Literature Adventures with Linguistic Inquiry and Word Count

Ms. Kristin L. Schaefer P.E., UH

Kristin Luthringer Schaefer is a licensed professional engineer (PE) and a certified secondary teacher (grades 6-12), both in Texas, as well as the owner of her own consulting firm, Schaefer Engineering. She obtained both her bachelor's and master's degrees in Mechanical Engineering (ME) from Texas A&M University (TAMU) and is currently pursuing a doctorate in ME at the University of Houston (UH). Her Ph.D. research interests are in STEM education, especially with underrepresented students of all ages, STEM mentors, and their motivations and/or persistence. The first part of her career was spent designing residential split system HVAC equipment and Indoor Air Quality (IAQ) unitsfor Trane in Tyler, TX. Kristin has taught about design, engineering, and manufacturing to students of all ages in various places including to preschoolers via Schaefer Engineering's STEM outreach, to senior mechanical engineering undergraduates at TAMU, to eighth graders in KatyISD at Beckendorff Junior High, and to freshmen mixed major undergraduates at UH. Kristin is also the mom of one smart teenage boy whose journey through learning differences and Type 1 Diabetes (T1D) has enabled her to connect with and support students with a broad spectrum of learning preferences.

Jorge Rosales

Dr. Jerrod A. Henderson, University of Houston

Dr. Jerrod A. Henderson ("Dr. J") is an Assistant Professor in the William A. Brookshire Department of Chemical and Biomolecular Engineering in the Cullen College of Engineering at the University of Houston (UH).

He began his higher education pursuits at Morehouse College and North Carolina Agricultural & Technical State University, where he earned degrees in both Chemistry and Chemical Engineering as a part of the Atlanta University Center's Dual Degree in Engineering Program. While in college, he was a Ronald E. McNair Scholar, which afforded him the opportunity to intern at NASA Langley. He also earned distinction as a Phi Beta Kappa member and an American Chemical Society Scholar. Dr. Henderson completed his Ph.D. in Chemical & Biomolecular Engineering at the University of Illinois at Urbana-Champaign. During his time as a graduate student, he was a NASA Harriet G. Jenkins Graduate Fellow.

Dr. Henderson has dedicated his career to increasing the number of students who are on pathways to pursue STEM careers. He believes that exposing students to STEM early will have a lasting impact on their lives and academic pursuits. He is the co-founder of the St. Elmo Brady STEM Academy (SEBA). SEBA is an educational intervention aimed at exposing underrepresented fourth and fifth-grade students and their families to hands-on STEM experiences. Henderson's research interests are in engineering identity development among Black men and engineering student success. He was most recently recognized by INSIGHT Into Diversity Magazine as an Inspiring STEM Leader, the University of Illinois at Urbana-Champaign with the College of Liberal Arts & Sciences (LAS) Outstanding Young Alumni Award, and Career Communications Group with a Black Engineer of the Year Award for college-level promotion of engineering education.

Literature Adventures with LIWC (Work-in-Progress)

1 Introduction and Purpose

A thematic literature review was conducted to inform a dissertation project that is using qualitative methods to study the experiences of female students and early-career engineers who participated in enrichment programs as middle school students and have persisted in engineering. That study seeks to understand if outreach could be a reason why 20% of bachelor's degrees awarded to women in 2010 was approximately fourteen thousand, but in 2020 was around thirty-three thousand [1], [2]. The study followed five commonly accepted systematic literature review steps, enabling a thorough and rigorous search for prior work on the topic [3], [4, Ch. 6]. To assist with step 2 of that literature review, we used SPIDER searching criteria (refer to Table 1 in Appendix A) to acquire an initial search return from the databases of Scopus, EBSCO, ProQuest, the university library search tool, and known-to-be-applicable articles [5]. Typical PRISMA filtering methods yielded twenty-seven original research and six review papers applicable to engineering identity (eID), female persistence in engineering, and/or middle school outreach [6]. Still, we wanted to determine if a faster path to evaluating literature was available via a novel thematic analysis methodology and the use of computational tools.

1.1 Computational Text Analysis: LIWC

The workings of the Linguistic Inquiry and Word Count (LIWC; pronounced "Luke") program allow it to analyze the language of a text file and categorize it by various attributes [7]. The traditional LIWC analysis has a "top-down" procedure that provides quantitative summary dimensions such as total word count, the number of "big" words (longer than 6 letters), and percentage of the dictionary's words appearing in the text, as well as various psychological categories, using both standard and custom dictionaries [8]. The psychological portion using the standard dictionary reports four summary measures (analytical thinking, clout, authenticity, and emotional tone) and nine dimensions (linguistic, drives, cognition, affect, social, culture, lifestyle, physical, perception, and conversation).

The four LIWC summary measures provide their analysis of the text as a dichotomy comparison from a normalized percentage of several variables [7]. Lower analytical thinking scores indicate more informal, personable writing and thinking styles, whereas higher analytical thinking scores indicate more logical, rigid writing and thinking styles [9]. Lower clout scores indicate more of a self-focus, a "follower" not caring as much about relative social status, whereas higher clout scores indicate a "leader" with more focus on dominating the others in a group [10]. While lower authenticity scores can reflect a measure of deception, they also indicate a prepared or socially cautious response, whereas higher authenticity scores indicate more spontaneous, complex, honest, and unfiltered conversations [11], [12]. Lower emotional tone scores indicate a more negative attitude, whereas higher emotional tone scores indicate a more positive outlook in the text [13]. LIWC provides their licensees with a copy of the words within each of their nine dimensions in a human-readable chart. Typically, linguistic words are "functional" such as pronouns and articles, drives words deal with motivation, cognition words include insight and memory, affect words contribute to emotions or a positive / negative tone, and social words deal with relationships. Both culture words and physical words deal with identity and human needs, but in slightly different ways. Lifestyle words are about work and leisure, perception words deal with senses and time, and conversation words include informal language like filler words. Our

team's final custom dictionaries analyze themes related to typical engineering education (EnEd) aspects of identity and persistence with thirty-eight dimensions and aspects of community cultural wealth (CCW) with three dimensions, refer to Table 5 in Appendix A for more details, including how this study updated the dimensions and categories.

