

**Implementing Natural Language Processing (NLP) in R to Evaluate Knowledge
Acquisition from a Course on Success and Well-Being**

Angeline Cranton

School of Psychology, University of Ottawa

PSY4274: Honours Thesis

Supervisor: Dr. Darcy Santor

April 11, 2025

Table of Contents

Contribution Statement	6
Abstract	7
Introduction.....	8
Importance and Limitations of Mixed Methods Research	8
Overview and Rationale of the Current Study	10
Review of Existing Text-Parsing Tools	11
Thematic Analysis.....	11
Saturation	12
NVivo	14
Natural Language Processing (NLP)	15
Review of NLP Text-Parsing Tools in Mixed Methods Research.....	16
Review of Mixed Methods Validation of NLP Tools in R.....	18
Objectives of the Current Study.....	19
Theoretical Framework	21
Hypothesis 1	21
Hypothesis 2	22
Hypothesis 3	22
Methodology	22
Questionnaire Administration	24

Quantitative Questions	24
Qualitative Questions	24
Data Cleaning.....	25
RStudio	25
Analytical Methods	26
NLP Metrics in R.....	27
Frequency Analysis.....	27
Sentiment Analysis (NRC Lexicon).....	29
Sentiment Shift Analysis (NRC Lexicon).....	29
NRC Clustering.....	30
Latent Dirichlet Allocation (LDA).....	30
Word Cloud.....	31
Statistical Analyses in R	31
Results.....	32
Part 1: Results for Quantitative Metrics.....	32
Part 2: Results for NLP Metrics	34
Frequency Analysis.....	35
Course Satisfaction	35
Course Helpfulness	38
Sentiment Analysis	40

Course Satisfaction	41
Course Helpfulness	42
Part 3: Bivariate Associations Between Quantitative Metrics and NLP Metrics	44
Quantitative Metrics and NLP Frequency Metrics	46
Course Satisfaction	46
Course Helpfulness	47
Convergent Validity	48
Quantitative Metrics and NLP Sentiment Metrics.....	49
Course Satisfaction	49
Course Helpfulness	49
Convergent Validity	50
Part 4: Comparison of Quantitative Metrics and NLP Metrics	51
Latent Dirichlet Allocation (LDA)	51
Sentiment Shift Analysis	55
NRC Clustering	58
Discussion	61
Comparison of NLP Methods.....	61
Convergent Validity.....	64
Implications for Program Evaluations.....	66
Limitations	69

Conclusion	71
Declaration of Generative AI and AI-Assisted Technologies in the Coding Process	71
References.....	73
Appendix A	81
Appendix B	87

Contribution Statement

Angeline Cranton contributed to the methodological design, background research, and preparation of the manuscript supporting the current thesis. She was primarily responsible for developing the natural language processing (NLP) script in R, which was integral to the analysis of quantitative and qualitative data. Additionally, she conducted all computational analyses, produced the figures and tables, and contributed substantially to the interpretation of results. Her involvement also included drafting and finalizing the manuscript.

Abstract

Despite broad support for integrating qualitative and quantitative research approaches, persistent barriers hinder the widespread adoption of mixed-methods research, particularly when sample sizes are large or analysis must be completed promptly (e.g., course and program evaluations). The current study aimed to develop and implement several text-parsing analytic tools in R using natural language processing (NLP) to facilitate efficient analysis of qualitative data from large samples. Various NLP-derived analyses were conducted on qualitative responses from undergraduate students' end-of-course feedback ($N = 144$), including frequency analysis, sentiment analysis, sentiment shift analysis, NRC clustering, latent Dirichlet allocation (LDA), and word-cloud visualization. All NLP methods in R were validated against traditional quantitative measures (e.g., Likert-scale ratings). NLP-based qualitative metrics obtained through sentiment analysis consistently demonstrated an acceptable degree of convergent validity with quantitative metrics ($\rho = .27-.69, p < .05$) across domains of course satisfaction and helpfulness. Insights garnered from sentiment shift analysis and LDA topic modeling provided evidence for the incremental utility of NLP methods over standard Likert-scale ratings. These findings illustrate the benefits of incorporating NLP-metrics alongside traditional quantitative survey methods and illustrate the potential of NLP in R to bridge the gap between qualitative insight and quantitative depth in mixed-methods psychological research.

