

Style Matters

Reihane Boghrati^a, Jonah Berger^{a,1}, and Grant Packard^b

^aUniversity of Pennsylvania; ^bYork University

This manuscript was compiled on February 12, 2021

1 **Why do some things succeed in the marketplace of ideas? While**
 2 **some have argued that an idea's content drives success, others sug-**
 3 **gest that style, or the way in which ideas are presented, also plays**
 4 **an important role. To provide a particularly stringent test of style's**
 5 **importance, we examine it in a context where one might imagine**
 6 **content is paramount and style is not: academic research. While**
 7 **scientists often see writing as merely a disinterested way to com-**
 8 **municate unobstructed truth, natural language processing of over**
 9 **75,000 articles from a range of disciplines indicates that writing style**
 10 **shapes research's impact. Ancillary analyses further suggest how**
 11 **style matters, highlighting the role of writing complexity, use of per-**
 12 **sonal voice, and temporal perspective (i.e., past vs. present tense).**
 13 **Taken together, the results provide empirical evidence that (writing)**
 14 **style matters, suggest how to boost academic research's impact, and**
 15 **highlight the value of natural language processing to shed light on**
 16 **drivers of cultural success.**

Language | Natural Language Processing | Cultural Success

1 **W**hy do some things succeed in the marketplace of ideas?
 2 One possibility centers on content. Certain things suc-
 3 ceed because they are better than what came previously (e.g.,
 4 relative advantage; 1). High speed internet replaced dial-up
 5 because it was faster and easier to use. Einstein's theory of gen-
 6 eral relativity replaced Newton's law of universal gravitation
 7 because it better explained the experimental evidence.

8 Another possibility, however, has less to do with content
 9 and more to do with style. The manner, or style, with which
 10 things are presented shapes their impact.

11 To provide a particularly stringent test of style's importance,
 12 we examine it in a context where one might imagine content is
 13 paramount and style is not: academic research. Science prides
 14 itself on being an objective exercise, where writing is merely a
 15 disinterested way to communicate unobstructed truth (2, 3).
 16 The notion is that some discoveries (e.g., general relativity) are
 17 simply more novel, groundbreaking, or valuable than others,
 18 and citations are seen as an unbiased measure of such quality
 19 (4).

20 Testing style's impact, however, is challenging. It's one
 21 thing to theorize that certain writing approaches are better,
 22 but actually measuring adherence to those approaches and
 23 linking them to a consequential outcome is difficult. Further,
 24 it can be tough to separate style from content. Even if papers
 25 that write certain ways (e.g., use more emotional language)
 26 are cited more, this could be driven by the subject matter
 27 discussed. Papers studying certain topics (i.e., emotions) likely
 28 use more language related to emotion, and thus topic, rather
 29 than writing style itself, could be driving impact.

30 To address these challenges, we focus on a small class of
 31 words that play a unique role in communication. Function
 32 words (e.g., conjunctions, grammatical articles, and preposi-
 33 tions, such as "and", "the", and "on") make up only a tiny
 34 portion of the human vocabulary (i.e., 0.04%; 5) but appear

in every sentence. They convey little semantic value on their
 own, but bind and enrich the nouns, verbs, and adjectives
 that make up communication content (6). Because they are
 largely meaningless without content, function words are often
 treated as junk by language scholars (7) and tossed out like
 meaningless garbage before text analysis is performed ("stop
 words;" 8).

But while function words tend to receive little attention
 from both scholars and communicators, they are particularly
 valuable here because they capture style rather than content.
 Indeed, researchers often refer to them as "style words" because
 they are seen as reflecting things about a communicator's
 linguistic style rather than anything about what is being
 discussed (7). Consequently, if function words help explain the
 impact of academic research, it suggests that style matters.

To test this possibility, we examine tens of thousands of
 articles from a range of disciplines. Controlling for article
 content, we examine whether writing style impacts citations,
 and if so, how.

Results

First, we examine whether style matters. Results of a negative
 binomial regression suggest that above and beyond the variance
 explained by non-language features ($R^2 = 0.136$), adding style
 features helps explain how many citations articles receive
 ($R^2 = 0.153$, $F = 14.600$, $p < .001$, see Table 1). Adding style
 features also adds predictive power even once article content
 is included ($R^2 = 0.224$ vs. 0.214 , $F = 2.620$, $p < .05$, see
 Table 1). Results persist controlling for other factors (e.g.,
 author prominence, where authors are from, and other content
 controls, Supplementary Information) and ancillary analyses
 suggest the effects are driven by more than just abstracts
 alone.