LIWC also has a "bottom-up" topic modeling procedure called the meaning extraction method (MEM analysis) that counts how many times words or phrases are used and in how many documents and outputs a frequency table for the meaningful words within the text [14]. The binary MEM table specifies which document contains what word, but we found the MEM frequency table to be the more useful output.

Our initial methodology for LIWC used both of these procedures, as they parallel the EnEd discipline constructs of *a priori* and *in vivo* coding, respectively [4, Ch. 8]. We postulated that LIWC could assist with understanding the themes used in EnEd as it opens multiple language analysis methods within a short timeframe.

1.2 Purpose, Positionality, and Research Questions

The first and last authors are not only education researchers, but also engineers at heart. We have the intersectional identities of White female engineering graduate student and Black male engineering faculty, respectively [16]. Our dual purposes were to utilize a novel methodology within the literature review process and to bring the second author, a Hispanic male engineering undergraduate student, into EnEd research.

The research question for this work is divided into two subparts:

(A) How does the use of LIWC affect the themes determined for a topic?

(B) What aspects of advice or other recommendations can be gleaned for future literature review improvements using computational text analysis tools?

Additionally, we felt it was important from a validity standpoint to do this work checking the LIWC methodology on the six applicable systematic literature review papers identified during the SPIDER search. By testing both LIWC analysis methods as well as our custom dictionary on this smaller sample, if any modifications were needed in our methodology or dictionary, we could adjust before embarking upon the thematic analysis of the twenty-seven papers [4, Ch. 10], [15, Ch. 6].

2 Methods

This section provides insight into the novel review analysis methodology using computerized text analysis.

2.1 Locating Resources

During Step 3 of the literature review we obtained the remainder of the articles [3]. All were saved in PDF format for reading by our team.

We chose to use these six review papers because two are seminal works on engineering identity [17], [18]; three papers have valuable lessons learned about the interaction between Middle

School STEM outreach and female STEM identity [19]–[21]; and the sixth one was returned via the SPIDER search terms in the ProQuest database [22].

2.2 Data Preparation

The typical synthesizing methodology for a literature review is reading and open coding, as well as gleaning study data, which we did. To attempt a novel method of analyzing the literature, we chose to use LIWC to rapidly find the salient themes of the primary studies set. Thus, we converted the set of literature PDF files into text (TXT) files for LIWC.

It was critical to include only the bodies of the articles, along with figure captions and table text, in the TXT files for analysis by the program. While LIWC has the capability to analyze PDF files, the references, titles, and abstracts from each paper were excluded from the TXT files generated for the LIWC analysis due to several reasons. We perceive that abstracts are applicable to screening purposes, not meaning analysis. References and titles were excluded because the word count could be misconstrued if they were included.

3 Results and Discussion

This section provides the results from using LIWC on six literature review articles.

3.1 "Top-Down" (*a priori*) Traditional LIWC Analysis

This type of LIWC output gives a dimension as the percentage of words within the text belonging to a word list, as well as providing the total Word Count (WC, the raw number of words), and the Words Per Sentence (WPS, the mean length of the sentences) [8]. Summary measure information for all six papers is in Table 6 in Appendix A [17]–[22]. An average of 79% of the words in all six papers were found in the standard LIWC dictionary. Unsurprisingly, the six review papers were very high in the analytic summary measures, at 85% or greater for each paper, and low in the authenticity summary measures, at an average score of 30% with a range from 14.31% to 38.98%. These measures are reflective of the scholarly nature of peer-reviewed journal articles as they are expected to be logical, prepared works. Both clout and emotional tone scores varied widely, as would be expected for works from different authors with varying positionalities. Roughly half of the words in all the literature review papers were linguistic in nature (articles, prepositions, pronouns, verb/adverbs, number/quantity, and conjugates). Notable dimensions present in eID literature are: lifestyle words surrounding the work dimension; cognition words from the cognitive processes dimension; social words around social behaviors, family, and gender identity; and perception words surrounding spatial and time dimensions. All these are as expected for eID in girls, with the exception of work, which relates more to careers and persistence. Refer to Table 7 for the full standard dimension information.

Shifting to how the standard LIWC dictionary compares to our custom eID and Persistence dictionary, we begin by stating that it was indeed necessary to update our custom dictionary upon review of its first run through LIWC and comparison with the standard dictionary results. We perceive that the drive and perception variables from the standard dictionary relate to our persistence category, the culture and physical variables relate to identity, and cognition, social and lifestyle variables relate to both persistence and identity. In Appendix A, refer to Table 5 for the changes in our dimensions, including how some word groupings were combined and others separated, including a separate dictionary file for CCW. Also refer to Table 9 for lessons learned for future custom dictionaries. There is a tool called the LIWC dictionary workbench that we

used to further improve our custom dictionaries [23]. Using the updated custom dictionary for Persistence and eID, we improved from an average of only 14% to an average of 25% of the words in all six papers being included in our dictionary. This means that our dimension percentages of the total words in the document were low, even for meaningful categories, but reasonable for analysis given that the standard dictionary had an average of 36% inclusion. Our reporting table thusly uses a sum of the dimensions within each category of the custom dictionary relative to the presence of the words, rather than the same % as the standard dictionary. This still gives a measure of relative usage when comparing across papers.

