Spelling errors and 'shouting' capitalization lead to additive penalties to trustworthiness of online health information: randomized experiment with laypersons

Witchel, Harry J, Thompson, Georgina A, Jones, Christopher I, Westling, Carina E I, Romero, Juan, Nicotra, Alessia, Maag, Bruno and Critchley, Hugo D (2020) Spelling errors and 'shouting' capitalization lead to additive penalties to trustworthiness of online health information: randomized experiment with laypersons. Journal of Medical Internet Research. ISSN 1439-4456 (Accepted)

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/89854/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
Title:

Spelling errors and 'shouting' capitalization lead to additive penalties to trustworthiness of online health information:

Randomized Experiment with Laypersons

Authors

Harry J Witchel, PhD 1,*, Georgina A Thompson, BMBS 1, Christopher I. Jones, PhD, MRSB CStat, CSci 1, Carina E I Westling, PhD 2,†, Juan Romero 3, Alessia Nicotra, PhD 3, Bruno Maag 3, Hugo D Critchley, MBChB, DPhil, FRCPsych 1

Addresses

1 Brighton and Sussex Medical School, Brighton BN1 9RY, UK
2 School of Media, Film and Music, University of Sussex, Brighton BN1 9RH, UK
3 Dalton Maag Ltd., London SW9 9SP, UK

Georgina Thompson ORCID: 0000-0001-6777-5730
Carina Westling ORCID: 0000-0002-6549-0974
Alessia Nicotra ORCID: 0000-0001-7775-0166
Bruno Maag ORCID: 0000-0002-2841-8439
Hugo Critchley ORCID: 0000-0002-2445-9284
Juan Romero ORCID: 0000-0001-5972-4001
Christopher Jones ORCID: 0000-0001-7065-1157

* Author for correspondence: h.witchel@bsms.ac.uk, ORCID 0000-0001-8404-3494
† Current address: Faculty of Media and Communication, Bournemouth University, Bournemouth BH12 5BB, UK

Short Title

"Trustworthiness, spelling and capitalization"

Keywords

trust, trustworthiness, credibility, incivility, information quality, grammatical errors, writing mechanics, capitalisation
Abstract

BACKGROUND. The written format and literacy competence of screen-based texts can interfere with the perceived trustworthiness of health information in online forums, independent of the semantic content. Unlike in professional content, the format in unmoderated forums can regularly hint at 'incivility', perceived as deliberate rudeness or casual disregard towards the reader, e.g. through spelling errors and unnecessary emphatic capitalization of whole words (online 'shouting').

OBJECTIVE. To quantify the comparative effects of spelling errors and inappropriate capitalization on ratings of trustworthiness independently of lay insight, and to determine whether these changes act either synergistically or additively on the ratings.

METHODS. In online experiments, 301 UK-recruited participants rated thirty-six randomised short stimulus excerpts (in the format of information from an unmoderated health forum about multiple sclerosis) for trustworthiness using a semantic differential slider. Nine control excerpts were compared to error-containing excerpts including 5 instances of misspelling, 5 instances of inappropriate capitalization ('shouting'), or a combination of the two. Data were analysed in a Linear Mixed Effects model.

RESULTS. The mean trustworthiness ratings of the control excerpts ranged from 32.59 to 62.31 (rating scale 0-100). Compared to the control excerpts, excerpts containing only misspellings were rated as being 8.86 points less trustworthy, inappropriate capitalization was 6.41 less, and the combination of misspelling and capitalization was 14.33 less. Misspelling and inappropriate capitalization show an additive effect ($P < .05$ for all).

CONCLUSIONS. Distinct indicators of incivility independently and additively penalize perceived trustworthiness of online text independently of lay insight, eliciting a medium effect size.
Introduction

Trustworthiness of online health information: background, context and importance

As of 2019, 90% of all British adults use the internet at least weekly [1], and as patients they often search online for health information in order to solve their medical problems [2]; furthermore, they are likely to be influenced by the online information, including changing their healthcare decisions and their frequency of ambulatory care visits [3]. In a survey of 2,200 adults with chronic health conditions in the USA who were active social media users, 57% used a health-condition specific website (e.g. specialising in multiple sclerosis or rheumatoid arthritis) on a monthly basis, and 5% used such sites daily; half of the patients surveyed had asked a health-related question to others online within the previous 6 months, and 87% of those were seeking responses from other patients with the health condition [4].

In response to this explosion of unvetted potential sources providing online healthcare information that is acted upon, researchers, experts and medical professionals have repeatedly expressed concern about the inaccuracy of online information and the limited ability of lay consumers to adequately assess its validity [2, 5, 6]. In particular, it has long been known that when lay users determine whether to use and trust online healthcare information, they are strongly influenced by non-medical criteria that experts do not use [7-11]. Broadly, while academics favour a checklist approach of transparency criteria [12], the approach of non-experts appears to be more variable, and situation-dependent, and it prioritises factors such as ease of understanding and the attractiveness of graphic design [9]. More generally, in research on the factors that influence judgments of trust, all aspects of trustworthiness can be relational and depend on the type of person (or group being studied); major relational factors include accessibility, both cognitive as well as physical [13], and correctly accommodating language to the intended audience [14].

