may have calibrated the amount of bias participants marked, improving overlap with the expert annotations without clearly improving sensitivity in distinguishing biased from unbiased words. The interpretation matches the observed F1 increase, driven by higher recall with stable precision. However, the limited or null effects for *Trust*, *Political*, and *Emotionality* should be interpreted cautiously given the study's slight underpowering. Together, the two models indicate that indicators were most effective when they either showed biased words directly or placed bias amount into an interpretable context.

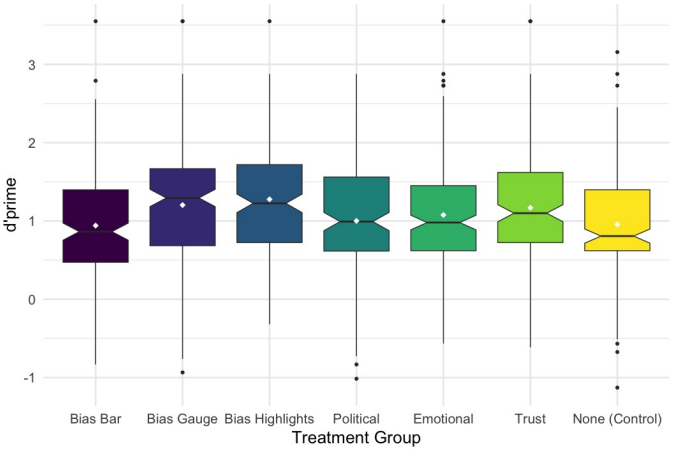

Figure 4: Boxplot with each group's d' score measuring the degree to which participants discern bias compared to the expert standard. Higher scores indicate higher ability to discern bias.

4.3 Bias Perception (RQ2)

We next examined whether indicators influenced self-reported bias perception. Participants' perceived bias increased with the ground-truth bias level of the statements, confirming that the selected material with its different bias levels was appropriate. However, the effect of indicators on perceived bias was weaker and less consistent than their effect on detection (Appendix Table 4).

The *Bias Highlights* group showed a significantly higher perception of bias, a small effect significantly greater than the *Control* group ($\beta = 0.19$, CI$[0.07, 0.74]$, $p = .024$). Conversely, the *Bias Bar* group recorded a significantly lower perception of bias across all statements, a small to medium effect ($\beta = -0.25$, CI$[-0.88, -0.18]$, $p = .004$). We discuss this further in Section 5.1. Apart from *Bias Highlights* and *Bias Bar*, no other group showed significant effects. Overall, a higher perception of bias was associated with more accurate detection of bias, as evident in the *Bias Gauge* group and visible in Appendix Figure 3 and Figure 5. In Figure 6, the *Bias Gauge* group shows the highest decrease in d' scores when perceived bias is low, yet records the highest d' scores when perceived bias is high. This suggests that when participants perceive less bias, their detection ability decreases. Gender showed a small but significant effect, with men perceiving slightly less bias than other genders.

4.4 Bias, Sentiment, Trust, and Sharing (RQ3)

To examine how perceived bias relates to other judgments, we computed correlations between bias perception, perceived emotionality, trust, and sharing intention. We found that most factors were moderately to strongly correlated (Table 3). The strongest relationship was between perceived emotionality and perceived bias ($\rho = .71$, $p < .001$), indicating that emotionally charged language was often interpreted as biased. This suggests that, in practice, readers might not separate linguistic bias from emotional tone, even though the two constructs are conceptually distinct. Participants in the *Emotionality* group rated statements as more emotionally charged ($M = 3.52$, $SD = 1.96$), whereas those in the *Bias*

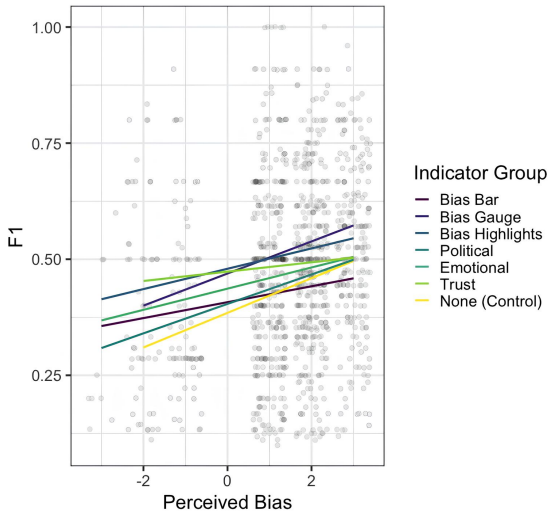

Figure 5: Perceived bias (question item rating bias in statement) vs. detected bias (F1 score) by indicator group. Higher F1 scores indicate higher bias detection accuracy.