Refer to Table 8 for our custom dimension findings related to the following discussion. All papers used generic study jargon (e.g., data, research, etc.). While all papers used some demographic jargon, they primarily communicated age, race, and sex dimensions of demographics rather than meaningfully discussing location or socioeconomic status. EnEd Jargon was unsurprisingly the highest category for most papers [17]–[19], [22]. Besides the generic eID Jargon, these papers tended to focus on the identity dimensions of attitude, intersectionality, and mentors rather than self-efficacy or competence. Besides the typical persistence substitute of career goals, which was relatively high for all papers, these authors focused more on dimensions of interest and motivation rather than knowledge or retention. Still, all our custom dictionary dimensions of eID and persistence were present in all six papers. All papers used enrichment program jargon, especially around activities and results, but not as much jargon around multiple touchpoints (ongoing activities with participants), resources, or funding. Ironically, considering the topic of Diversity Pathways in STEM [19], the paper that appeared to have very little identity and persistence jargon did have the most jargon surrounding the STEM pipeline and enrichment programs!

Our second EnEd custom dictionary is about community cultural wealth (CCW), but at this time, we checked for the six types of capital and asset-based versus deficit-based jargon. LIWC, with this custom dictionary, found that three of the six literature reviews contained a much higher percentage of asset-based language over deficit-based language [17], [21], [22]. Still, only one review in Table 8 would be considered reflective of both usages [20]. Future Work could consist of creating another separate custom dictionary specifically for CCW themes and a more meaningful summary metric that reflects the degree of asset-based versus deficit-based language within a text.

3.2 "Bottom-Up" (*in vivo*) LIWC Meaning Extraction Method (MEM)

This style of text analysis, counting words utilized, did indeed show that the EnEd review literature for the given search is focused on the experiences of women in EnEd. Most of the literature review papers use both "interest" and "identity" in the contexts of both school and outreach programs. It is notable that "self-efficacy" was found twenty-four times in only four documents within the six-document set of literature reviews, and "persistence" was found only eight times in two documents (persist/persistent add four more mentions but only in one document). Also of note is that while "al" was excluded from the top-20 overall and the top-20 in all six documents, the exclusion of citation years and authors only became necessary with the top listed words shared between fewer documents. This could indicate a reliance on specific researchers for certain topics, e.g., on Dweck for mindset [24] or Eisenhart for sociocultural theory [20]. This frequency of words allows for more visual representations of salient themes, such as creating Figure 1 in Appendix A.

When using MEM, we kept the standard stop list and 1-gram setting [14]. We discovered various data conversion flaws ("artifacts"), so we manually adjusted the top-20 and top-10 tables to exclude non-meaningful words such as the "al" and various authors/years from citations within the texts. Refer to Table 2 and Table 3, respectively.

A weakness that these results exposed is one of typographical and PDF-to-TXT translation errors. So, for future work, we recommend a side-by-side reading of the papers and the TXT file before running LIWC in order to correct these issues. Also, more understanding of MEM analysis is needed to determine if meaningful words that share the same basis and are important to EnEd can be combined into one MEM count. For instance, the 20th word in all 6 documents is used 119 times, "development" from Table 2, but there are related words that could have changed its relative importance (e.g., develop, developmental, developer, etc.).

The MEM analysis also provides the opportunity to check the actual frequency of use for the literature review SPIDER search terms [3], [5]. It is interesting to note from Table 4 that four of the search terms are not present in any of the documents due to Boolean searching "OR". The notes column contains additional support for why our novel methodology procedure needed tweaking before embarking upon the full literature review for the dissertation and supports the reasoning behind this work-in-progress. Low frequencies might be a good indicator of saturation when that low-frequency search term is part of capturing synonyms around a research topic. In this usage, saturation is where no new information is gained by looking for additional literature resources in other databases. If all synonym terms are low frequency, that could be an indicator of a knowledge gap in the research. This check also exposes an interesting phenomenon with EnEd jargon typography in that "mixed" was not found, but "mixed-method" was. Another example of this is that "woman" was used 330 times in all 6 documents, but "women" was used less frequently.

This *in vivo* method of literature analysis brings a quantitative lens to the discussion around the academic jargon in EnEd literature about women in outreach. It also allows for the exploration of variant terms. The MEM analysis does indicate that science is stressed more than math and that high school has been the focus age over middle school in EnEd literature. With these results, the set of papers found for the dissertation have been shown to relate directly to the research topic.

3.3 Tips For Using the LIWC Program

In answering subpart B of our research question for this work-in-progress, we gathered important tips to guide research using computational text analysis. Items marked with an "⏚" are steps that were added due to the results previously discussed.

- Have administrative rights to your device to download, install, and update the licensing for the program.
- When using the internal standard LIWC dictionary, use the "View Internal Dictionary" button to download their human-readable chart.
    - We highly recommend reviewing this document prior to both analyzing LIWC results and creating a custom dictionary, to better understand which words belong to what categories.
    - Run the standard LIWC dictionary analysis first to understand the psychological tone. ⏚
- When using a user-defined dictionary:

