CAMBRIDGE
UNIVERSITY PRESS

COMMENTARY

Descriptive statistics and advanced text analytics: A dual extension

Emily D. Campion[1]* and Michael A. Campion[2]

[1]Old Dominion University, Strome College of Business and [2]Purdue University, Krannert School of Management
*Corresponding author. Email: ecampion@odu.edu

The main tenet of Murphy's (2021) argument is that we should be leveraging the valuable descriptive information that is offered in Table 1 (sample sizes, means, standard deviations, and intercorrelations) and not relegate it to a single line in our manuscripts. Whereas Murphy focused largely on the value of this information for interventions, moderators, and mediators (presumably used in regression-based models), our goal in this commentary is to *extend* his credo in two ways: (a) applying it to the quickly emerging method of advanced text analytics and (b) using advanced text analytics to systematically analyze and include more narrative sample and contextual information to bolster a study's descriptive statistics.

Extension 1: Advanced text analytics

Advanced text analytics broadly refers to a family of techniques that is used to identify and summarize topics, themes, or other dimensions of interest in a corpus of text using computer software (e.g., SPSS Modeler, Python). Two of the most well-known advanced text analytics algorithms are latent Dirichlet allocation (Antons et al., 2019; Blei et al., 2003; Hannigan et al., 2019) and its predecessor latent semantic analysis (Landauer et al., 1998).

In practice and academia, advanced text analytics serve a range of purposes from discovering previously uncaptured constructs and offering triangulated measurement of existing constructs to providing a quicker and more cost-effective method of summarizing text data. Text data can be sourced directly from workers via open-ended questions in engagement surveys, candidate application materials (Campion et al., 2016), employee performance narratives (Speer, 2018), organizational websites (Banks et al., 2019), social media (e.g., Twitter; Min et al., 2021; Zhang et al., 2021), and the like. Advanced text analytics and machine learning have become some of the most popular topics at the Society for Industrial and Organizational Psychology (SIOP) annual conferences and have appeared in some fashion in SIOP's annual Top 10 Trends list since the list began in 2014. Now is the time to encourage those who use advanced text analytics to report traditional descriptive statistics of their text data.

Advanced text analytics, and machine learning more broadly, have been charged with being too much of a "black box," meaning that we vaguely understand how the algorithms are depicting the relationships among variables. This concern is largely justified—most industrial-organizational (I-O) psychologists are not trained on machine learning techniques, and those in the fields who developed machine learning are not trained in traditional statistical theories and psychometric methods that would afford translation among the disciplines. Moreover, machine learning's sole focus on prediction to the neglect of interpretability reinforces Murphy's (2021) concern about complex analyses.

Yet, those who develop such models regularly know the truth is simpler. We can use our traditional methods to generate the statistics that are reportable in Table 1 to develop a deeper understanding of what our text models are capturing. Put differently, by translating this complex analysis into I-O language, such as the descriptive statistics in Table 1, we can help make the "black box" more translucent and demystify advanced text analytics in the process.

For this extension to Murphy's (2021) advocacy of Table 1, we recommend the following:

1. Include all text variables in Table 1. We suspect all researchers who conduct advanced text analytics do this already.
2. Include means and standard deviations of all first-order text variables (often referred to as "features" or "concepts," depending on your software and training) in Appendix Table A1 or an online supplement. When conducting an advanced text analysis, our focal text variables often comprise many lower-order text variables and we should include these first-order text variables in an appendix to improve interpretation and transparency. This is similar to how we routinely aggregate Likert-type items to create measures of constructs and report those items in appendices.
3. Include zero-order correlations between the features and the criterion variable in Table A1 or an online supplement. This is particularly important in studies where the correlations between the first-order text variables and criterion are used to determine first-order text variable retention. Those who are unfamiliar with advanced text analytics should expect small means and correlations. This is similar to the relationships we observe between individual test questions and criteria: The individual correlations are small and the gains are accrued when they are aggregated into a measure. Nevertheless, these small correlations have interpretive value in understanding the nature of the text variables and those that are most predictive.
4. Conduct traditional psychometric quality analyses of the text variables, such as internal consistency reliability, and perhaps item analyses and factor analyses, as appropriate. Variables that are created from advanced text analysis made up of features are measures of constructs just like other psychological measures made up of survey items or test questions, so our psychometric analyses apply and will make the description of the variables more interpretable. These results should be reported in either the manuscript or an appendix.