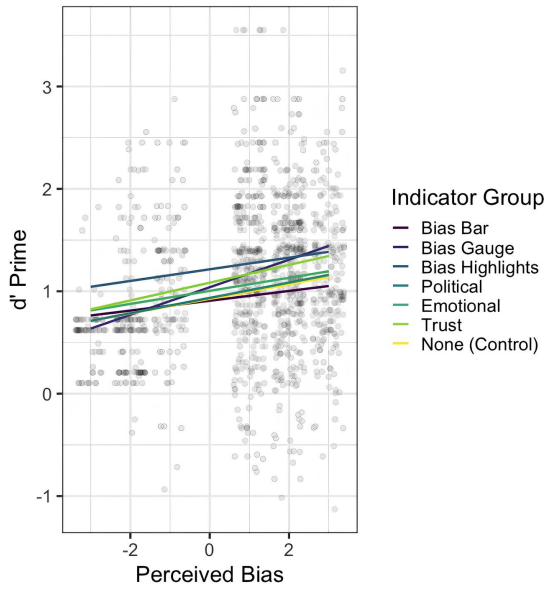

Figure 6: Perceived bias (question item rating bias in statement) vs. detected bias (d' score) by indicator group. Higher d' scores indicate higher bias discernment abilities.

Bar group perceived them as least emotional ($M = 2.88$, $SD = 1.73$). Similarly, the F1 detection model showed a large effect of ground truth sentiment which the d' and perception models did not mirror.

Bias perception was also related to trust and sharing intention. Statements perceived as more biased tended to be judged as less trustworthy ($\rho = -.61$, $p < .001$) and less suitable for sharing ($\rho = -.24$, $p < .001$). Trust was positively correlated with sharing ($\rho = .38$, $p < .001$). Political congruence was the strongest predictor of higher trust, and low-bias statements were trusted more than high-bias statements. For example, liberal participants reported higher trust for liberal low-bias statements ($M = 5.04$) compared to liberal high-bias statements ($M = 4.12$). Across experimental groups, the *Trust* group reported the lowest average trust ($M = 3.80$, $SD = 1.48$), while the *Political* group reported the highest trust ($M = 4.37$, $SD = 1.43$).

Sharing, transformed from the 7-point scale to -3 to 3, was lowest for highly biased statements (e.g., right-leaning, high-bias, $M = -2.51$, $SD = 1.04$) and highest for low-bias statements (e.g., centrist, low-bias, $M = -1.46$, $SD = 1.71$). Participants in the *Trust* group reported the lowest sharing intentions ($M = -2.21$, $SD = 1.35$), while the *Bias Highlights* ($M = -1.63$, $SD = 1.56$) and the *Control* ($M = -1.67$, $SD = 1.48$) groups report the highest. The relationship between trust and sharing decisions connects to findings from misinformation research, where they are also correlated [10] and perceptions are shaped by quick, automatic judgments [45].

Table 3: Correlation of bias perception, sentiment perception, and willingness to share with statement bias, statement sentiment, and participants' trust (indication phase with indicator).

	Sentiment	Sharing	Trust
Bias Perception	.71***	-.24***	-.61***
Sentiment		-.14***	-.48***
Sharing			.38***

Note. '***': ≤ .001, '**': ≤ .01, '*': ≤ .05, '.': ≤ .1, ' ': > .1

4.5 Political Congruency (RQ4)

Political leaning and congruency were strong factors influencing bias perception and detection, often stronger than the impact of an indicator. In the F1 bias detection model (Appendix Table 2), participants' congruence with the statements' political leaning influence on F1 scores showed a medium effect ($\beta = 0.34$, CI$[0.02, 0.12]$, $p = .009$). The d' detection model mirrored these findings (Appendix Table 3). Here, congruency showed a significant small to medium effect ($\beta = 0.25$, $t(1046) = 2.31$, CI$[0.03, 0.33]$, $p = .021$). However, since congruency was coded across three levels (liberal, moderate, conservative), we focus our analysis on the interaction effects of liberal and conservative congruency, as these represent the most theoretically distinct positions.