  - Create a human-readable chart first, then create the machine-readable table for creating a DIC file, this helps in organizing dimensions by similar thematic ideas (categories) and assigning the keywords. Note that custom dictionaries cannot create the higher category automatically, which is why we opted for a gauge of relative importance using a sum across dimensions within a category, divided by the % of dictionary-included words.
  - Avoid duplicate keyword entries between different dimensions, if possible. ⏚
  - Do not repeat entries within a dimension, noting no differentiation between letter cases.
  - Use a numbering system that assigns a number to each dimension, and then assign the keywords to the corresponding category number.
  - Use a mouse and keyboard hotkeys, rather than a touchpad and right-click shortcuts, this can make creating the machine-readable table take less time.
  - In the human-readable chart, alphabetize the words by their categories to fix leading spaces – and do not leave a space after the last word.
  - With wildcard endings (e.g., *), if the root word will capture the expanded word, delete the expanded word; if there is truly a different meaning, add a note to the human-readable chart, or consider creating a separate custom dictionary. ⏚
  - Refer to Table 9 in Appendix A for other pitfalls and fixes discovered about custom dictionaries when we ran the LIWC dictionary workbench.
- Large documents or transcripts can be segmented to determine how one text changes over time, including the use of a LIWC feature called "Narrative Arc", and the LIWC analyses can be run on multiple texts at once for cross-analysis.
- Data Cleaning:
  - Some PDF documents will not be able to copy exactly to a TXT file, which may necessitate use of Optical Character Recognition software. Even if the PDF copies, the Table and Figure captions will likely need to be copied into the TXT file separately.
  - Certain papers, especially those using APA style, need removing "et al." in order to keep word count meaningful (e.g., the citation text "al" frequency was 274 times in all 6 docs). Choosing to do this is in keeping with qualitative data preparation precepts such as correcting misspellings in transcripts.
    Find-and-Replace "et al." with "." in each file with the keyboard shortcut "Ctrl+H". ⏚
  - Read each TXT file along with the PDF (side-by-side) to find and to remove header or footer text and any artifacts (data flaws) from the PDF conversion. Be especially cognizant of words typographically split by syllable over a line. ⏚
- During the MEM analysis:
  - Start with 1-gram (single word) topic modeling. Increasing to phrases (2-gram or more) will exponentially increase run times.
  - Ensure the "Stop List" is enabled (filtered words without contribution to shared meaning). Be sure to create a custom stop list if your research has a particular set of words that are not important but occur at high frequency, or if the standard stop list includes words with known meaningfulness to your topic.
  - Alternately, one can re-order the MEM analysis results to exclude custom stop words (as we show in Table 2 and Table 3).
- When exporting LIWC results into a separate file for statistical analysis or Excel-based visualization, if the dataset is too large, there are options to reduce the size.
- Once the analysis has been run, by using built-in visualization tools like color coding and word clouds, one can better understand a text beyond the mere numerical analysis.

- The creator of the program has collected a series of tutorials on how to make the most of the analyses [25]. This resource provides an excellent visual experience of the program besides teaching on its use.
- A final important thing to note about the program is that, while there is a "demo" version available on the internet, one must have a paid license to use all the features and functionality.
    - Commercial and non-academic licenses are also available; these are being used to develop multiple technologies, including those that assist law enforcement in analyzing suspect interviews [26].
    - License duration begins at the time of purchase, so it is wise to buy LIWC only after review of the tutorial information and once a project is identified.

3.4 Estimate of Time Saved

While it took about ten hours to convert all the PDF files into TXT files, and to check that the Figure/Table captions copied correctly, it only takes a few seconds to run thirty files on an 8GB RAM laptop with a 3GHz processor and a 64-bit operating system. This allows researchers to move quickly to a comparison of results once papers are in the program. As a body of TXT files used by a specific group are converted and verified for typographical errors, future work should have a reduction in the conversion time as certain papers remain meaningful to a particular EnEd topic.

The initial gathering of words into categories and the subsequent translation of the list into a LIWC-readable dictionary took approximately two days' worth of work, and another four days evaluating with the workbench and reorganizing the custom dictionary for more effective LIWC results, but this step no longer must be repeated for future literature reviews. A recommendation for groups who develop their own custom dictionaries is to periodically revisit their list and groupings to ensure the desired characteristics are being counted in the LIWC analysis.

The LIWC user-interface was well-designed, but there are tedious specific steps that must be performed, or the analysis may not be as expected from the tutorials [25]… as the GIGO adage applies to any computer program, the settings for each type of analysis matter, as Garbage In will give you Garbage Out for both the top-down and the bottom-up methods!

4 Limitations

This work divulges the methodology of using computational text analysis to aid in quickly determining the themes within a set of literature review papers. Both the LIWC standard and the research team's own custom dictionaries were used to evaluate six texts, as well as the LIWC meaning extraction method. While this novel methodology does not eliminate the need to read abstracts and critical text thoroughly, it can indeed save time when reviewing multiple studies and determining the major focus of reported information. The use of LIWC unexpectedly uncovered a dimension of work within the standard dictionary's Lifestyle category. Thus, we recommend that even if a research group uses a custom dictionary, that the standard LIWC analysis is also run in order to gauge themes on a psychologically verified basis [8]. Also unexpected from the MEM and related to the importance of choosing search terms wisely was the prevalence of "woman" over "women" in the literature [5], [14].

As with any word-counting *in vivo* method or coding *a priori* method, there is a possibility of confirmation bias, however that is why we still advocate a critical read-through of the study before the LIWC analysis and meaning extraction methods are completed. Additionally, we stopped the meaning extraction method process at the step where LIWC generates a table of which document contains what common words. To fully realize the power of this type of analysis, in future work we could utilize a statistical software package to perform a Principal Components Analysis, and we could determine how to collate meaningful words with the same basis (e.g., develop). Another limitation was found with the conversion of the documents from PDF to TXT, as there may have been additional artifacts from optical character recognition or other data cleaning processes that were not caught before analysis, as original procedure had the critical read-through slated for after the LIWC analyses. Thus, future work should include in the procedures a human reading both the PDF and TXT files side-by-side for any errors before starting LIWC analyses. As we advocate a critical reading of the text anyway, this is merely a change in the order of steps to add an additional layer of validation.

5 Conclusions

As with any computer program, we sought to validate the procedures for LIWC on a subset before continuing the venture with a larger body of work. While we discovered updates to our novel methodology, LIWC allowed for a more rapid journey through eID literature. For a holistic understanding of the state of any topic, one must read to capture the necessary insights from search results, so its use will never replace that step of the process. Still, we postulate that the use of computational text analysis in engineering education reviews will become more prevalent, and we advocate both for the LIWC platform for our methodology discussed here as well as for the five commonly accepted systematic literature review steps with the SPIDER search term framework.

Our team was able to quantitatively check the emergent themes with the repeated use of language from the texts themselves. The traditional LIWC analysis with the standard dictionary gives a psychologically verified and impartial look at the language, while with the custom dictionary it gives a measure of the known signposts for a topic. The MEM analysis gives the emergent themes within the topic. When these two *a priori* and *in vivo* thematic computational methods arrive at similar landmarks, researchers can be confident that although this took less time, the adventure has not only been worth the computations, but it has also arrived in the correct place.