Understanding lay assessment of trustworthiness of online health information is important because false online information presented to the general public, if believed, has the potential to undermine correct medical advice [15], to elicit unhealthy behaviour [16] as well as influence socio-political discourses on healthcare and other topics [17, 18, 19]. Sbaffi and Rowley's [20] recent review of how lay people assess the trustworthiness and credibility of online health information concluded that thus far much less research has been performed to understand what interferes with trust (we describe these as penalties to trustworthiness [21]), compared to what causes trust.
Trust and credibility are closely related to one another and to information quality; although there remains disagreement among researchers as to the exact definitions of these terms. In general the definitions emphasise likelihood of information use, believability, reliability and dependability [22, 20]. Correspondingly, there is little evidence to suggest that the general population reliably make fine conceptual distinctions between trustworthiness and credibility of information.

Online health support often is divided into seeking information or emotional reassurance, and can be gender-specific (e.g. prostate cancer versus breast cancer) and person-specific [23, 24]. Trustworthiness remains pertinent to online emotional reassurance and sharing, as shown by the many occurrences of large-scale hoaxes designed to manipulate emotions [8]. For example, between 2010 and 2011 a Macmillan cancer forum was inundated with posts in response to an elaborate hoax by a purported mother about her six-year-old daughter struggling with cancer. Upon the exposure of the hoax (perpetrated by a lonely sixteen-year-old girl), many users of the forum – who had formed close online relationships with the supposed mother -- refused to believe it had all been a hoax [25]. In light of the range of uses that online health information has for the general public, we contend that it is important to build knowledge of the factors that influence how trust and its absence are formed online beyond source analysis and fact checking. As this study shows, both linguistic and meta-linguistic factors affect how trustworthiness is instinctively rated.

**Theoretical Underpinnings for Factors Influencing Trustworthiness**

The range of elements that influence online trustworthiness includes issues related to the source (e.g. author identification and the absence of advertising), issues related to the content (e.g. a date stamp and inclusion of medical evidence), issues relating to design and engagement (e.g. inclusion of images) and issues affecting all of the above (e.g. the absence of typographical errors) [20, 26]. For some time, credibility has been subdivided into aspects such as source credibility, message credibility and media credibility – which strongly influence each other [22]. All such aspects of credibility include accessibility, which can be relational and cognitive as well as physical [13]. More recently, Sun et al. [6] have divided the elements that influence consumer evaluation of online health information into 25 *criteria* and 165 *quality indicators*; criteria are rules that reflect notions of value and worth (e.g. expertise and objectivity), while quality indicators are properties of information objects to which criteria are applied to form judgments (e.g. the owner of the website and inclusion of statistics). In line with Diviani et al. [5], Sun et al. [6] have suggested that indicators can be
positive or negative (in terms of trustworthiness), and that consumers’ perceived online health information quality could conceivably be measured by a small set of core dimensions (i.e. a few groups of criteria might incorporate many of the quality indicators that explain most of the trustworthiness judgments).

Thus far, Lederman et al. [8] have proposed a five-category model that highlights verification processes via the comparison of different web sites or other online statements. This model includes argument quality, source credibility, source literacy competence and crowd consensus. An extended six-category model [26] proposes that the lay reader may assess some or all of the following: reputation, endorsement, consistency with other sources, self-confirmation (agreement with the reader’s own opinions), persuasive intent, and expectancy violation. For the purposes of our research, we have adopted Lederman et al.’s [8] model of credibility; however, the choice among these models for our current research is moot because all the models [20, 26, 5, 6, 22] are concordant with the idea that spelling errors will detract from judgments of trustworthiness (i.e. they are negative quality indicators).

**Incivility, Literacy Competence and Errors of Writing Mechanics**

Although institutionally produced web sites and curated online health content (cf. [6]) will usually be both grammatically correct and civil, the responses by the general public may be uncivil [27], and patient-authored text is known to have many misspellings [28] (and so do discharge summaries written by doctors [29]). Inappropriate capitalization (including inappropriate capitalization of entire words, sometimes called online ‘shouting’) and misspelling are examples of errors in literacy competence [8] and/or writing mechanics [30]; writing mechanics is defined as elements of a language that only manifest when communication is in written form. Both inappropriate capitalization and misspelling have been highlighted by qualitative investigations as explicit criteria used by lay readers in judgments of online credibility [8, 31]. The rationales given to explain why these two error types undermine trustworthiness are that either (1) the errors imply a lack of intelligence (expertise, ability, authority, education) [32, 8, 33], or (2) that they suggest a lack of motivation (objectivity, attention to detail, conscientiousness) to be trustworthy [34, 35]. The term incivility is used to describe this latter lack of motivation or effort to make statements that are compliant with rules of communication. Hargittai [33] distinguishes between spelling errors (due to levels of education or social inequalities) and typographical errors (errors due to accidentally hitting the wrong keys on the keyboard). The fact that these errors are not corrected during proofreading is another form of incivility. Incivility implies a lack of
respect for the reader, the platform, and/or the rules of social exchange [36, 37], and it is fundamentally relational. Quantitative research on incivility in the mainstream world wide web demonstrates that civil statements are rated as more trustworthy and influential than uncivil ones [38, 39].

