Words of Estimative Probability in Academic Text

Megan Church

The College of William & Mary

Abstract

Words of Estimative Probability (WEP's) are terms such as, "frequently" and "likely", used to convey the likelihood of an event. Although several times have been made to standardize a paradigm for WEP's, there has been a little progress. there is no current literature on the use of WEP's in academic writing. This paper takes a first step at exploring the frequency with which WEP's occur in academic text. A subset of PubMed central journal articles related to Educational Psychology were tokenized for analysis. The results suggest that WEP's are not frequently used in academic text, however, when and how they are used goes unexplored. Future applications include exploring, text, from, various industries, the implications of WEP's on psychology's replication crisis, using vector embeddings to provide insight into context, and predictive modeling to assign numerical values to WEP's.

Words of Estimative Probability in Academic Text

In the early 1950s, the CIA's Office of National Estimates released a statement saying there was a "serious possibility" of a Soviet attack on Yugoslavia within the year. A higher-up at the Office of National Estimates, Sherman Kent, was unsure what a "serious possibility" meant, but he interpreted it as around a 65% chance of attack. However, when he asked board members, he got responses ranging from 20% to 80% (Mauboussin & Mauboussin, 2020). This wide range of interpretations is not ideal for such a serious matter.

In an attempt to clarify the meaning of these probabilistic terms, or, Words of Estimative Probability (WEP's), Kent mapped the relationship between these terms and their assumed probabilities. He showed sentences that included probabilistic terms to around 20 military officers and asked them to translate the terms into numbers. Some words were assigned similar numerical values across all participants; however, other terms resulted in a wide range of assigned numerical values (Mauboussin & Mauboussin, 2020). This finding eventually led Kent to develop a paradigm relating WEP's to numerical odds as shown in Table 1 (CIA, 2002, Sherman Kent).

Kent is not the only advocate for standardizing WEP's. Several studies have successfully replicated his results (Barnes, 2016; Mosteller & Youtz, 1990), leading to the development of new paradigms. The National Intelligence Council recommended the use of a paradigm similar to Kent's (Table 2) which combines confidence levels with the scope and quality of supporting information (Prospects, 2007). After three decades of data collection, The Mercyhurst College Institute for Intelligence Studies developed their own paradigm (Table 3), which reduces Kent's schema to its least ambiguous words (Wheaton, et al., 2008). Despite these efforts, there has been repeated failure to implement a standardized method of reporting using WEP's in the intelligence community (Blair, 2004; Schrage, 2005).

A disconnect between wording and implied numerical value is not unique to the intelligence community. We continue to utilize vague, probabilistic terms in our discussions on more general, although still important, matters such as business and weather. In fields like politics, where “vague verbiage” provides safety, probabilistic words are almost expected. However, other fields, like academia, value explicit and precise wording. This paper documents the methods used to take an initial look at the use of WEP’s in academic writing and explores areas of further research and potential application.

Methodology/Dataset

A list of WEP’s was curated for analysis (Appendix B). Terms form each of the paradigms previously discussed were included as well as more common terms identified in The Harvard Business Reviews’ article “If You Say Something Is “Likely,” How Likely Do People Think It Is?” (2018).

The corpus consists of 1140 PubMed central journal articles resulting from the search string “Educational Psychology”. These articles were scraped from the PubMed Central Open Access Subset, which contains journal articles that have been made available under creative comments licenses that allow reuse (PMC, 2003). The journal articles were scraped using selenium (version 4.1.0) and parsed using beautiful soup (version 4.12. 0). The tags collected were the article title, date, and body text. The body text was then broken into sections, based on the section headings in the HTML. For example, Abstract, Introduction, Methods, Results, etc.

The raw data consisted of a PMCID, article title, date published, and full body text. The article title was dropped for pre-processing and the publication year was extracted from the date using regular expressions. The body text for each article was separated by section heading. For example, the abstract, introduction, and methods section were all split and labeled accordingly. This was done using a series of regular expressions. Due to the structure of the body text, in some cases, the last word of the section header was left in the body text. For example, "ParticipantsParticipants were recruited…". In this case, it is assumed that the first word of a section will not be a WEP.

The separated raw text was then pre-processed by tokenizing, lemmatizing, and removing stop word using the Spacy library. Spacy's stop words include some terms of interest, "always, often, and never", so these were removed from Spacy's list of stop words before preprocessing began. Punctuation was removed using Pythons string library. The tokens were then exploded and subset to only include the key WEP's, PMCID and section still attached. The term frequency was then calculated for each term by document (PMCID) and section.