6 References

[1] ASEE, "Engineering by the Numbers, 2010," American Society for Engineering Education, Washington, DC, Profiles of Engineering and Engineering Technology Colleges, 2011.

[2] ASEE, "Engineering & Engineering Technology by the Numbers, 2020," American Society for Engineering Education, Washington, DC, Profiles of Engineering and Engineering Technology Colleges 2020, 2021. Accessed: Dec. 13, 2021. [Online]. Available: https://ira.asee.org/wp-content/uploads/2021/11/Total-by-the-Number-2020.pdf

[3] M. Borrego, M. J. Foster, and J. E. Froyd, "Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields," *Journal of Engineering Education*, vol. 103, no. 1, pp. 45–76, Jan. 2014, doi: https://doi.org/10.1002/jee.20038.

[4] J. W. Creswell, *Qualitative Inquiry & Research Design: Choosing Among Five Approaches*, Fourth edition. Los Angeles: SAGE Publications, Inc, 2018.

[5] A. Cooke, D. Smith, and A. Booth, "Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis," *Qual Health Res*, vol. 22, no. 10, pp. 1435–1443, Oct. 2012, doi: https://doi.org/10.1177/1049732312452938.

[6] M. J. Page *et al.*, "The Prisma 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *Systematic Reviews*, vol. 10, no. 1, p. 89, Mar. 2021, doi: https://doi.org/10.1186/s13643-021-01626-4.

[7] Pennebaker Conglomerates, Inc, "Linguistic Inquiry and Word Count (LIWC)," *Welcome to LIWC-22*. https://www.liwc.app/ (accessed Aug. 18, 2022).

[8] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, Mar. 2010, doi: https://doi.org/10.1177/0261927X09351676.

[9] K. N. Jordan, J. Sterling, J. W. Pennebaker, and R. L. Boyd, "Examining Long-Term Trends in Politics and Culture Through Language of Political Leaders and Cultural Institutions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 9, pp. 3476–3481, Feb. 2019, doi: https://doi.org/10.1073/pnas.1811987116.

[10] E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser, "Pronoun Use Reflects Standings in Social Hierarchies," *Journal of Language and Social Psychology*, vol. 33, no. 2, pp. 125–143, Mar. 2014, doi: https://doi.org/10.1177/0261927X13502654.

[11] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying Words: Predicting Deception from Linguistic Styles," *Pers Soc Psychol Bull*, vol. 29, no. 5, pp. 665–675, May 2003, doi: https://doi.org/10.1177/0146167203029005010.

[12] S. C. Kalichman and J. M. Smyth, "'And You Don't Like, Don't Like the Way I Talk': Authenticity in the Language of Bruce Springsteen," *Psychology of Aesthetics, Creativity, and the Arts*, Jun. 2021, doi: https://doi.org/10.1037/aca0000402.

[13] D. Monzani *et al.*, "Emotional Tone, Analytical Thinking, and Somatosensory Processes of a Sample of Italian Tweets During the First Phases of the COVID-19 Pandemic: Observational Study," *Journal of Medical Internet Research*, p. e29820, Oct. 2021, doi: https://doi.org/10.2196/29820.

[14] C. K. Chung and J. W. Pennebaker, "Revealing Dimensions of Thinking in Open-Ended Self-Descriptions: An Automated Meaning Extraction Method for Natural Language," *Journal of Research in Personality*, vol. 42, no. 1, pp. 96–132, Feb. 2008, doi: https://doi.org/10.1016/j.jrp.2007.04.006.

[15] J. A. Maxwell, *Qualitative Research Design: An Interactive Approach*, 3rd edition. in Applied social research methods series, no. 41. Thousand Oaks, Calif: SAGE Publications, Inc, 2013.

[16] S. Secules *et al.*, "Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community," *Journal of Engineering Education*, vol. 110, no. 1, pp. 19–43, 2021, doi: https://doi.org/10.1002/jee.20377.

[17] S. L. Rodriguez, C. Lu, and M. Bartlett, "Engineering Identity Development: A Review of the Higher Education Literature," *International Journal of Education in Mathematics, Science and Technology*, vol. 6, no. 3, Art. no. 3, May 2018, doi: https://doi.org/10.18404/ijemst.428182.

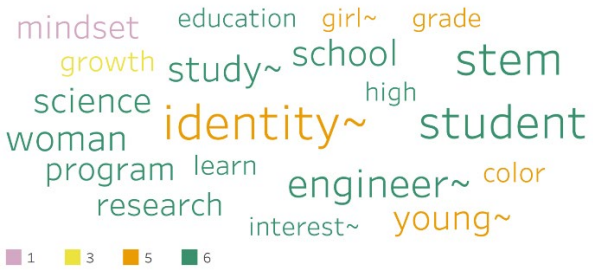
[18] A. Simpson and Y. Bouhafa, "Youths' and Adults' Identity in STEM: a Systematic Literature Review," *Journal for STEM Educ Res*, vol. 3, no. 2, pp. 167–194, Jul. 2020, doi: https://doi.org/10.1007/s41979-020-00034-y.

[19] N. L. Cagle, L. Caldwell, and R. Garcia, "K-12 Diversity Pathway Programs in the E-STEM Fields: A Review of Existing Programs and Summary of Perceived Unmet Needs," *Journal of STEM Education: Innovations and Research*, vol. 19, no. 4, pp. 12–18, Sep. 2018.