Integration versus Heuristics: Lay Judgments Based on Multiple Cues

The reader who must judge unvetted online health information is faced with a wealth of cues that indicate the degree of trustworthiness, and those cues may have contradictory effects (e.g. a cogent message that is misspelled). There are three broad theories for how individuals (both lay and expert) make judgments based on multiple cues. The rational approaches are represented by the Information Integration model [40], in which an individual accounts for all the different pieces of information by a complex (but often subconscious) mathematical process that is usually based on addition, multiplication or averaging; extensive observations of integration in judgments occur across cultures and individuals. This is the process that Sun et al. [6] allude to when they propose that the consumer integrates the relevant trustworthiness criteria and quality indicators in a "complex cost-benefit analysis". In 2003 Fogg’s Prominence Interpretation Theory proposed explicit mathematical relationships for how to predict the effects of multiple factors on credibility, but the theory never detailed how to measure the relevant quantities independently [41]. Computational models typically assume that the elements of incivility (e.g. inappropriate capitalization) act together either additively or non-linearly [42, 43, 44], although hypothesis-led proof for this assumption is minimal. The elucidation of this integration is only just starting in the literature [45], and the relative importance of each indicator in this intuitive cost-benefit analysis remains unknown; the relative values for each cue may be elucidated empirically by statistical methods such as regression, but it is unlikely that these values would be explicit or transparent in the minds of most lay decision makers. This is an area that needs to be further researched.

An alternative theory to costs-benefit analysis (i.e. for how judgments are made based on multiple cues) is the process-level cognitive perspective [46]; this has been made famous by the heuristics and biases research program from behavioural economics [47]. Heuristics are rapid cognitive strategies/shortcuts (either explicit or subconscious) formulated as practical, bounded rational decision systems for multiple cues that can be more transparent than complex cost-benefit analyses, for example, hierarchical lexicographic decision models [48]. In a take-the-best judgment [49] first a single cue (the most important one) is searched for in the environment and considered independently of all others, and if a tie or no clear result
occurs, then the second most important cue is sought and decided upon, and so on. For example, when deciding whether you have the right of way when driving your car through an intersection, first you follow the signal of any policemen present, and only if there are no policemen present do you seek and consider a traffic light (including a temporary traffic light for construction), and only if there are no traffic lights present do you then consider static road markings and the positions of the other cars.

In biased heuristics, only a limited subset of information (often only one cue) is used to make a fast and ecological judgment outside of conscious awareness [46]; when biased, these heuristics are used to support preferred or preconceived outcomes. With biased heuristics, the prioritization of the cue, and even the cue's basic validity for the judgment being made, is dubious. Such biased heuristics have been used to explain seemingly irrational preferences that individuals make in situations involving slot machines and organ donation [50].

Examples of biased heuristic processes include representativeness -- where some cues are weighted disproportionately compared to their real representativeness -- and availability -- where a cue that is easily recalled (such as occurs with recency and news) determines the judgment. None of these judgment models so far proposed have explicitly assessed specific issues within literacy competence.

**Determining the Criteria and Weighting of Judgments Based on Multiple Cues**

A key feature of heuristics is that they are typically made subconsciously, and that *post hoc* explanations for such choices are often self-serving justifications or rationalizations [51]. That is, in the case of heuristics, the decision maker does not have privileged information on how the judgment was made, and, furthermore, the decision maker can be wrong about themselves [52]. For example, university students have been shown to greatly over-estimate how much they actually learned from excellent lecturers (conflating it with how much they feel they learned, which is discounted by effort and exertion) [53]. This creates potential issues for researchers, in which insight-based techniques (both qualitative interviews as well as quantitative questionnaires) can lead to judgments and explanations of causes that are inaccurate because of demand characteristics or social desirability [54, 55]. It has long been known that when evaluating credibility (e.g. website privacy policy), the importance of factors that lay individuals say they would use to make their judgment diverge from the factors they are observed to use [41]. Recent sophisticated measurements suggest that, although multiple factors can interact when quantifying perceived credibility (of information for an online health forum), these interactions do not support Petty and Cacioppo's
Elaboration Likelihood Model [56]. Furthermore, there is a gap in the literature for any data showing how spelling errors interact with other writing mechanics violations that might also affect trustworthiness [57]. This suggests that lay insight into the influences on their perceptions of trustworthiness are imperfect, and that research on trustworthiness should be supplemented by approaches that do not rely upon such insight.

In order to avoid insight bias from our lay participants, the approach of this study is to compare and contrast *marginal differences* in penalties to trustworthiness elicited by different combinations of literacy or writing mechanics violations. Our experiments on changes in marginal trustworthiness were specifically organised so that the participating healthy volunteers were unaware that the experiment tested the effects of capitalization or misspelling *per se* (with full institutional ethical approval). We did not explicitly ask lay individuals for their beliefs regarding how their process for judging message credibility incorporates quality indicators relating to source credibility and media credibility. That is, instead of asking directly, "How much less would you trust a web post that is misspelled?", we simply presented the participants with some posts that had misspellings and asked the same question as usual, "How trustworthy do you find this information?" while subtly varying misspelling and capitalization (i.e. ‘shouting text’, not acronyms or beginnings for sentences).

**Research Questions**

RQ1: Do errors in writing mechanics and incivility lead to marginal changes (i.e. without signposting) in judgments for message trustworthiness?

RQ2: Are the marginal effects of incivility, such as with inappropriate capitalization, on trustworthiness judgments quantitatively comparable to the known effects of writing mechanics such as spelling errors?

RQ3: How, if at all, do the effects of writing mechanics (e.g. spelling errors) and incivility (e.g. inappropriate capitalization) integrate? Is there a ceiling effect or a binary effect in which once a message source is judged as incompetent, no further trustworthiness penalty is added to the judgment (similar to a take-the-best heuristic) [49]? Or is there some additive, multiplicative, synergistic or otherwise integrative function that increases the penalty on the judgment to a new higher level when both cues are present [40]?