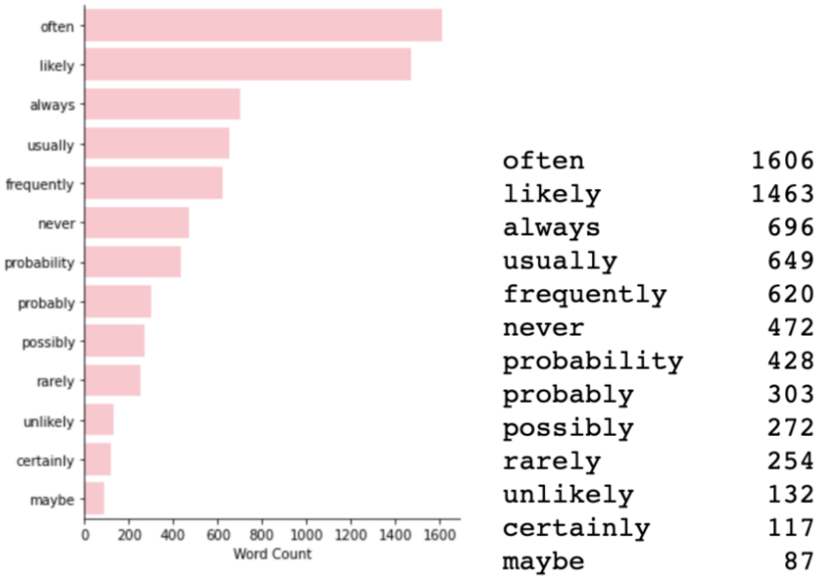
Results

"Often" and "likely" were used significantly more than other WEP's, with frequencies of 1610 and 1471, respectively. Figure 1 displays all of the term frequencies. Most documents contain about five WEP's, as shown in Figure 2. The document with the highest use of WEP's was titled *"Meta-Intelligence: Understanding, Control, and Interactivity between Creative, Analytical, Practical, and Wisdom-Based Approaches in Problem Solving"* and contained 28 key WEP's. Additionally, eight of the 1140 documents (3.6%) do not contain any key WEP's.

Analysis could not be performed using the parts because they did not always correspond to a traditional paper section. Even after normalizing the parts using these key words (abstract, introduction, method, result, discussion, conclu, future, limitation, footnote, background, analysis), there were 1208 unique sections.

Figure 1

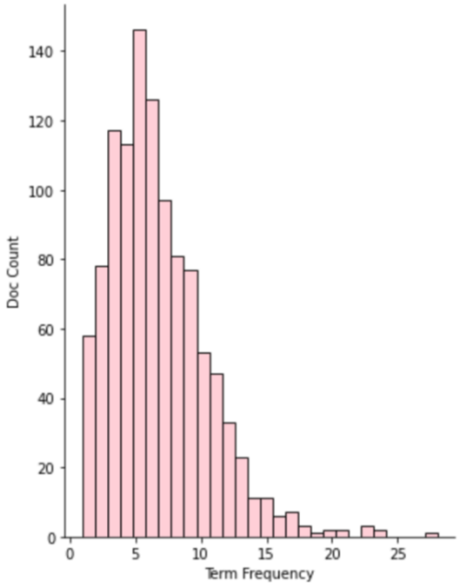
Distribution of Probability Terms



```
often          1606
likely         1463
always          696
usually         649
frequently      620
never           472
probability     428
probably        303
possibly        272
rarely          254
unlikely        132
certainly       117
maybe            87
```

Figure 2

Distribution of Probability Term Frequency by Document



Discussion

The results suggest that WEP's are used, but not frequently, in academic text relating to educational psychology. However, the terms that are used the most, "often" and "likely", are considered relatively easy to interpret.

Since this study appears to be the first of its kind, there is no data to compare these results to. Further analysis of different academic fields, and different types of text would need to be analyzed before any strong conclusions can be reached.

Conclusion

This study explored the use of WEP's in academic writing. Although academic text has not yet been subject to this particular flavor of analysis, there is evidence that WEP's may be overused in situations where explicitness is required(cite). At present, the study is limited to journal articles pertaining to educational psychology. The next step in this research would be to look at articles from different academic fields compare their use of WEP's to that of educational psychology's. Additionally, the analysis of different writing forms would allow for comparison across industry (i.e. business reports, white papers, and policy briefs).