[20] M. Eisenhart and C. D. Allen, “Addressing Underrepresentation of Young Women of Color in Engineering and Computing Through the Lens of Sociocultural Theory,” *Cultural Studies of Science Education*, vol. 15, no. 3, pp. 793–824, Sep. 2020, doi: https://doi.org/10.1007/s11422-020-09976-6.
[21] A. Y. Kim, G. M. Sinatra, and V. Seyranian, “Developing a STEM Identity Among Young Women: A Social Identity Perspective,” *Review of Educational Research*, vol. 88, no. 4, pp. 589–625, Aug. 2018, doi: https://doi.org/10.3102/0034654318779957.
[22] M. Stohlmann, “Growth Mindset in K-8 Stem Education: A Review of the Literature Since 2007,” *Journal of Pedagogical Research*, vol. 6, no. 2, pp. 149–163, Apr. 2022, doi: https://doi.org/10.33902/JPR.202213029.
[23] *LIWC-22 Tutorial 4: The Dictionary Workbench*, (Oct. 10, 2022). Accessed: Feb. 25, 2023. [YouTube]. Available: https://www.youtube.com/watch?v=zSCetEXINjM
[24] C. S. Dweck, *Mindset: The New Psychology of Success*, Updated Edition. New York: Ballantine Books, 2007.
[25] *LIWC Tutorials*. Accessed: Nov. 08, 2022. [YouTube]. Available: https://www.youtube.com/user/jwpennebaker/videos
[26] R. Thirumalainambi and C. C. Jorgensen, “United States Patent: 8337208 - Content Analysis to Detect High Stress in Oral Interviews and Text Documents,” 8337208, Dec. 25, 2012

7 Appendix A

Figure 1: MEM Results Overall Top-20 Word Cloud



Note: Word size is directly proportional to frequency; Word color denotes how many documents used the term; ~means SPIDER search term.

Table 1: SPIDER criteria

Search Strategy	Search Terms Used
S (Sample)	(women OR girl* OR female* OR gender) AND ("middle school" OR adolescent* OR teenager* OR "young adult")
PI (Phenomenon of Interest)	engineering AND (identity OR† persist* OR grit OR pipeline)
D (Design)§	"questionnaire*" OR "survey*" OR "interview*" OR "focus group*" OR "case stud*" OR "observ*"
E (Evaluation)	(competenc* OR self-efficac* OR belong* OR attitude* OR mentor*) AND (career* OR interest* OR motivat* OR knowledg* OR retent*)
R (Research type)§	"qualitative" OR "mixed method*" OR "quantitative"

** Denotes wildcard (e.g., multiple endings for the term).*

† Changed AND to OR between Identity/Persistence because initially no results were found in EBSCO, only 9 in Scopus.

§ In all databases, no results found with full SPIDER, thus D & R terms excluded.

Note: Inclusion criteria included:

- Articles written in English
- Articles published between January 2001 through December 2021
- Peer Reviewed, Scholarly Journals only

Table 2: Top 20 Words From The Majority of Engineering Education Review Literature

Word	Frequency	Docs with Word	Word	Frequency in 6 Docs	Word	Frequency in 5 Docs
identity	608	5	stem	569	identity	608
stem	569	6	student	547	young	292
student	547	6	engineer	388	color	170
engineer	388	6	woman	330	grade	161
woman	330	6	study	315	girl	158
study	315	6	school	315	field	133
school	315	6	science	312	college	125
science	312	6	program	260	mathematics	106
young	292	5	research	246	discipline	103
mindset	264	1	learn	180	female	98
program	260	6	interest	158	intervention	95
research	246	6	high	158	gender	91
growth	195	3	education	157	environment	87
learn	180	6	experience	153	change	75
color	170	5	social	143	see	73
grade	161	5	work	141	theory	70
interest	158	6	math	128	american	66
high	158	6	group	123	class	62
girl	158	5	review	121	journal	61
education	157	6	development	119	researcher	60

Table 3: Top 10 Words Appearing in Some of the Engineering Education Review Literature

Word	Frequency in 4 Docs	Word	Frequency in 3 Docs	Word	Frequency in 2 Docs	Word	Frequency in 1 Doc
Middle	63	growth	195	compute	97	mindset	264
White	60	scholarship	54	free	65	e-stem	34
Peer	56	discourse	49	pipeline	36	technique	29
Target	44	band	44	ingroup	35	fix	21
Major	37	boy	42	prototype	33	post-secondary	21
Suggest	37	family	38	external	30	promise	21
Great	36	parent	36	narrative	24	k-8	20
Stereotype	34	effort	35	environmental	24	interactional	16
Motivation	34	youth	33	conference	21	scientifc	16
Encourage	34	women	32	k-12	20	afterschool	16

Table 4: Search Term Frequency in Set of Documents

SPIDER Term	MEM Frequency	# of Docs	Note:
women	32	3	Variant adds: woman +330
girl*	158	5	Freq. Sum/# Max (girl), "girls" not appearing
female*	98	5	Freq. Sum/# Max (female), "females" not appearing; Variant adds: feminine +6, femininity +2, feminism +1
gender	91	5	
middle	63	4	
school	315	6	Variant adds: afterschool +16, after-school +2, in-school +1, Out-of-school +10, school-* +2
adolescent*	17	2	Freq. Sum/# Max (adolescent, adoles, adolescence), "adolescents" not appearing; "adoles" shows the typographical error (word break artifact)
young	292	5	Variant adds: youth +33
adult	28	3	Variant adds: adulthood +1
engineer†	388	6	"engineering" wasn't found in any document; but there were 32 "ing", 4 "engi", and 1 "engine" which also shows the word break artifacts that reading first should catch
identity	608	5	Variant adds: identification +15, identify +63, identities-* +8, identity-* +5; note there were 19 misspellings in at least 3 documents, including word break artifacts.
persist*	12	2	Freq. Sum/# Max (persist, persistence, persistent), "persisted" not appearing; Variant adds: perseverance +1, persevere +5
pipeline	36	2	Variant adds: path +4, pathway +29 We acknowledge "pathway" to be a variant of "pipeline" for those unfamiliar with the bends and turns available to the designers of fluids systems.
survey*	8	3	Freq. Sum/# Max (survey), "surveyed, surveys" not appearing
interview	11	4	
focus	109	6	Variant adds: stem-focused +5
group*	123	6	Freq. Sum/# Max (group), "grouped, groups, subgroups, grouping" not appearing
case	24	4	
stud*	863	6	Freq. Sum/# Max (study, student, student-oriented), "studying, studied, studies" not appearing; note "student, student-oriented" were unintended inclusions per *
observ*	16	3	Freq. Sum/# Max (observation, observe, observational, observa), "observations, observed, observing" not appearing; "observa" shows the typographical error
competen*	20	3	Freq. Sum/# Max (competence, competency), "competencies, competent" not appearing; Variant adds: self-competence +2, incompetence +1, incompetency +1
self-efficac*	24	4	Freq. Sum/# Max (self-efficacy), "self-efficacies" not appearing; Variant adds: self-concept +12