**Methods**
The project was approved by our local ethics committee (Brighton and Sussex Medical School's Research Governance and Ethics Committee (RGEC, University of Sussex), approval 16/044/WIT). All experiments were performed according to the Declaration of Helsinki. All individuals provided informed consent via a welcome page in each online study. Participants were recruited via the prolific.co website. We specified that the recruitment should focus on UK-based members of the public. Participants consented to participate with the understanding that the research concerned "responses to online text"; none of the advertising, web URLs or experimental information to the participants mentioned that the experiment was related to graphics, formatting, incivility, spelling, etc. This feature of the advertising was approved by our ethics committee, not least because the paragraphs were not considered misleading or potentially emotionally adverse.

Study Design Process

The study design was a confirmatory, cross-sectional experiment with a balanced incomplete block design; it was a randomized experiment with lay participants, each of whom experienced only a limited number of the possible options. The total number of experimental excerpts that were tested in the entire cohort was 36; there were 9 excerpts, each in 4 possible versions: no errors, inappropriate capitalization only, misspelling only, or the combination of both errors. The stimulus texts (and the versions) are shown in the supplementary methods 01. There were two additional paragraphs that always appeared as the first two stimulus texts, which were training stimuli. The training stimuli were not labelled as being different from other stimuli in any way, and were never included in the statistical analyses. The only purpose of the training stimuli was to allow participants to familiarise themselves with the rating task, and with the range of the trustworthiness scale: Training stimulus 1 was quite believable (mean trustworthiness rating = 54.33 ± 23.39, mean ± sd, n = 301), and training stimulus 2 was less plausible (mean trustworthiness = 40.20 ± 25.80, n = 301, \( P < .001 \), paired t test).

Each participant experienced and rated only 9 (of the possible 36) experimental excerpts, and those 9 included exactly one version of each excerpt, and among those 9, that participant would experience a mix of different error types (Figure 01). For example, participant number 001 experienced the capitalization-only version of excerpt E07, then the "both errors" version (capitalization and misspelling) version of excerpt E02, etc. This double randomization prevented any participant from seeing the same excerpt twice.
The goal was to present to participants a set of coherent text excerpts of similar lengths (70-100 words) on the topic of multiple sclerosis, each excerpt being a coherent answer to a question. The rationale for presenting excerpts in a question-then-answer format was that it was possible to ask whether readers trusted the advice enough to believe it or act upon it. To make these excerpts, a range of comments were found in the public domain (see Table 01); these texts often had to be edited substantially to fit within the word count or to avoid explicitly endorsing commercial products (see Supplementary Methods 02 for a comparison of the original and presented texts). Initially the excerpts and the presentation system were tested online by a small group of testers, who then provided verbal feedback on the test to the experimental team. After that a cohort of 40 participants was recruited online to test the paragraphs and demonstrate that the paragraphs elicited similar standard deviations in ratings (19 to 28 units out of 100) and elicited a wide range of mean trustworthiness ratings (from 20 to 70). These excerpts were to be presented either as they were ("no errors" or the negative control) or in one of the three error versions stated above.

Table 01. Sources of excerpts on multiple sclerosis.
Text Interventions

For the error versions of the texts (inappropriate capitalization or misspelling), we wanted to include 5 of the relevant errors per excerpt (10 errors total for the combination of both errors version), with the errors spread throughout the excerpt (rather than bunched together). In excerpts where inappropriate capitalization was required, there would be five "sets" of words/phrases. Normally a set was one or two words, although one of the five sets had to be a 4-word series. The priorities for selecting words to capitalize were (in this order):

1. adverbs (especially those suggesting extremity such as "very" or "never")
2. judgments ("rubbish", "hopeless", "horror")
3. strong emotions ("worried", "angry")
4. words implying danger ("fatal", "death")
5. amounts ("all", "every", "ten")
6. adjectives (rather than nouns)
7. action verbs (especially gerunds)
8. conjunctions ("and")

To verify that each word that we capitalized was naturally capitalized on the web, we analyzed words that were capitalized online on Twitter. We used the Claritin corpus, which is a crowd sourced data set of all the Twitter tweets that contained the word "claritin" in the month of October 2012 [58]. This corpus has some 4,900 tweets, and we used Matlab to find all the words that were in all capitals (which did not have a hashtag or an at-sign in them); this resulted in a list of 343 capitalized words (see Supplementary Methods 03), many of which were short words, acronyms and internet memes. From this list of words spontaneously capitalized on the web, we selected words in our excerpts to capitalize.

The rationale for how we selected words to misspell was that misspellings should be quite noticeable, and that the meaning of the words should remain clear to the reader even when misspelled. We avoided homonyms and words that looked plausibly English when misspelled. To make sure that misspelled words were noticeable, short words were preferred, or we placed the misspellings in the first syllable of a multi-syllable word. In addition, one of the misspelled words had to be in the first five words of a paragraph. The misspelled word had to be completely understandable (in the absence of other words or context) even when misspelt. Thus, a misspelled word with missing or added letters should be pronounceable in English (e.g. "yu" plainly means "you"). The types of misspellings were:
1. swap one letter for another letter that is next to it on a qwerty keyboard ("pisitive")
2. double a consonant ("esstimate")
3. double a vowel, or add an extra vowel ("theere")
4. leave out a vowel ("expsure")

To verify that each word that we misspelled was naturally misspelled on the web, we searched for the misspelled word along with the words "health" and "forum"; if we could not find at least two examples of a misspelling on online health forums, we did not use it. A complete listing of the misspellings and where we found them online is in supplementary methods 04.