While liberal participants achieved significantly higher F1 scores ($\beta = 0.34$, $p < .011$), the F1 model shows a large negative interaction effect between liberal participants and liberal statements ($\beta = -0.73$, $p < .001$), substantially larger than the non-significant conservative–congruency interaction ($\beta = -0.17$, $p = .411$). The asymmetry was not visible in the d' model. Both liberals ($\beta = -0.39$, CI$[-0.53, -0.05]$, $p = .02$) and conservatives ($\beta = -0.38$, CI$[-0.52, -0.03]$, $p = .027$) showed significant medium negative interactions. The d' model indicates that for conservative and more so for liberal participants, bias detection accuracy significantly decreased when statements align with their political beliefs. The perception model did not find general congruency effects (Appendix Table 4). However, liberal participants perceived significantly more bias in congruent statements compared to the baseline of moderates' congruency interaction ($\beta = 0.2$, CI$[0.01, 0.40]$, $p = .037$), diverging from the detection models. Concluding, while some visual indicators affected perception, they did not override partisan bias and congruency effects. Even effective indicators operate within a broader interpretive context shaped by prior beliefs, highlighting that user background is one of the main factors impacting bias perception and effectiveness of interventions.

5 Discussion

This study examined how different visual indicators support the detection and perception of linguistic bias in short news statements and which other factors, such as political congruency, sentiment, and credibility, influence bias perception. The results show that indicators that support pattern learning through examples and direct cue localization [24] (*Bias Highlights*) or interpretable reference frames [14, 64] (*Bias Gauge*), improve linguistic bias detection more than abstract indicators. However, outcomes are more strongly influenced by political congruence.

The study further indicated a close connection between detected bias, perceived bias, actual bias in the statement, trust, sharing discernment, and perceived emotionality. Their connection suggests that linguistic bias is embedded in a broader judgment process. Readers do not evaluate bias in isolation. Instead, they might connect it to emotionality,

political stances, and credibility. Thus, representation choices may shape multiple downstream judgments beyond the main bias task.

We have also seen differences between bias perception and bias detection. For example, the *Bias Bar* indicator did not impact detection but impacted perception. The difference matters for future tool designs. While detection and perception are related, they are not the same outcome. An indicator may help readers identify biased words without fully changing how biased they say a statement feels. For this reason, evaluating bias-mitigation tools requires both subjective perception measures and more objective test tasks.

Often, visual systems are evaluated as if information display alone can correct biased interpretation. While our findings show that indicators can help readers notice one important layer of media bias, they do not address the broader problem of bias on their own. Linguistic bias indicators can be helpful but not sufficient. They can support awareness, yet operate inside social and cognitive contexts that influence what users are willing to notice or accept. Interpretation and impact on decision making remain strongly influenced by one's demographics, political worldview, the framing of the content, and prior information exposure. Our work positions bias indicators not as standalone correctness tools, but as visual supports within a larger sensemaking process. Political worldview, emotional language, and trust judgments continue to shape how news is interpreted, even when visual support is provided. From a visualization perspective, the question is therefore not only whether an indicator changes a judgment, but how it supports users in forming more reflective and better grounded judgments under real-world cognitive and political conditions.

5.1 Indicator Effects on Bias Metrics (RQ1, RQ2)

In-text encodings that externalized the bias signal at the phrase level or translated quantitative values in an interpretable reference frame tended to raise bias detection more than abstract bias, trust, or sentiment summary encodings. *Bias Highlights* and, partly, *Bias Gauge* produced the strongest improvements in word-level bias detection, whereas more abstract summary indicators with cues related to political leaning, sentiments and trust showed no or limited effects (RQ1, RQ2). Information labels like *Political* and *Emotionality*, commonly used by news aggregators [4], showed no significant effects on bias perception or detection. While political classifications may be informative for curious readers, our study and prior work suggest that their value for increasing bias awareness is limited [49, 58]. *Trust* showed only minimal positive effects, and because the detection analyses were slightly underpowered, these findings should not be taken to mean that emotionality, political, or trust cues are ineffective in general. However, abstract summary scores, a derived data type, require an additional mental step to map back to source content [48]. The mapping might be more fragile when prior beliefs are strong. Overall, the findings are in line with previous research and suggest that indicators are most effective when they highlight concrete biased wording in-text [61] or present overall bias with reference frames in an interpretable form [14, 64].

The two most effective indicators, *Bias Highlights* and *Bias Gauge*, rely on different mechanisms of support. *Bias Highlights* likely help because they make the biased phrases explicit in-text. Despite gray highlights being a lower-salience choice [61], the effect suggests the mechanism is example-based learning [24] rather than perceptual salience alone. *Bias Gauge* may partially work because it provides a contextualized interpretation of bias amount which allows quick assessment through heuristics [55, 63] linked to the example text. In contrast, *Bias Bar* may convey the same underlying information but be harder to interpret because it lacks a clear reference frame, consistent with previous work [14, 53, 64]. This may even reverse its effect on bias perception (Section 4.3), as bars filled to only 30-50% may anchor judgments around a summary score that readers interpret as 'not biased' when no explicit categories are shown. The reference frame best suited for linguistic bias should be further investigated.