SPIDER Term	MEM Frequency	# of Docs	Note:
belong*	31	3	Freq. Sum/# Max (belong, belongingness), "belonging" not appearing
attitud*	30	5	Freq. Sum/# Max (attitude, atti), "attitudes"/"attitudinal" not appearing; "atti" shows the typographical error
mentor*	18	4	Freq. Sum/# Max (mentor, mentorship), "mentors, mentored" not appearing
career*	77	6	Freq. Sum/# Max (career, career-ready, career-related), "careers" not appearing
interest*	159	6	Freq. Sum/# Max (interest, interestingly), "interested, interesting, interests" not appearing; Variant adds: disinterest +1
motivat*	43	4	Freq. Sum/# Max (motivate, motivation, motiva), "motivated, motivational, motivations" not appearing; "motiva" shows the typographical error; Variant adds: motive +1
knowledg*	30	6	Freq. Sum/# Max (knowledge, knowl), "knowledgeable" not appearing; "knowl" shows the typographical error
retent*	10	5	Freq. Sum/# Max (retention), "retentions, retentive, retentivity" not appearing; Variant adds: retain +8
qualitative	7	4	
method*	30	5	Freq. Sum/# Max (method, methodology), "methodological, methods" not appearing; Variant adds: mixed-method +1, mixed-methods +1
quantitative	8	4	

Search terms not appearing in MEM analysis: teenager, grit, questionnaire*, mixed.*

† Different form of original search term, this could advocate for a modified search term in future work.

** Denotes wildcard (e.g., multiple endings for the term) that added to frequency count but used max for number of documents; MEM-found Variants would add to frequency count, possibly informs future search terms.*

Table 5: Comparing Custom Dictionaries

Old Custom Dictionary (dimensions)		New Custom Dictionary (*categories* and dimensions)
		Study Jargon
n (numbers)	t (time scale)	SJ_time
	Theory	SJ_Theory
Study Variables	Study Type	SJ_Research Type
Metrics	Methods	SJ_Methods Metrics
Variable Scale	Program/Tool	SJ_Program Database
	Validation Jargon	SJ_Validation
		Demographic Jargon
Age Studied	Age/Timing	DJ_Ages
	Location	DJ_Location
	Demographic Jargon: Race	DJ_Race
	Demographic Jargon: Sex	DJ_Sex
	Demographic Jargon: SES	DJ_SES
	Phenomenon	*Engineering Education Jargon*
	Educational Jargon: Pedagogy	EnEd_Pedagogy
	Educational Jargon: Social	EnEd_Social
	Engineering Profession	EnEd_Engineering Profession
	Computer Science Profession	EnEd_Computer Science Profession
	Identity	*EnEd_Identity*
		ID_Self-efficacy
		ID_Attitude
		ID_Competence
	EP Features/ People	ID_Mentors
		ID_Intersect
	Persistence	*EnEd_Persistence*

Old Custom Dictionary (dimensions)		New Custom Dictionary (*categories* and dimensions)
Type of Influence	Pipeline	P_Career Goals
		P_Interest
		P_Motivation
		P_Knowledge
		P_ Retention
	Enrichment Program (EP) Jargon	*Enrichment Program Jargon*
	EP Features/ Activities	EP_Features Activities
	EP Features/ Branded tools & Resources	EP_Resources
	EP Features/ Funding	EP_Funding
	Enrichment Program Synonyms	EP_Synonyms
	EP Results Jargon	EP_Results
	EP studied	EP_Names
	Multiple Touchpoints	EP_Multiple Touchpoints
	Found CCW Jargon	*CCW Jargon*
		Asset-based Perspective
	Deficit-based Perspective	Deficit-based Perspective

Note: CCW category was separated into its own dictionary file to avoid the "repeated word" error

Table 6: LIWC Analysis Summary Measures

First Author's Last Name	Word Count	Words Per Sentence	Big Words	Punctuation	Standard Dictionary	Analytic	Clout	Authentic	Tone
Cagle	3915	23.6	39.0	25.6	80.9	93.2	49.0	33.8	40.9
Kim	12932	27.2	31.2	21.0	79.9	90.2	73.8	25.7	50.6
Rodriguez	4908	26.1	42.0	25.9	76.1	93.6	44.5	32.4	38.9
Eisenhart	15661	27.1	30.6	17.6	81.5	85.3	65.4	32.0	50.6
Simpson	9397	23.1	34.4	24.8	75.9	90.9	53.4	39.0	24.3
Stohlmann	7506	20.8	35.9	18.6	78.1	87.1	59.5	14.3	66.8
	#	#	%	Σ(% dims)	%	norm %	norm %	norm %	norm %
Max	15661	27.2	42.0	25.9	81.5	93.6	73.8	39.0	66.8
Min	3915	20.8	30.6	75.9	75.9	85.3	44.5	14.3	24.3
Range	11746	6.4	11.4	5.6	5.6	8.3	29.3	24.7	42.5
Mean	*9053*	*24.6*	*35.5*	*78.7*	*78.7*	*90.0*	*57.6*	*29.5*	*45.4*
Standard Deviation	*4578*	*2.6*	*4.4*	*2.4*	*2.4*	*3.3*	*10.9*	*8.6*	*14.3*