These results are intriguing, but one could wonder whether
 they might somehow be driven by the modeling approach
 used. While results are the same using a penalized regression,
 maybe there are non-linear relationships between the non-
 language or content controls and citations, for example, or
 interactions between these variables that, once included, would
 wipe out any effect of style features. Even using a more
 sophisticated two-layer feed-forward neural network to predict
 citations (Supplementary Information), however, including
 style features still adds additional predictive power ($R^2 = 0.234$
 vs. 0.209 , $F = 7.000$, $p < .001$).

Taken together, these results suggest that style matters.
 Style features increase the variance explained by 1.0-2.5%,
 which is 4-11% of the overall variance explained and 20-27% of

J.B. designed the theory. R.B. collected data and performed the computations. R.B., G.P., and J.B. verified the analytical methods. J.B., G.P., and R.B. wrote the manuscript.

Authors have no competing interests to disclose.

²To whom correspondence should be addressed. E-mail: jberger@wharton.upenn.edu

81 the variance explained by language content (i.e., LDA topics).
82 This result suggests that fewer than 500 style words that
83 contain no ideas or content on their own hold up against over
84 7 million content words in explaining an idea's success.

85 Second, we examine *how* style matters: whether certain
86 ways of writing increases impact. Analyses suggest complexity,
87 personal voice, and temporal perspective may all play a role.

88 *Complexity.* While academic ideas are often quite complex,
89 communicating them more simply should increase their impact.
90 This may be particularly important in the beginning of an
91 article where authors are laying out their thinking and ex-
92 plaining how their work relates to prior research. To test this
93 possibility, we extract each article's front end (i.e., literature
94 review and theorizing, Supplementary Information).

95 Results indicate that papers whose front ends use less
96 complex writing are cited more. *Articles* and *prepositions* are
97 two types of style words linked to cognitive complexity (9–11).
98 Grammatical articles ask readers to make distinctions between
99 a single case or class of something (e.g., the car vs. a car) while
100 prepositions describe the nature of linkages between nouns,
101 pronouns, or phrases (e.g., growth *despite* inflation; ate more
102 candies *except* when). Consistent with the notion that less
103 complex writing boosts citations, papers that use fewer articles
104 ($\beta = -0.030, p < .001$) or prepositions ($\beta = -0.016, p < .01$) in
105 the front end are cited more. Traditional readability measures
106 show similar effects (Supplementary Information).

107 We do not mean to suggest that complexity is always
108 bad. Complexity in the methods and results may sometimes
109 be useful or even required. Indeed, the cost of complexity
110 seems to weaken ($\beta_{articles} = -0.009, p = .10$) or even reverses
111 ($\beta_{prepositions} = 0.029, p < .001$), in the middle section where
112 methods or results are discussed.

113 *Personal Voice.* Academic writing guides have long sug-
114 gested that authors should write in a manner that is distant,
115 objective, and devoid of self-reference (e.g., first-person pro-
116 nouns like *I* or *we*; 12, 13). But is that actually more effective?

117 In contrast to prior suggestions, results suggest that per-
118 sonal voice may sometimes be beneficial. Papers whose
119 front ends are written with more function words used for
120 self-reference (i.e., first person pronouns) are cited more
121 ($\beta = 0.083, p < .001$). Papers written with more first-person
122 pronouns in the middle, usually empirical section, however,
123 are cited less ($\beta = -0.033, p < .05$). This may reflect the
124 degree to which personal ownership is valuable in different
125 parts of a paper. Taking personal ownership of arguments,
126 hypotheses, and contributions (e.g., “we suggest” versus “the
127 present research suggests”) that are empirically supported
128 may make the authors seem more prescient, increasing the
129 perceived authority of the research. Taking personal own-
130 ership of methods and results (e.g., “we asked participants
131 to do X,” versus “participants did X”), however, may make
132 methodological choices seem more subjective. Similarly, saying
133 “we show” rather than “results show” may make it seem that
134 the results are driven more by author choices, and thus less
135 objective. Along these lines, impersonal pronouns (e.g., it or
136 that), which remove a personal actor from the results (e.g.,
137 “this shows” rather than “we show”) are also linked to greater
138 citations in the middle section ($\beta = 0.077, p < .001$).