For *Bias Gauge*, the significant F1 effect without a corresponding d' effect or perception effect suggests that it influenced how much bias participants labeled rather than how accurately they identified it. The gauge may have served as a calibration cue that helped participants match the expected amount of bias rather than improving their ability to detect biased wording like *Bias Highlights*. The difference in cognitive support might support this, with *Bias Highlights* providing explicit location of evidence, whereas *Bias Gauge* operates at statement level, providing aggregate magnitude with context.

The differing effects observed in the bias perception model, compared to the detection models, suggest that noticing biased words and reporting a statement as biased are related but distinct processes. Detection can improve through exposure and learning, while explicit judgment remains more closely tied to prior attitudes and broader evaluative impressions. This reinforces the importance of evaluating visual interventions with both test and subjective perception measures rather than relying on a single measure alone.

5.2 Perceived Emotionality, Trust, and Sharing (RQ3)

The strong correlation between perceived emotionality and perceived bias suggests limited differentiation between emotionally charged language and linguistic bias. Dual-process theory could explain this conflation, as fast, affect-driven processing registers emotional tone as a signal of bias, while the slower reasoning needed to disentangle the two is rarely used during routine news reading [44]. Accordingly, participants in the *Emotional* condition perceived statements as most emotionally charged, yet showed no gains in detection accuracy, suggesting that emotional salience amplified the conflation rather than resolving it. This implies that indicators showing emotional content may not reliably improve bias perception and detection.

The positive trust–sharing correlation replicates a documented pattern in misinformation research [41]. Notably, participants in the *Trust* condition reported the lowest trust and sharing intentions of any group. The low trust suggests a backfire effect where explicit credibility cues might have triggered generalized skepticism rather than active assessment of bias [28]. Designers might want to treat such indicators with caution. Trust-based indicators may reduce engagement with legitimate content rather than improving critical evaluation of biased material [4].

5.3 Political and Demographic Factors (RQ4)

Political leaning and political congruence were among the strongest factors across our models. Participants from the left or right tended to overlook bias in politically congruent statements but found them more trustworthy, regardless of their bias content. This tendency remained visible even when indicators were present, meaning that visual support did not neutralize partisan bias effects. Instead, indicator effects might have been filtered through readers' existing orientations.

While Garrett and Bond [19] document greater conservative susceptibility to politically congruent but misleading information and Spinde et al. [58] reported stronger congruency effects regarding bias perception for conservatives, our results show a more nuanced pattern. Liberals perceived congruent statements as significantly more biased, yet their word-level detection accuracy dropped in one of the two models, suggesting they recognized bias as a general quality of the text but struggled to localize it at the word level. We hypothesize that liberals may have a heightened general awareness of bias in congruent content, but this awareness does not translate into precise identification of biased language. Familiar, ideologically congruent phrasing may be harder for liberals to flag at the word level because it reads as natural, even when they consciously sense the text is biased overall. Conservatives, by contrast, showed no equivalent perception shift, suggesting that for them, congruency operates more automatically, reducing detection without triggering explicit bias awareness. However, as our sample contained more liberal participants, we cannot draw clear conclusions for conservatives. Customized media literacy messages depending on the participant's political leaning and topic opinion from prior work [66] may be a strategy worth testing for media bias. While demographic variables were less influential than political factors, their inclusion remains important as bias judgments are not made in a vacuum and should not be modeled as if users were interchangeable.

5.4 Design Implications

Prior work on bias and credibility indicators clusters around two principles: make evidence concrete and localized rather than abstract [61], and pair quantitative signals with interpretable reference frames [14,64]. Our results align with these principles but add empirical weight for linguistic bias (Table 4). *Bias Highlights* confirms that in-text localization outperforms summary-level encodings for transfer-based detection, consistent with earlier work on highlights [6, 56, 58] as pre-attentive cues [61]. *Bias Gauge* confirms that categorical segmentation helps magnitude interpretation beyond raw bar encodings [14, 64], though its effect appeared in F1 but not d, suggesting calibration rather than genuine discrimination gains. However, effects of icon-based credibility cues [33, 63] do not clearly carry over to bias detection. *Trust* showed only a marginal positive tendency, and its low sharing and trust scores across participants hint at backfire effects [28]. We draw three design implications for bias mitigation tools.

(1) Provide contextual learning opportunities. Indicators that show examples or interpretable context outperform those showing only summary metrics. Designers should favor cues that help users understand *why* content may be biased, not only *how much*. Concretely, highlights, annotated examples, and interpretable scales are more promising than isolated scores or icons.