Table 7: LIWC Analysis Standard Dictionary Dimensions

First Author's Last Name	Linguistic	Drives	Cognition	Affect	Social	Culture	Lifestyle	Physical	Perception	Conversation
Cagle	48.1	4.1	11.9	2.2	8.2	2.4	12.6	0.2	8.5	0.3
Kim	49.4	4.4	11.0	2.9	12.1	2.2	10.3	0.3	8.4	0.1
Rodriguez	45.4	3.1	12.9	1.5	6.6	4.0	12.1	0.1	8.4	0.0
Eisenhart	54.5	5.1	10.9	3.5	11.2	2.1	8.8	0.4	9.1	0.1
Simpson	48.0	4.1	13.9	1.8	8.6	1.3	7.9	0.1	7.9	0.2
Stohlmann	50.8	5.4	13.7	3.6	10.0	0.8	12.3	0.3	6.2	0.0
	%	%	%	%	%	%	%	%	%	%
Max	54.5	5.4	13.9	3.6	12.1	4.0	12.6	0.4	9.1	0.3
Min	45.4	3.1	10.9	1.5	6.6	0.8	7.9	0.1	6.2	0.0
Range	9.1	2.3	3.0	2.1	5.5	3.3	4.8	0.2	3.0	0.3
Mean	*49.4*	*4.4*	*12.4*	*2.6*	*9.5*	*2.1*	*10.7*	*0.2*	*8.1*	*0.1*
Standard Deviation	*3.1*	*0.8*	*1.3*	*0.9*	*2.0*	*1.1*	*2.0*	*0.1*	*1.0*	*0.1*

Table 8: LIWC Analysis Custom Dictionary Dimensions

First Author's Last Name	Original Custom Dictionary	Revised Custom Dictionary	Study Jargon Used	Demographic Jargon Used	Engineering Education Jargon Used	EnEd Identity Used	EnEd Persistence Used	Enrichment Program Jargon Used	CCW % Asset	CCW % Deficit
Cagle	11.6	25.5	6.8	23.5	33.1	4.9	13.6	18.1	59%	39%
Kim	17.7	28.0	7.9	36.1	24.9	11.7	11.3	8.0	77%	22%
Rodriguez	11.8	23.5	15.7	14.7	29.1	22.3	9.5	8.7	62%	34%
Eisenhart	12.3	21.9	4.1	39.1	27.4	7.2	13.3	8.8	50%	50%
Simpson	10.9	21.7	15.2	13.2	33.3	22.7	6.5	9.0	59%	38%
Stohlmann	16.7	27.1	6.9	19.5	39.9	9.4	10.5	13.9	87%	13%
	% included	% included	Σ/(%)	Σ/(%)	Σ/(%)	Σ/(%)	Σ/(%)	Σ/(%)		
Max	17.7	28.0	15.7	39.1	39.9	22.7	13.6	18.1	87%	50%
Min	10.9	21.7	4.1	13.2	24.9	4.9	6.5	8.0	50%	13%
Range	6.8	6.3	11.5	26.0	15.0	17.8	7.1	10.1	38%	37%
Mean	*13.5*	*24.6*	*9.4*	*24.4*	*31.3*	*13.0*	*10.8*	*11.1*	*66%*	*32%*
Standard Deviation	*2.9*	*2.7*	*4.8*	*10.9*	*5.3*	*7.7*	*2.6*	*4.0*	*14%*	*13%*

Note: EnEd = Engineering Education , CCW = Community Cultural Wealth

Table 9: Key Custom Dictionary Tips for Reducing Workbench Errors

Workbench Error	Tip to Fix
Duplicated entries with different categories	Try to use each entry word once, consider carefully if the meaning is truly in both categories
Leading/ Trailing Spaces	Alphabetize categories in human-readable file to fix leading spaces - don't leave a space after the last word
Overlaps Entries	If the root word will capture the expanded word, delete the expanded word; If there's truly a different meaning, add note to human-readable file or consider separate dictionary
Duplicated entries with same categories	LIWC does not differentiate between capital and lowercase letters; no need to repeat

8 Abstract

Background: One way to broaden the participation of women in engineering beyond the commonly reported 20% proportion of degrees awarded is through providing outreach (e.g., enrichment programs) for young learners. Yet, we do not know the full impact of outreach, especially how it impacts persistence and engineering identity (eID) among girls, because these enrichment programs often happen in silos. Therefore, with the fast propagation of engineering education (EnEd) research, there is a need to **quickly** evaluate relevant research to identify gaps in our knowledge of eID development via outreach.

Purpose: A traditional (i.e., by hand) thematic literature review was conducted as a part of an ongoing study on Middle School outreach, eID, and persistence for women in engineering. However, we wanted to understand the viability and accuracy of a computer-driven analysis, Linguistic Inquiry and Word Count (LIWC), as a resource for fast, reliable analysis of literature.

Scope/Method: The program LIWC was used as an analysis tool to quickly gather data on a set of six literature review papers, with both user-defined and built-in dictionaries, as well as a topic modeling procedure, to refine the methodology for this novel approach.

Results: The use of LIWC to conduct a thematic literature review on a subset of articles confirmed the same themes that arrive via traditional coding methods, yet the novel computational method took less time and offered a few surprises. Thus, *a priori* codes using traditional LIWC analysis, with both the standard dictionary and our custom dictionary, and *in vivo* codes using LIWC meaning extraction method (MEM analysis), allowed us to quickly analyze how many papers used the same terms.

Conclusions: While the available computational tools allow us to quickly focus on the most salient of themes in the literature and come to inter-rater consistency faster, its use does not replace the need to read. Novel tools like LIWC might be the future for rapidly understanding the language of EnEd research and could help researchers more easily categorize prior research in their areas.

Keywords: *Systematic Literature Review, LIWC, Engineering tools, Computational tools, Text analysis*