Extension 2: Narrative data

Narrative data typically refer to text data, which can complement numeric data by contextualizing or otherwise improving understanding of the phenomenon under study by describing it in words. These underused data are valuable because they are in the respondent's own words and unfiltered by the imposition of the researcher's language—unlike survey items that are usually in the researcher's words. Thus, the data are descriptive statistics in the sense of describing the phenomenology of what is being studied and can perhaps be as descriptive as means and standard deviations from traditional rating scales. Narrative data are typically collected in the practice and science of I-O psychology.

For example, most consultants know that clients often understand their problems deeply and may already have solutions. Consultants can bring value by simply listening, summarizing, and presenting the narrative data back to clients. Similarly, researchers sometimes begin by conducting informational interviews or collecting open-ended responses to help them ground their research in the phenomenon, identify the appropriate constructs and theory, and develop survey items or other standardized measures. However, they rarely report the raw data. These narrative data are important to report in any study, at least in a summary format.

The inclusion of narrative data is less common in psychology journals than in management journals, and this may be because such methods are viewed as less scientifically rigorous. Yet, if analyzed in a systematic, replicable way using advanced text analytics (extension 1), we can demonstrate more rigor because all aspects of data collection and analysis can be explicitly described such that other researchers can critique and/or replicate it. Researchers may also consider reporting simple quotes as illustrative of the categories that are derived from advanced text analysis to provide additional descriptive richness. Furthermore, narrative data that are subjected to advanced text analysis are crucial not only to the primary study but also to future systematic reviews that can use such descriptive data to test for contextual relationships in meta-analyses and offer opportunities for theoretical expansion in narrative systematic reviews.

For this extension to Murphy's (2021) advocacy of Table 1, we recommend the following:

1. Most research studies and consulting projects start by asking subject matter experts about the nature of the phenomenon. They often have a deep understanding of what is happening and can save researchers time. This should be part of every study, perhaps especially academic studies where it is less common.
2. Every study should ask respondents directly about the phenomenon under investigation, and the narrative data should be analyzed and reported in Table 1 or an appendix, even when the target variables are not derived from text data.
3. Researchers should analyze these data using the same rigor as any analysis to ensure transparency and replicability. With advanced text analytics so readily available, this should be no more difficult than describing other analyses.

Concluding remarks

We learn the most from our descriptive statistics. It's what we use to ensure that our data are clean, it's what we look forward to examining as our initial interpretation of the data, and it's what many of us often focus on when we review or read an article. We echo Murphy (2021) and ask our fellow practitioners and researchers to not neglect what descriptive statistics on narrative data can offer, and we encourage them to include more information on their text variables and the nature of their study's context and phenomenon through the words of respondents to improve our accumulated knowledge.

References

Antons, D., Joshi, A. M., & Salge, T. O. (2019). Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*, **45**(7), 3035–3076. https://doi.org/10.1177/0149206318774619

Banks, G. C., Woznyi, H. M., Wesselen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management*, **45**(2), 476–509. https://doi.org/10.1177/0149206318764295

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, **101**(7), 958–975. https://doi.org/10.1037/apl0000108

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, **13**(2), 586–632. https://doi.org/10.5465/annals.2017.0099

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. *Journal of Applied Psychology*, **106**(2), 214–229. https://doi.org/10.1037/apl0000886

Murphy, K. (2021). In praise of table 1: The importance of making better use of descriptive statistics. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **14**(4), 461–477.

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, **71**(3), 299–333. https://doi.org/10.1111/peps.12263

Zhang, C., Yu, M. C., & Marin, S. (2021). Exploring public sentiment on enforced remote work during COVID-19. *Journal of Applied Psychology*, **106**(6), 797–810. http://dx.doi.org/10.1037/apl0000933