(2) Design for political congruence effects. Politically aligned content is consistently rated as less biased across the spectrum, so indicator designs must account for this asymmetry. Combining wording cues with comparative context (alternative viewpoints, source diversity, framing contrasts) and adapting indicators per target group [66] are two promising directions. Systems supporting easy cross-perspective comparison may further reduce congruence-driven blind spots.

(3) Use bias cues, not sentiment cues. Despite a strong correlation between perceived bias and perceived emotionality, the *Emotionality* indicator did not directly impact bias perception. Tools should not treat emotional language as a proxy for bias and instead make the distinction between emotional tone and biased wording explicit and visible.

Linguistic cues address only one part of media bias. We suggest that future tools should connect word-level bias indicators with source-level, framing-level, and cross-article context. This layered approach would move beyond single labels toward richer visual analytic support for news interpretation. Lastly, we argue that designers should treat bias indication as a reflective aid, not as an automated correction mechanism. The goal is not to replace the reader's own judgment, but to improve the conditions under which the judgment is formed.

5.5 Limitations

This study has several limitations. We focused on short statements rather than full, multi-modal articles. This choice improved experimental control and reflected short-form news exposure, but it does not capture all forms of media bias, especially those that emerge through structure, omission, or article-level framing. Material topic choices might also impact results. Because transfer was tested jointly with a material topic change, our results estimate transfer under changed material conditions rather than pure same-item retention. Findings only generalize to first-exposure effects of indicator type on our material format, but not to full articles, non-English contexts, or long-term use.

Researcher subjectivity, political stances, and Western-centric biases remain limitations, as no method can establish a fully objective gold standard for bias. Although we achieved higher inter-annotator agreement than prior work on word-level bias annotation [60], the task remains challenging and inherently subjective. Consequently, any ground truth is an approximation affecting evaluation. The study tested immediate effects in a controlled setting and therefore cannot show how indicator use develops over longer periods of real-world news consumption. The sample was somewhat more liberal and more educated than the general U.S. population. Further, the study was slightly underpowered with small differences in group size. We decided against resampling because news content can become outdated quickly, which could influence the results.

Including indicators based on bias, political leaning, sentiments, or trust reflects real-world practice, but it also complicates interpretation because they were originally developed for different goals. Our indicators covered a range of commonly used constructs but differed in their visual encoding, data type, abstraction level, and the cognitive support they provide, e.g., *Bias Highlights* exposed biased words in-text, while other indicators summarize bias at the statement level. This limited the analysis of individual design features and may not be representative for the whole design space of bias indicators. Effects attributed to an indicator reflect the combined influence of its data type, visual encoding, cognitive support, and nature of the cue type or strategy rather than any single visual encoding. While we standardized the designs for comparability, the differences may still have influenced participant responses. We therefore frame this study as exploratory rather than focused on individual encoding effects.

Table 4: Predictors and their effects on detection (F1, d') and perception (P.). ↑ positive, ↓ negative, ( ) non-significant effect.

Predictor	F1	d'	P.	Result / Implication
Bias Bar	–	–	↓	Lowered bias perception significantly. Avoid designs without context frames.
Bias Gauge	↑	–	–	Supported task, may provide reference frames for what is high/low bias.
Bias Highlights	↑	↑	↑	Best supported task, examples might increase example-based learning.
Political Classification	–	–	–	While often used online, we found no effects.
Emotionality	–	–	–	While often used online, we found no effects.
Trust Indicator	–	(↑)	–	Non-significant effect in the d' model. Needs further research.
Higher Perceived Bias	↑	↑	–	Perceiving more bias was connected to detecting it more accurately.
Higher Bias in Statement	↓	↓	↑	On statements with higher bias, detection accuracy decreased, but bias perception increased.
More Sentiment in Statement	↑	–	↑	Only in F1, not in d'. Needs further research.
Liberals	↑	–	–	Liberals detected bias more accurately, only significant in F1 LMM.
Congruency Liberals	↓	↓	↑	Decreased detection but increased perception of bias in congruent statements.
Congruency Conservatives	–	↓	–	Decreased detection in d' model. Less represented in samples, more research needed.
Gender (Male)	(↓)	(↓)	↓	Men detected and perceived less bias.
Higher Age	–	(↓)	–	Higher age decreased bias detection in the d' model (non-significant).

6 Conclusion

We compared six visual indicators for linguistic bias and found that indicators providing direct examples or interpretable context, especially *Bias Highlights* and *Bias Gauge*, best supported bias detection. Our results extend prior findings by showing that in-text localization and categorical reference frames improve transfer-based detection of linguistic bias while still interacting strongly with political congruence and perceived emotionality. Political congruence remained a strong influence on both detection and perception, showing that visual indicators can support reflection but cannot solely remove broader interpretive biases. Our study position linguistic bias indicators as one useful component of a larger bias mitigation strategy, highlighting the value of designing visual tools that support learning, comparison, and reflection rather than relying on summary scores or binary judgments.