**Study Delivery**

The study was presented to participants using the Qualtrics portal, which allows for a wide range of question types and keeps track of answers and total response time. A full description of the survey in the CHERRIES format [59] is included as part of the supplementary methods 05 for this paper. The online study welcome page explained in brief what the study was about and what it entailed (estimated 8 minutes participation time, including reading the ethics and filling in demographics), the ethics of the study (include the ability to withdraw at any time), and a brief complaints procedure. The ethics page explicitly excluded participants under 18 years of age, or those from vulnerable populations. A pointer to a full length Participant Information Sheet (3 pages) was shown; the welcome/ethics page had an "I agree" button at the bottom. After the welcome page, participants filled in a brief multiple choice demographics page, which included questions on gender, age, field of work (e.g. healthcare, agriculture, retired), and familiarity with English language/Roman alphabet. All demographics questions included an option for "rather not say". After the demographics page, participants saw the instructions page, and then were launched into the experimental ratings pages.

Each ratings page consisted of a short excerpt of text (which was randomized as to whether or not it had the spelling or capitalization errors), followed by a horizontal slider for rating how "trustworthy" the participant found the statement to be; the slider had anchors of "completely untrustworthy" (left) and "completely trustworthy" (right, see figure 02). Although the data collection was numerical (0-100, left to right), there were no numerical cues or tic marks seen by the participants. The instructions to the participants for the trustworthiness ratings were, "If you find something trustworthy, you would be prepared to act upon it; an untrustworthy statement you would ignore, and a rating in the middle represents information where you
would want more proof or confirmation that it is correct”. As explained in the instructions to participants, each stimulus excerpt was written as if it was an answer to a question written on an unmoderated health forum, with a specific focus on multiple sclerosis. Multiple sclerosis was chosen as a topic because the information was obviously important, but healthy participants would be unfamiliar with the veracity of each statement; thus, we predicted that the trustworthiness ratings would be more susceptible to non-verbal cues. The questions were:

1. Is multiple sclerosis preventable?
2. How risky is Tecfidera as a treatment for multiple sclerosis?
3. Does multiple sclerosis decrease intelligence/IQ?

The nominal responses to these questions were the experimental stimulus excerpts being rated.

![Figure 02. Unnumbered horizontal slider for trustworthiness ratings.](image)

**Study Design, Analysis and Statistics**

**Sample size**

To detect a difference in rating scores between two groups (i.e. no errors versus misspellings, inappropriate capitalisation, or both) with 80% power at the 5% significance level, assuming the standard deviation within each group is 25 and the difference between groups is 15 (equivalent to a medium effect size of 0.517, would require 60 participants per group (120 total). Each participant was asked to rate 9 text excerpts, randomly divided between the 4 conditions and assuming an intraclass (i.e. within participant) correlation coefficient of 0.185 gives a design effect of 2.48. The product of the design effect and the sample size for a non-repeating experiment is $2.48 \times 120 \approx 300$ total.

**Modelling**

Linear mixed effects models were fitted in Stata version 16.0 using the "mixed" command. Residuals were checked for normality and homoscedasticity at the cluster and individual levels. Where the assumption of residual homoscedasticity was not appropriate, robust
standard errors were used to allow for calculation of appropriate 95% confidence intervals and p-values [60].

**Data Available Online**

Supplementary methods as well as raw data is available online at: https://github.com/harry-witchel/CapitalizationMisspelling

**Results**

**Differences between the Paragraphs**

We ran an experiment in which we gathered data from 301 volunteers, each making 9 experimental observations (2,709 ratings in total). We estimated that this task (rating 11 excerpts plus instructions and demographics) would take 8 minutes; in fact, it took a mean time of 6.25 minutes (375.74 ± 12.94 s, mean ± s.e.m.). The median trustworthiness ratings of each of the excerpts (E01 to E09) in the negative control condition (i.e. without any errors or incivility) are shown in the box and whisker plot in Figure 03. On each box, the red line in the centre indicates the median, and the bottom and top edges of the box indicate the first and third quartiles, respectively. When the notches for two boxes do not overlap, this implies that the true medians do differ with 95% confidence. The whiskers extend to the most extreme data points not considered outliers, while outliers are shown using red plus signs; outliers are any points that fall more than 1.5 × the interquartile range away from the main box. Excerpt 01 generally elicits low ratings of trustworthiness (median = 27.5), while excerpts 08 and 09 generally elicit high ratings of trustworthiness (medians = 62 and 63, respectively). As illustrated by the non-overlapping notches, these two sets of excerpts elicit significantly different ratings of trustworthiness, which has been true in all our previous cohorts rating these excerpts for trustworthiness (data not shown). Excerpts 03, 04, 05 and 06 all elicited median ratings in the middle range from 45 to 55, while excerpt 02 is a transitional excerpt between E01 and the middle range and E07 is a transitional excerpt between the middle range and the most trustworthy excerpts.
Figure 03. Trustworthiness ratings of the different excerpts in the negative control (no errors) condition. For each box N = 75.