139 *Temporal Perspective.* Journal style guides, and academics
140 themselves, commonly recommend describing research using
141 the past tense (2, 12, 14). But is that actually more effective?

142 Analysis of a function word category (i.e., auxiliary verbs),
143 as well as temporal language more generally, suggests past-
144 focused language may actually reduce citations. Auxiliary
145 verbs (e.g., had or will) modify content verbs and can signal
146 their framing in time (e.g., “had considered” or “will consider”).
147 While a paper's content (i.e., theorizing, methods, and results)
148 occurred in the past, using present tense may make that
149 content seem more current and in the moment (15). To test
150 this possibility, judges coded each auxiliary verb based on
151 whether it referenced the past, present, or future.

152 Results indicate that while papers written with more past-
153 focused auxiliary verbs are cited less ($\beta = -0.100, p < .001$),
154 those written with more present-focused auxiliary verbs are
155 cited more ($\beta = 0.072, p < .01$). Temporal language shows
156 similar effects (Supplementary Information).

157 50,000+ Additional Articles

158 One might wonder whether the results are somehow driven by
159 the particular journals used. To test this possibility, we col-
160 lected an additional dataset of 52,633 articles from 27 journals
161 in areas such as chemistry, biology, physics, medicine, engi-
162 neering, and computer science (Supplementary Information).
163 Results are extremely similar. Style features still help explain
164 how many citations articles receive, above and beyond article
165 content. Further, style matters in similar ways. Even in this
166 alternate set of journals, writing complexity, personal voice,
167 and temporal perspective seem to play a role in research's
168 impact.

169 Discussion

170 Academics and practitioners alike have long debated about
171 why things succeed in the marketplace of ideas. But while
172 content certainly matters, the present work suggests that style
173 also plays an important role. Even in a domain like academic
174 research, where writing is often seen as merely a disinterested
175 way to communicate truth, writing style helps explain the
176 impact (i.e., number of citations) ideas achieve.

177 Ancillary analyses shed light on how style matters. While
178 academic ideas are often complex, results suggest that ex-
179 plaining things simply may be important, particularly in a
180 paper's front end. While some have suggested that first per-
181 son voice is bad because it distracts from the paper's content
182 (12), results suggest that there are times for taking personal
183 credit for writing (i.e., front end) and others for letting the
184 paper's content stand on its own (i.e., methods and results).
185 And while journal style guides often suggest using past tense,
186 results suggest that using the present tense (e.g., we theorize
187 instead of theorized) may be more beneficial.

188 These findings have clear implications. Peer-reviewed re-
189 search often adopts a dry, dense, and impersonal style that
190 can be challenging to both read and understand (16, 17). But
191 while academics across disciplines have intermittently theo-
192 rized about what counts as “better writing” (12, 18), little
193 work has actually tested these suggestions. This work suggests
194 that a few relatively simple shifts in writing (e.g., simplicity)
195 may help boost research's impact. In addition, the findings
196 also highly the importance of style in the marketplace of ideas
197 more generally. Ideas succeed not only based on their content,
198 but how they are framed and conveyed.

199 Future research might delve into other ways language shapes
200 impact. Expressing more certainty, for example, may be bene-

Table 1. Style words and citations

	Baseline	+Style	Baseline + Content	+Style	Baseline + Content	+Style
Overall R2	0.136	0.153	0.214	0.224	0.209	0.234
Style Features		yes		yes		yes
Content Controls						
LDA Topics			yes	yes	yes	yes
Non-Language Controls						
Publication Year	yes	yes	yes	yes	yes	yes
Journal	yes	yes	yes	yes	yes	yes
Article Length	yes	yes	yes	yes	yes	yes
Abstract Length	yes	yes	yes	yes	yes	yes
Title Length	yes	yes	yes	yes	yes	yes
Article Order	yes	yes	yes	yes	yes	yes
Num Authors	yes	yes	yes	yes	yes	yes
Author Gender	yes	yes	yes	yes	yes	yes
Num References	yes	yes	yes	yes	yes	yes
Article Type	yes	yes	yes	yes	yes	yes

Given interest in the predictive power of models with different numbers of predictors, model comparisons use predicted R2 values from out-of-sample 10-fold cross validation (90% training and 10% test). Results also replicate performing in-sample comparisons with adjusted R2 values.