Supplemental Materials

All supplemental materials are available on Zenodo, DOI 10.5281/zenodo.19347505, released under a CC BY 4.0 license. They include the R code, the study material, study screenshots, all indicator graphics, and the literature review table.

Figure Credits and Copyrights

All figures remain under the authors' own copyright, with the permission to be used here. We also share them under a CC BY 4.0 license.

Acknowledgments

This work was supported by the Hanns-Seidel Foundation, the German Federal Ministry of Education and Research (BMBF) through the DAAD (German Academic Exchange Service), the Bavarian State Ministry for Digital Affairs in the project "XR Hub" (Grant A5-3822-2-16), and partially supported by JSPS KAKENHI Grant JP24H00732, by JST CREST Grants JPMJCR20D3 and JPMJCR2562 including AIP challenge program, and by JST K Program Grant JPMJKP24C2 Japan.

References

[1] J. An, M. Cha, K. Gummadi, J. Crowcroft, and D. Quercia. Visualizing media bias through twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(2):2–5, 2021. doi: 10.1609/icwsm.v6i2.14343 2, 4

[2] A. Appelman and S. S. Sundar. Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, 93(1):59–79, 2016. doi: 10.1177/1077699015606057 4

[3] A. Ardèvol-Abreu and H. G. de Zúñiga. Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news. *Journalism & Mass Communication Quarterly*, 94(3):703–724, 2017. doi: 10.1177/1077699016654684 1

[4] K. Aslett, A. M. Guess, R. Bonneau, J. Nagler, and Joshua A. Tucker. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18):eabl3844, 2022. doi: 10.1126/sciadv.abl3844 2, 4, 8

[5] C. Batailler, S. M. Brannon, P. E. Teas, and B. Gawronski. A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science*, 17(1):78–98, 2022. doi: 10.1177/1745691620986135 4

[6] E. P. S. Baumer, F. Polletta, N. Pierski, and G. K. Gay. A simple intervention to reduce framing effects in perceptions of global climate change. *Environmental Communication*, 11(3):289–310, 2015. doi: 10.1080/17524032.2015.1084015 1, 2, 4, 9

[7] M. M. Bhuiyan, M. Horning, S. W. Lee, and T. Mitra. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021. doi: 10.1145/3479571 2

[8] A. Bodaghi, K. A. Schmitt, P. Watine, and B. C. M. Fung. A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems*, 11(4):5119–5145, Aug. 2024. doi: 10.1109/tcss.2023.3289031 1, 2

[9] C. Budak, S. Goel, and J. M. Rao. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, Apr. 2016. doi: 10.1093/poq/nfw007 1

[10] M. Chan, C. Vaccari, and M. Yamamoto. Cognitive drivers of misinformation belief and sharing on social media: A cross-national comparison. *Mass Communication and Society*, 29(1):142–160, 2026. doi: 10.1080/15205436.2025.2577294 7

[11] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095, 2020. doi: 10.1007/s11109-019-09533-0 4

[12] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc.*, 79(387):531–554, Sept. 1984. doi: 10.2307/2288400 2, 3

[13] Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, and D. Rand. The social media context interferes with truth discernment. *Science advances*, 9(9):eabo6169, 2023. doi: 10.1126/sciadv.abo6169 4

[14] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3):110–161, Dec. 2021. doi: 10.1177/15291006211051956 2, 3, 7, 8, 9

[15] N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Jarvelin et al. An Information Nutritional Label for Online Documents. *ACM SIGIR Forum*, 51(3):46–66, 2018. doi: 10.1145/3190580.3190588 4

[16] M. Färber, V. Burkard, A. Jatowt, and S. Lim. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3007–3014. ACM, 2020. doi: 10.1145/3340531.3412876 5

[17] Gallup. Party affiliation. https://news.gallup.com/poll/15370/party-affiliation.aspx, 2024. 5

[18] M. Gao, H. J. Do, and W.-T. Fu. An intelligent interface for organizing online opinions on controversial topics. In *Proceedings of the 22nd international conference on intelligent user interfaces*, IUI '17, pp. 119–123. Association for Computing Machinery, 2017. doi: 10.1145/3025171.3025230 4

[19] R. K. Garrett and Robert M. Bond. Conservatives' susceptibility to political misperceptions. *Science Advances*, 7(23):eabf1234, 2021. doi: 10.1126/sciadv.abf1234 8

[20] Google Cloud. Natural language basics. https://cloud.google.com/natural-language/docs/basics?hl=en, 2022. 4