Cumulative Distributions Shifted Left by Errors

Figure 04 shows how the errors in writing mechanics and incivility led to changes in the cumulative probability distributions for each of the excerpts. As expected, for each excerpt, compared to the negative control with no errors (in black), all three alterations (inappropriate capitalization in blue, misspelling in magenta, and both errors together in red) led to a decrease in the ratings of trustworthiness (i.e. a leftward-upward shift in the curve). For most of the excerpts, at most points on the cumulative distribution curve, the combination of errors (both inappropriate capitalization plus misspelling) led to decreases in trustworthiness ratings (i.e. a shift left and up) compared to either of the single errors. That is, the lines for inappropriate capitalization only (blue) and misspelling only (magenta) fall between the black line (no errors) and the red line (both errors); this results suggests that the two errors together have an additive or integrative effect.
Figure 04. Cumulative probability distribution plots for each excerpt (E01 to E09), comparing the alternative writing errors (coloured lines) to negative control (black).

**Mixed Linear Effects Model with Four Conditions of Alteration**

We tested this data in a mixed linear effects model (Model 1) where trustworthiness rating was the dependent variable, while alteration (no errors/misspellings/inappropriate capitalisation/both misspellings and inappropriate capitalisation) and Excerpt (1-9) were included as fixed effects. A random effect for participant was included in the model to account for clustering of observations within volunteers. The reference group/condition for this model was E05 with no literacy errors. This model (and the following model) were calculated with robust standard errors [60] to allow for the heteroskedasticity of the residuals. The results of Model 1 are shown in Table 02. The intracluster correlation (correlation within the individuals) coefficient estimate for Model 1 is 0.334 (95% CI: 0.287 to 0.384).

<table>
<thead>
<tr>
<th>Categorical Variables</th>
<th>Coefficient</th>
<th>Robust Std. Err.</th>
<th>P-value</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. Error</td>
<td>df</td>
<td>t value</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>------------</td>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>Caps Only</td>
<td>-6.41</td>
<td>1.30</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Misspelled</td>
<td>-8.86</td>
<td>1.40</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Both Errors</td>
<td>-14.33</td>
<td>1.42</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Excerpt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E01</td>
<td>-11.94</td>
<td>1.71</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>E02</td>
<td>-8.29</td>
<td>1.58</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>E03</td>
<td>-0.70</td>
<td>1.56</td>
<td>1</td>
<td>0.653</td>
</tr>
<tr>
<td>E04</td>
<td>0.03</td>
<td>1.54</td>
<td>1</td>
<td>0.983</td>
</tr>
<tr>
<td>E06</td>
<td>1.80</td>
<td>1.62</td>
<td>1</td>
<td>0.269</td>
</tr>
<tr>
<td>E07</td>
<td>6.19</td>
<td>1.71</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>E08</td>
<td>14.75</td>
<td>1.53</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>E09</td>
<td>14.22</td>
<td>1.46</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>constant</td>
<td>47.06</td>
<td>1.53</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 02. Mixed Effects Model 1 for trustworthiness rating, with errors and excerpts as fixed effects and a random effect for the clustering of data by participant.

There is strong evidence against each of the null hypotheses that inappropriate capitalization only (Result 1), misspelled only (Result 2) and both errors (Result 3) do not affect trustworthiness ratings compared to the negative control group; stated positively, our data suggest that there is a statistically significant penalty to trustworthiness for inappropriate capitalization (Result 1), misspelling (Result 2), and for both errors together (Result 3). Inappropriate capitalization reduces trustworthiness ratings by -6.41 (95% CI: -8.96 to -3.86) and misspelling reduces trustworthiness ratings by -8.86 (95% CI: -11.61 to -6.11). The effect on trustworthiness ratings of combining both inappropriate capitalization and misspellings together is -14.33 (95% CI: -17.11 to -11.55), which appears to be an additive effect.

Our further analysis aimed to test whether there was likely to be either an additive or integrative effect [40] of combining inappropriate capitalization and misspelling. Such an effect should lead to a significantly larger trustworthiness penalty when both error types are combined compared to either error individually. To test for this, an alternative specification of the LME model of the same data was formulated (Model 2). In Model 2 the condition 
"both errors" was specified as an interaction between two binary variables, (1) misspelling and (2) capitalisation errors, instead of considering all the errors as four conditions in a single independent variable. The specification of the rest of the model was identical to Model 1. In this model (Table 03) the main effects and an interaction between them provide no evidence for an interaction effect between the two variables; that is, the main effects for inappropriate
capitalization only and for misspelled only were as in the original Model 1 (Table 02), while
the coefficient for the interaction was not significantly different from zero. This supports the
interpretation that the effects of the two error types are additive, rather than partially
summative or than synergistic.

<table>
<thead>
<tr>
<th>Categorical Variables</th>
<th>Coefficient</th>
<th>Robust Std. Err.</th>
<th>P-value</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capitalized</td>
<td>-6.41</td>
<td>1.30</td>
<td>&lt;0.001</td>
<td>-8.96 to -3.86</td>
</tr>
<tr>
<td>Misspelled</td>
<td>-8.86</td>
<td>1.40</td>
<td>&lt;0.001</td>
<td>-11.61 to -6.11</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.94</td>
<td>1.68</td>
<td>0.576</td>
<td>-2.34 to 4.24</td>
</tr>
</tbody>
</table>

Table 03. Model 2: Alternatively specified LME model with two binary variables for
capitalisation and spelling errors and an interaction term to account for combining
both types of error. All unlisted values are identical to Table 02 above.