The psychological sciences are big advocates of study replicability, in part due to the failure to replicate some major, field defining studies (Sussex Publishers). Although several factors contribute to this failure, the inability to reproduce older studies is often attributed to small sample size, vague reporting of the methodology, and incomplete statistics. These studies are often over half a century old and were not subject to the rigorous standards of the American Psychological Association that are now in place. Comparison between the use of probabilistic terms in replicable, and non-replicable studies may provide some insight into the difference in reporting techniques used.

WEP's can be used in a variety of contexts, some that do not necessarily require a numerical value. For example, this paper contains the probabilistic term "often" four times. While some instances probably should be replaced with a concrete value (ex. "These studies are often over half a century old…"), there are some occasions when words like "often" get the job done (ex. This failure is often attributed to …). Vector embedding can provide further insight into when and how probabilistic terms are used. Clustering the WEP's based on their semantic similarity could help identify patterns in how

WEP’s are used in different contexts. Vector embedding can also be used to identify the words or phrases that are most closely associated with each WEP, and tag if they would be better off being replaced with a numerical value.

Another avenue for exploring this, and similar, datasets would be creating a predictive model that inputs some amount of text and outputs the implied numerical value of the probabilistic terms used in the text. This model would be a combination of a machine learning model, trained on a corpus to identify probabilistic terms, and a rule-based system that assigns a numerical value to each phrase. These two approaches would need to work in tandem due to the wide range of assumed numerical values some probabilistic phrases assume. This model could then be used to bring clarity to WEP’s in cases where their assumed probability is unclear.

References

Alan Barnes (2016) Making Intelligence Analysis More Intelligent: Using Numeric Probabilities, Intelligence and National Security, 31:3, 327-344, DOI: 10.1080/02684527.2014.994955

Blair, Bruce (2004), *The Logic of Intelligence Failure* (PDF), Center for Defense Information, p. 11, retrieved 2008-04-21

Mauboussin, A., & Mauboussin, M. J. (2020, October 29). *If you say something is "likely," how likely do people think it is?* Harvard Business Review. https://hbr.org/2018/07/if-you-say-something-is-likely-how-likely-do-people-think-it-is

Mosteller, F., & Youtz, C. (1990). Quantifying Probabilistic Expressions. *Statistical Science, 5*, 2-12.

PMC Open Access Subset [Internet]. Bethesda (MD): National Library of Medicine. 2003 – [cited YEAR MONTH DAY]. Available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

"Prospects for Iraq's Stability: A Challenging Road Ahead" (PDF), *What We Mean When We Say: An Explanation of Estimative Language*, National Intelligence Council, p. 5, 2007, retrieved 2008-04-21

Schrage, Michael (20 February 2005), "What Percent Is 'Slam Dunk'? Give Us Odds on Those Estimates", *Washington Post*, p. B01

Sherman Kent and the Profession of Intelligence Analysis, Center for the Study of Intelligence, Central Intelligence Agency, November 2002, p. 50, archived from the original on June 12, 2007

Sussex Publishers. (n.d.). *Replication crisis*. Psychology Today. https://www.psychologytoday.com/us/basics/replication-crisis

Wheaton, K.; Chido, Diane (2008), "Words of Estimative Probability", *Competitive Intelligence Magazine*, Alexandria, VA: Society of Competitive Intelligence Professionals, ISSN 1521-5881

Appendix A

Table 1

Kent’s Words of Estimative Probability

Certain	100%	Give or take 0%
Almost Certain	93%	Give or take about 6%
Probable	75%	Give or take about 12%
Chances About Even	50%	Give or take about 10%
Probably Not	30%	Give or take about 10%
Almost Certainly Not	7%	Give or take about 5%
Impossible	0	Give or take 0%

Table 2

National Intelligence Council Probabilistic Terms

Almost Certainly
Probably/Likely

Even Chance
Unlikely
Remote

Table 3

Mercyhurst Probabilistic Terms

Almost Certain
Highly Likely
Likely/Probable
Unlikely
Almost Certainly Not

Appendix B

Words of Estimative Probability used in the Study

'always',

'usually',

'certainly',

'likely',

'frequently',

'probably',

'often',

'maybe',

'possibly',

'probability',

'unlikely',

'rarely',

'never'