[21] A. Guess, J. Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1):eaau4586, 2019. doi: 10.1126/sciadv.aau4586 6

[22] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler et al. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, 117(27):15536–15545, 2020. doi: 10.1073/pnas.1920498117 2

[23] C. Harris. Searching for Diverse Perspectives in News Articles: Using an LSTM Network to Classify Sentiment. In *ESIDA '18*, 2018. 4

[24] C. G. Healey and J. T. Enns. Attention and visual memory in visualization and computer graphics. *IEEE transactions on visualization and computer graphics*, 18(7):1170–1188, 2012. doi: 10.1109/TVCG.2011.127 2, 7, 8

[25] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 203–212. Association for Computing Machinery, 2010. doi: 10.1145/1753326.1753357 2, 3

[26] S. Hinterreiter, T. Spinde, S. Oberdörfer, I. Echizen, and M. E. Latoschik. News Ninja: Gamified Annotation of Linguistic Bias in Online News. *Proc. ACM Hum.-Comput. Interact.*, 8(CHI PLAY), 2024. doi: 10.1145/3677092 4

[27] S. Hinterreiter, M. Wessel, F. Schliski, I. Echizen, M. E. Latoschik, and T. Spinde. NewsUnfold: Creating a News-Reading Application That Indicates Linguistic Media Bias and Collects Feedback. In *Proceedings of the international AAAI conference on web and social media (ICWSM'25)*, vol. 19. AAAI, June 2025. 2

[28] E. Hoes, B. Aitken, J. Zhang, T. Gackowski, and M. Wojcieszak. Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 8(8):1545–1553, June 2024. doi: 10.1038/s41562-024-01884-x 8, 9

[29] C. Hube and B. Fetahu. Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 195–203. ACM, 2019. doi: 10.1145/3289600.3291018 1

[30] A. Kause, T. Townsend, and W. Gaissmaier. Framing climate uncertainty: Frame choices reveal and influence climate change beliefs. *Weather, Climate, and Society*, 11(1):199–215, 2019. doi: 10.1175/WCAS-D-18-0002.1 1

[31] M. P. Kenning, R. Kelly, and S. L. Jones. Supporting Credibility Assessment of News in Social Media using Star Ratings and Alternate Sources. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2018. doi: 10.1145/3170427.3188489 4

[32] P. Kherwa, A. Sachdeva, D. Mahajan, N. Pande, and P. K. Singh. An approach towards comprehensive sentimental data analysis and opinion mining. In *2014 IEEE International Advance Computing Conference (IACC)*, pp. 606–612, 2014. doi: 10.1109/IAdCC.2014.6779394 4

[33] A. Kim and A. R. Dennis. Says who? the effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, 43(3):1025–

1040, 2019. doi: 10.25300/MISQ/2019/15188 2, 4, 9

[34] A. Kim, P. L. Moravec, and A. R. Dennis. Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. *Journal of Management Information Systems*, 36(3):931–968, 2019. doi: 10.1080/07421222.2019.1628921 2, 4

[35] J. Kirchner and C. Reuter. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), 2020. doi: 10.1145/3415211 2

[36] N. Lee, Y. Bang, T. Yu, A. Madotto, and P. Fung. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias, 2022. doi: 10.48550/arXiv.2204.04902 1

[37] J. Liedke and J. Gottfried. U.S. adults under 30 now trust information from social media almost as much as from national news outlets, 2022. 2

[38] A. P. C. Lixun Su and M. F. Walsh. Trustworthy blue or untrustworthy red: The influence of colors on trust. *Journal of Marketing Theory and Practice*, 27(3):269–281, 2019. doi: 10.1080/10696679.2019.1616560 2, 4

[39] M. Luo, J. T. Hancock, and D. M. Markowitz. Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media: Effects of Truth-Bias and Endorsement Cues. *Communication Research*, 49(2):171–195, 2022. doi: 10.1177/0093650220921321 2

[40] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, Apr. 1986. doi: 10.1145/22949.22950 2, 3

[41] P. Majerczak and A. Strzelecki. Trust, media credibility, social ties, and the intention to share towards information verification in an age of fake news. *Behavioral Sciences*, 12(2):51, 2022. doi: 10.3390/bs12020051 8

[42] C. Martel, G. Pennycook, and D. G. Rand. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47, 2020. doi: 10.1186/s41235-020-00252-3 4

[43] H. Martínez Alonso, A. Delamaire, and B. Sagot. Annotating omission in statement pairs. In *Proceedings of the 11th linguistic annotation workshop*, pp. 41–45. Association for Computational Linguistics, Apr. 2017. doi: 10.18653/v1/W17-0805 1