The output for Model 1 shows the following comparisons: (1) capitalization only versus no
errors, (2) misspelled only versus no errors and (3) both errors versus no errors. Table 04
shows the additional comparisons between (4) capitalization only versus misspelled only, (5)
both errors versus misspelled only and (6) capitalization only versus both errors.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Coefficient</th>
<th>Robust Std. Err.</th>
<th>P-value</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Caps only vs. no errors</td>
<td>-6.41</td>
<td>1.30</td>
<td>&lt;0.001</td>
<td>-8.96 to -3.86</td>
</tr>
<tr>
<td>(2) Misspelled vs. no errors</td>
<td>-8.86</td>
<td>1.40</td>
<td>&lt;0.001</td>
<td>-11.61 to -6.11</td>
</tr>
<tr>
<td>(3) Both errors vs. no errors</td>
<td>-14.33</td>
<td>1.42</td>
<td>&lt;0.001</td>
<td>-17.11 to -11.55</td>
</tr>
<tr>
<td>(4) Caps only vs. misspelled</td>
<td>2.45</td>
<td>1.31</td>
<td>0.061</td>
<td>-0.12 to 5.02</td>
</tr>
<tr>
<td>(5) Both errors vs. misspelled</td>
<td>-5.47</td>
<td>1.20</td>
<td>&lt;0.001</td>
<td>-7.83 to -3.11</td>
</tr>
<tr>
<td>(6) Caps only vs. both errors</td>
<td>7.92</td>
<td>1.20</td>
<td>&lt;0.001</td>
<td>5.56 to 10.28</td>
</tr>
</tbody>
</table>

Table 04. All possible alterations tested in between group comparisons (Model 1).

Comparison (4) indicates there is weak evidence ($P = .061$) against the null hypothesis of no
difference between the effects of capitalization only and misspelled only. That is, misspelled
only leads to a larger trustworthiness penalty by -2.45 (95% CI: -5.02 to 0.12) compared to
capitalization only. There is strong evidence against the null hypothesis of there being no difference between the effects of capitalization only versus both errors combined (comparison 6). Compared to capitalization only, the combination of both errors significantly reduces trustworthiness ratings by a further -7.92 (95% CI: -10.28 to -5.56). Likewise, there is strong evidence against the null hypothesis of there being no difference between the effects of misspelled only and the combination of both errors (comparison 5). Compared to misspelling only, the combination of both errors leads to a further penalty to trustworthiness ratings of -5.47 (95% CI: -7.83 to -3.11).

**Statistically Significant Differences between the Excerpts**

In Model 1, when compared to E05, the effect of the various excerpts' contents on trustworthiness ratings varies from -11.93 to 14.75 (a range of 26.68). This range is roughly twice as large as the effect of both errors (-14.33), suggesting that the errors in incivility and writing mechanics that we tested can together have an overall effect of nearly half of the effects of the content of the excerpts we tested.

**Discussion**

**Original contributions**

This study sought to quantitatively determine how two different errors of writing mechanics (contributing to incivility) combine to penalize subjective ratings of trustworthiness, in the medically relevant context of materials typical of an unmoderated online health forum. Using an LME model of a suitably powered study, we found that all three interventions (inappropriate capitalization, misspelling, and the combination of the two) were significantly different from the negative control (no added errors or incivility), which clearly answers RQ1. The data also show that (for these 70-100 word long excerpts about multiple sclerosis) the trustworthiness penalty for five instances of inappropriate capitalization was of a similar magnitude to the penalty for five instances of misspelling. Note that there was a trend for the penalty of misspellings to be larger, but as a generalized rule, the two are similar in magnitude and the precise difference will depend on exactly how many words and which words are capitalized or misspelled. This finding answers RQ2. Finally, with a combination of different Linear Mixed Effects Models, the data show that the combination of two different types of errors had a significantly greater trustworthiness penalty than either of the error types
alone, and that the effect in this study was almost perfectly additive (RQ3); thus, the effects of the combination of errors was integrative [40] rather than a simplified heuristic such as take-the-best [48], multiplicative, or affected by ceiling effects in this study. This begins to answer the question recently posed of how spelling interacts with other quality indicators on credibility [57]. To the best of our knowledge, this is the first study that was specifically designed to test and quantify these kinds of specific additive effects on trustworthiness independently of the lay participants' insight.

We also showed that the stimulus excerpts that we designed are appropriate for studies that test trustworthiness of online health information independently of lay insight. While our study did not preclude lay insight (i.e. participants might notice that some words were misspelled, etc.), the study was not dependent on such insight, which is useful for interrogating intuitive evaluations of information (i.e. cost-benefit analyses [6]). In this study, the semantic content of the excerpts engendered consistent effects on trustworthiness ratings (at least among this type of online psychology experiment cohort, see Figure 03). Also, this is the first time that numerical values have been gathered for isolated effects of inappropriate capitalization.

The Results in Context

This study reaffirms an earlier observation that incivility decreases message credibility [38, 39]. As suggested previously, inappropriate capitalization (shouting) is histrionic, and induces a strong effect of incivility on readers' subjective ratings [37]. In our study, the effect of shouting text showed a trend for eliciting a smaller trustworthiness penalty than the effect of similar quantity of misspelling. One could easily speculate about new experiments where we might change the quantities of literacy errors; our experiment used either five misspellings or five shouting phrases, but one could run experiments to titrate errors, for example to determine the relative effects of three misspellings or ten inappropriate capitalizations. Furthermore, the effects of text shouting may be moderated by whether the statements are controversial [38]. We deliberately chose statements about multiple sclerosis that would be unfamiliar to the general public. This lack of familiarity makes the content/context not particularly emotional or controversial, likely weakening the effect of inappropriate capitalization.