201 ficial (because it increases the perception that a phenomenon
 202 is true) or detrimental (if it seems unwarranted; 19). Using
 203 more familiar language may help if it makes things easier to
 204 read. Compared to using definite articles (i.e., “the”, which
 205 specifies a singular, identified member of the type, e.g., violat-
 206 ing the norm of ...) using indefinite articles (i.e., “a” or “an”
 207 which means that any member of that type is being discussed,
 208 e.g., violating a norm of ...), may make content seem broader
 209 and generally applicable, which may increase citations. Lan-
 210 guage’s impact may also vary by discipline. While jargon may
 211 generally decrease readability (20), and thus citations, it may
 212 increase impact in disciplines where it is seen as a signal of
 213 credibility.

214 Finally, this work highlights the value of using natural
 215 language processing to study culture. Cultural items like aca-
 216 demic articles, songs, books, and movies often succeed or fail,
 217 but understanding why requires being able to quantify their
 218 underlying features or dimensions. Automated text analysis
 219 makes this possible, and as a result, will hopefully unlock a
 220 range of interesting insights.

221 **Materials and Methods**

222 We compiled a corpus of 28,988 full-text peer-reviewed articles
 223 from 22 different journals from 1990-2018 (full data and method
 224 details are presented in Supplementary Information). Writing style
 225 was measured using the incidence rate (proportion of words) of
 226 each of the nine categories of function words (i.e., auxiliary verbs,
 227 conjunctions, grammatical articles, impersonal pronouns, negations,
 228 personal pronouns, prepositions, quantifiers, and common adverbs)
 229 in each article, and we collected the number of citations each article
 230 achieved. We control for a variety of factors beyond the article text
 231 that are linked to citations, and given certain content (i.e., research
 232 topics or areas) might be cited more (19), we control for this as
 233 well. To examine whether style matters, we test whether beyond
 234 the variance explained by article content and non-language features,
 235 including style features adds explanatory power (i.e., helps explain
 236 the number of citations articles receive).

237 **ACKNOWLEDGMENTS.** Acknowledgement

- 238 1. EM Rogers, Elements of diffusion. *Diffusion innovations* 5 (2003).
 239 2. AP Association, *Publication manual of the American psychological association.* (American
 240 Psychological Association Washington, DC), (2020).

3. S Pinker, Why academics stink at writing. *The chronicle higher education* 61 (2014). 241
 4. DS Hamermesh, P Schmidt, The determinants of econometric society fellows elections. *Econometrica*, 399–407 (2003). 242
 5. RH Baayen, R Piepenbrock, L Gulikers, The celex lexical database (cd-rom) (1995). 244
 6. ME Ireland, JW Pennebaker, Language style matching in writing: Synchrony in essays, cor- 245
 7. C Chung, JW Pennebaker, The psychological functions of function words. *Soc. communica- 246*
 8. RTW Lo, B He, I Ounis, Automatically building a stopword list for an information retrieval sys- 247
 9. D Biber, B Gray, K Poonpon, Should we use characteristics of conversation to measure gram- 248
 10. JW Pennebaker, LA King, Linguistic styles: language use as an individual difference. *J. 249*
 11. JW Pennebaker, MR Mehl, KG Niederhoffer, Psychological aspects of natural language use: 250
 12. DJ Bem, Writing the empirical journal in *The compleat academic: A practical guide for the 251*
 13. W Strunk Jr, E White, The elements of style (1999). 252
 14. Nature, Writing scientific papers (2014). 253
 15. N Liberman, Y Trope, SM McCrea, SJ Sherman, The effect of level of construal on the tem- 254
 16. B Freeling, ZA Doubleday, SD Connell, Opinion: How can we boost the impact of publica- 255
 17. A Ruben, How to read a scientific paper (2016). 256
 18. B Mensh, K Kording, Ten simple rules for structuring papers (2017). 257
 19. DL Willis, CD Bahler, MM Neuberger, P Dahm, Predictors of citations in the urological litera- 258
 20. JW Pennebaker, RL Boyd, K Jordan, K Blackburn, The development and psychometric prop- 259
 260
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272