2.4 Providing Quality Score Feedback to Revise Content

To explore the incremental value of using NLG relative to SEO experts with access to the quality score further, we conducted the following additional study. Following the content that was initially produced by the real SEO experts we offered them the opportunity to improve their content by providing them with the quality scores of their initial content and the real search engine ranked top 10 content. The experimental setup mirrors A/B testing and is as follows: The tests were conducted using an online survey (including personal explanations of the task and a Q&A section). The survey contained the task description, the principal investigator’s contact details, and an incentive of 40€ for the revision / feedback round per piece of content. The participants were introduced to the quality score and provided with an explanation of each quality score component and how to interpret it. They were provided with the quality score (both on each component and overall) for the content they produced initially. For comparison purposes, they were also provided with the top 10 ranked content for the specific keyword and their associated quality scores. Study participants entered their revised text in an open text field. We extended the original study reported in the main manuscript using 30 keywords for the real SEO experts (instead of just 9), so the testing was conducted for 30 pieces of SEO expert produced content.

Table W2.4.1 shows that the SEO experts changed their original content by 10.24% (~77.50 words), ranging between 12 words changes and 176 word changes. Table W2.4.2 compares the achieved quality scores of the original SEO experts’ content to the revised SEO experts’ content for each quality score component. We find no statistically significant differences between them, suggesting that the SEO experts were not able to improve the quality of their content, likely due to the associated complexity (i.e., dozens of word distributions, numbers, and abstract concepts). This suggests that the semi-automated procedure not only reduces the time/cost associated with content production, but also performs better than human experts on tasks involving the generation of content for a specific purpose.

Table W2.4.1: Descriptives for Real SEO Experts Content Revision

Dimension	Groups	Descriptives			
		Median	(IQR)	Min	Max
Produced content length (in words)	Original SEO Expert Content	729.5	(47.25)	587	819
	Revised SEO Expert Content	760.5	(58.75)	546	930
Content change (original vs. revised)[1]	Change in %	10.24	(5.84)	1.62	24.24
	Change in words	77.50	(49.25)	12	176

[1] This includes every possible change between the original SEO experts and revised SEO experts content such as added words, deleted words, and words with at least one changed letter (including changed letter capitalization).

Table W2.4.2: Quality Score: Original Real SEO Experts Content vs. Revised Real SEO Experts Content

Quality Score Component	**Group**	**Descriptives**			**Wilcoxon rank sum**[1]			
		Median (IQR)	Min	Max	W	z	*r*	*p*
Topic (s_a)	Original SEO Experts	**.41 (.18)**	**.30**	**.72**				
	Revised SEO Experts	.44 (.18)	.30	.72	423	.39	.05	.697
Keywords (s_k)	Original SEO Experts	**.46 (.20)**	**.16**	**.79**				
	Revised SEO Experts	.46 (.15)	.16	.80	423	.39	.05	.697
Uniqueness (s_d)	Original SEO Experts	**.97 (.06)**	**.08**	**1.00**				
	Revised SEO Experts	.96 (.07)	.63	.99	560	-1.62	-.21	.105
Readability similarity (s_r)	Original SEO Experts	**.72 (.54)**	**.02**	**1.00**				
	Revised SEO Experts	.66 (.71)	.02	1.00	469	-.27	-.04	.784
Naturality similarity (s_n)	Original SEO Experts	**.67 (.39)**	**.08**	**1.00**				
	Revised SEO Experts	.58 (.25)	.17	1.00	477.5	-.40	-.05	.687

[1] Two-tailed tests between original vs revised real SEO experts quality scores, statistical significance codes: *0.05 level, **0.01 level;

Appendix References

Baayen RH, Shafaei-Bajestan E (2019) Analyzing linguistic data: A practical introduction to statistics. Package ‘languageR’. Version 1.5.0. *CRAN*. Accessed May 20, 2019, https://cran.r-project.org/web/packages/languageR/languageR.pdf

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A, (2018) “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*. 3(30). https://doi.org/10.21105/joss.00774

Berger J, Sherman G, Ungar L (2020b) TextAnalyzer. Accessed November 11, 2020, http://textanalyzer.org

Bronnenberg BJ, Kim JB, Mela CF (2016) Zooming in on choice: How do consumers search for cameras online? *Marketing Science*. 35(5):693-712.

Danaher PJ, Mullarkey GW, Essegaier S (2006) Factors affecting website visit duration: A cross-domain analysis. *Journal of Marketing Research*. 43(2):182-194.

Edelman B, Zhenyu L (2016) Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research*. 53(6):881-900.

Flanigan, AJ, Metzger, MJ (2007) The role of site features, user attribtues, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*. 9(2):319-342. https://doi.org/10.1177/1461444807075015

Jerath K, Ma L, Park YH (2014) Consumer click behavior at a search engine: The role of keyword popularity. *Journal of Marketing Research*. 51(4):480-486.

Kamoen N, Holleman B, Bergh H (2013) Positive, negative, and bipolar questions: The effect of question polarity on ratings of text readability. *Survey Research Methods*. 7(3):181-189.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*. 37(6):930-952.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao ELT, Palma MA (2020) Basic robust statistics. Package ‘robustbase’. Version 0.93-6. *CRAN*. Accessed May 20, 2020, https://cran.r-project.org/web/packages/robustbase/robustbase.pdf

Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2015) Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates. Accessed November 1, 2020, www.LIWC.net.

Pitler E, Nenkova A (2008) Revisiting Readability: A unified framework for predicting text quality. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 186-195.

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. OpenAI.

Roberts C (2010) Correlations among variables in message and messenger credibility scales. *American Behavioral Scientist*. 54(1):43-56.

Rocklage MD, Rucker DD, Nordgren LF (2018) Persuasion, emotion and language: the intent to persuade transforms language via emotionality. *Psychological Science*. 29(5):749-760.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomze AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. 1-15.