The debate between how accessibility (i.e. cost and speed) versus information quality (i.e. accuracy and presumed benefit) quantitatively affect the use of (and search for) information
Categorical frameworks have long been proposed to provide a theoretical underpinning for the factors, grouping specific elements that influence perceptions of trustworthiness in a variety of ways. The most well-known two-category grouping of factors affecting how people respond to communication is Petty and Cacioppo’s Elaboration Likelihood Model (ELM) for persuasion [56], in which elements contributing to a central pathway (e.g. argument quality) are complemented by seemingly less rational elements that contribute to a peripheral pathway (e.g. website design) [41, 20]. Another two-category persuasion model that has been used to explain online trustworthiness is Chaiken’s dichotomy of heuristics versus systematic information [61, 62]. In both the Heuristic-Systematic Model and the ELM, diminishing motivation/involvement (or diminishing user dependency [45]) is associated with a switch from focusing on the effortful systematic evaluation of information (quality) to low-effort heuristics (accessibility). In purchasing decisions, the type of product affects how strongly grammar and mechanics errors affect credibility [30]. In particular, ‘experience goods’ (non-technical items like body lotion that are used personally) are more affected by grammar and writing mechanics than ‘search goods’ (technical items like printers). The implication is that when objective signals about trustworthiness are absent, heuristics play a stronger role [61, 56]. In the current study, where laypersons made judgments about unfamiliar medical issues, we might expect to find a stronger response to misspelling and capitalization. A necessary future approach is to repeat this experiment with multiple sclerosis patients who would exhibit user-dependence when evaluating the statements [45].

Limitations

Our study had several important constraints. We deliberately tested laypersons judgments on unfamiliar ideas about multiple sclerosis, and showed that literacy errors can have a strong effect. Nevertheless, this effect may be smaller in a cohort that is more dependent on knowing the information. In particular, if readers are dependent, then pre-existing points of view and feeling of ‘homophily’ [63] will influence perceived credibility, in a way that would not influence the general public with less interest in statements for and about multiple sclerosis.

It is also notable that in the current study that the highest mean trustworthiness rating was 62.3 (0-100 scale), despite the statement being medically correct. Multiple factors may account for why this mean rating is not higher for trustworthiness. For example, participants
in this psychology experiment saw the statements in vacuo, so that they could not verify the statement, check source credibility, or look for crowd consensus [8]. In real life situations, other aspects of trustworthiness may dominate, and among some individuals there may be ceiling and floor effects within this data set, given the very wide standard deviations.

**Conclusions**

Incivility and literacy competence are proposed factors in how lay web users assess trustworthiness of online health information. These results support Lederman et al.'s [8] theory of credibility assessment in online forums; the results also fit with Anderson's [40] description of integrative assessment of multiple cues. Here, we have shown that literacy competence errors have additive effects. The additive effects are strikingly precise. This implies that when people make seemingly rapid judgments about the trustworthiness of online text -- no more than rough estimates -- that their intuitive estimates seemingly add up quite accurately -- at least at the population level. This completely vitiates binary models of judgment where a writer is judged as either competent or incompetent, with no further penalty for additional errors in writing mechanics [57]. The implication for writers is that for this level of errors (five in one paragraph), multiple errors can add up.

Many other factors also contribute to trustworthiness, notably the argument quality of the content (logic), verification with other sources, reference credibility, and crowd consensus [8]. How these additional factors interplay with literacy competence will require further extensive research. A start would be to determine how persons with multiple sclerosis (with similar education levels to our current cohort) respond to these stimulus excerpts, as those readers would understand the information in a more self-relevant context.

**Acknowledgements**

We gratefully acknowledge: Tom Ormerod and The School of Psychology at the University of Sussex for access to Qualtrics. We acknowledge Matthieu Raggett for IT set up. We also acknowledge the BSMS Independent Research Project programme for funding. Finally, we acknowledge the late Fayard Nicholas for the original idea on civility.

**Competing Interests**

BM owns and JR worked for a commercial enterprise (Dalton Maag) that creates typefaces and graphical designs for commercial clients. AN acts as an ad-hoc Scientific Consultant for
Dalton Maag. This enterprise and those authors will not benefit directly or indirectly from data supporting incivility, misspelling, or shouting text.

HJW, GAT, CEIW, CIJ and HDC have no competing interests to declare.

Abbreviations

ELM – the Elaboration Likelihood Model
LME – Linear Mixed Effects (statistical model)
SE – standard error
s.e.m. – standard error of the mean
vs. – versus

Author Contributions

HDC, HJW, GAT, BM, AN & CEIW conceived the experiments. HJW, GAT, CEIW and JR conducted the experiments. HJW, CIP & GAT analysed the results. CIP performed the statistical analyses. HJW, GAT and HDC drafted the manuscript. All authors reviewed the manuscript.

References


652  Research 2015;17(5):e112. PMID: 25953147


682  15. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm–An overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012 May; 28;30(25):3778-89. PMID: 22172504


691  18. Pickard, V. Media failures in the age of Trump. The Political Economy of Communication 2017:


61. Chaiken S. Heuristic versus systematic information processing and the use of source versus