[44] P. Moravec, R. Minas, and A. R. Dennis. Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All, 2018. doi: 10.2139/ssrn.3269541 8

[45] P. L. Moravec, A. Kim, and A. R. Dennis. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, 31(3):987–1006, 2020. doi: 10.1287/isre.2020.0927 2, 4, 7

[46] S. Mullainathan and A. Shleifer. Media bias. Working paper 9295, National Bureau of Economic Research, Oct. 2002. Series: Working paper series. doi: 10.3386/w9295 1

[47] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009. doi: 10.1109/TVCG.2009.111 3

[48] T. Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, Dec. 2014. doi: 10.1201/b17511 2, 8

[49] S. Park, S. Kang, S. Chung, and J. Song. NewsCube: delivering multiple aspects of news to mitigate media bias. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 443–452, 2009. doi: 10.1145/1518701.1518772 2, 8

[50] D. M. W. Powers. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. *arXiv*, 2, 2008. doi: 10.48550/arXiv.2010.16061 4

[51] M. pui Sally Chan, C. R. Jones, K. H. Jamieson, and D. Albarracín. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546, 2017. doi: 10.1177/0956797617714579 2

[52] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1650–1659. Association for Computational Linguistics, Aug. 2013. 1

[53] P. Shah and J. Hoeffner. Review of Graph Comprehension Research: Implications for Instruction. *Educational Psychology Review*, 14(1):47–69, Mar. 2002. doi: 10.1023/A:1013180410169 2, 8

[54] D. Simkin and R. Hastie. An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, 82(398):454–465, June 1987. doi: 10.1080/01621459.1987.10478448 2

[55] C. Speier. The influence of information presentation formats on complex task decision-making performance. *International Journal of Human-Computer Studies*, 64(11):1115–1131, Nov. 2006. doi: 10.1016/j.ijhcs.2006.06.007 8

[56] T. Spinde, F. Hamborg, K. Donnay, A. Becerra, and B. Gipp. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, pp. 389–392. Association for Computing Machinery, 2020. doi: 10.1145/3383583.3398619 1, 2, 9

[57] T. Spinde, S. Hinterreiter, F. Haak, T. Ruas, H. Giese, N. Meuschke et al. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv*, 2023. doi: 10.48550/arXiv.2312.16148 1, 4

[58] T. Spinde, C. Jeggle, M. Haupt, W. Gaissmaier, and H. Giese. How do we raise media bias awareness effectively? Effects of visualizations to communicate bias. *PLOS ONE*, 17(4):1–14, 2022. doi: 10.1371/journal.pone.0266204 1, 2, 4, 8, 9

[59] T. Spinde, C. Kreuter, W. Gaissmaier, F. Hamborg, B. Gipp, and H. Giese. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021. doi: 10.1109/JCDL52503.2021.00018 2, 4

[60] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-emnlp.101 1, 4, 5, 9

[61] H. Strobelt, D. Oelke, B. C. Kwon, T. Schreck, and H. Pfister. Guidelines for Effective Usage of Text Highlighting Techniques. *IEEE transactions on visualization and computer graphics*, 22(1):489–498, Jan. 2016. doi: 10.1109/TVCG.2015.2467759 2, 3, 4, 8, 9

[62] M. Sultan, A. N. Tump, N. Ehmann, P. Lorenz-Spreen, R. Hertwig, A. Gollwitzer et al. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors, 2024. doi: 10.31234/osf.io/7u4fg 1, 2

[63] S. S. Sundar. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital Media, Youth, and Credibility*, (Foundation Series on Digital Media and Learning):73–100, 2008. 2, 8, 9

[64] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. doi: 10.1207/s15516709cog1202_4 2, 3, 7, 8, 9

[65] C. Thornhill, Q. Meeus, J. Peperkamp, and B. Berendt. A Digital Nudge to Counter Confirmation Bias. *Frontiers in Big Data*, 2, 2019. doi: 10.3389/fdata.2019.00011 2

[66] T. G. van der Meer and M. Hameleers. Fighting biased news diets: Using news media literacy interventions to stimulate online cross-cutting media exposure patterns. *New Media & Society*, 23(11):3156–3178, 2021. doi: 10.1177/1461444820946455 8, 9

[67] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann Publishers Inc., 2004. doi: 10/d3gg9c 1, 2, 3

[68] M. Wessel, T. Horych, T. Ruas, A. Aizawa, B. Gipp, and T. Spinde. Introducing MBIB - the first media bias identification benchmark task and dataset collection. In *Proceedings of 46th international ACM SIGIR conference on research and development in information retrieval*. ACM, 2023. doi: 10.1145/3539618.3591882 1