Keywords: mixed methods research, natural language processing, R, machine learning, text parsing, course evaluations, course satisfaction, course helpfulness

Introduction

Importance and Limitations of Mixed Methods Research

Mixed methods research is an increasingly prevalent approach that combines and analyzes quantitative and qualitative data (Fàbregues et al., 2021; Maxwell, 2016; Onwuegbuzie et al., 2022). Typically, quantitative data refers to data collected through self-report surveys and experiments yielding a range of numerical results, including the age of participants, the degree to which respondents agree or disagree with a set of statements, or the amount of satisfaction or difficulty severity a person experienced, and percentages. Conversely, qualitative data comprises non-numerical data, such as text, often concerning similar topics, including the degree to which respondents liked a certain event, how they confronted challenges or setbacks, or their reasoning behind a course of action or decision. Though painstaking to collect and analyze, qualitative data provides researchers with a deeper understanding of a research question or phenomenon, whereas quantitative data facilitates large-scale, efficient comparison of groups and individuals on numerous metrics (Parks & Peters, 2023).

While these methods present advantages and disadvantages, researchers have argued that the most thorough understanding of a domain of interest and lived experiences is achieved by combining both quantitative and qualitative methods (Chang et al., 2021; Fàbregues et al., 2021; Parks & Peters, 2023). Furthermore, mixed methods research is argued to assist in improving the dissemination of research problems and refinement of measurement instruments (Fàbregues et al., 2021). Although mixed methods research has been promoted and utilized for decades, conducting research through this approach has been historically challenging due to the complexities and difficulties of merging two highly different datasets (Chang et al., 2021; Griffith et al., 2024; Maxwell, 2016). As a result, the global practice of mixed methods research

has been subject to criticism, limited, and underutilized compared to its methodological counterparts (Fàbregues et al., 2021; Maxwell, 2016). When used, researchers typically prioritize quantitative data, focusing solely on a small facet of collected qualitative data, often using these insights to contextualize numerical findings and thus failing to integrate both elements altogether (Onwuegbuzie et al., 2022). Accordingly, most information gathered from qualitative reports is substantially overlooked. These notable challenges suggest that mixed methods data analysis, by virtue of its time-consuming nature, is impractical and should be avoided regardless of the detail offered by quantitative *and* qualitative data collection.

Despite these challenges, the collection and utilization of qualitative information, largely from individual comments and feedback, remains a central component of evaluation in countless settings, including consumer satisfaction in industry, program satisfaction and effectiveness in clinics and hospitals, and course evaluations at colleges and universities (Chang et al., 2021; Kabir et al., 2020; Shaik et al., 2022; Weissman et al., 2019; Yuan & Hu, 2014). At the end of every academic semester at most colleges and universities across North America, students are commonly asked a series of quantitative questions concerning course content and instructional effectiveness, alongside open-ended qualitative prompts requesting suggestions to improve the course and personal experiences in those classes (Shaik et al., 2022). Although routinely collected, this qualitative information is rarely used in a systematic manner, nor seen by anyone besides the student and course instructor (Yuan & Hu, 2014). In large courses with up to or exceeding 200 students, conducting a systematic and thorough evaluation of individual comments would be impractical (Shaik et al., 2022). Consequently, comments submitted by students, particularly in large courses, are unlikely to be utilized to the degree that they could be and were intended to be used at an institutional level. Even if read, the influence of several

factors, including reader fatigue or rater bias, can be expected to undermine the validity and reliability of any inferences drawn from this type of information. Indeed, utilizing more of this information and doing so in a more reliable manner will depend upon developing more efficient means to extract, evaluate, and interpret monumental quantities of qualitative data routinely collected in classrooms across North America.

Recent developments in statistical methods capitalizing on machine learning natural language processing (NLP) models have produced a series of extraordinarily powerful and efficient tools for text analysis and synthesis, resembling the traditional focus of qualitative data analysis (Chowdhary, 2020; Shaik et al., 2022; Yuan & Hu, 2014; Zaki et al., 2023). These tools promise to address several of the concerns with qualitative data analysis and mixed methods research. However, despite the rapid pace at which NLP models have been developed and deployed across numerous industries and social domains, they have yet to be adequately investigated and implemented in the analysis of qualitative student responses in North American educational contexts (Yuan & Hu, 2014).

Overview and Rationale of the Current Study

Accordingly, the objective of the current study was to develop and validate numerous machine learning text-parsing and analytical tools to enable researchers and policymakers to benefit from the volumes of qualitative data frequently provided by students on an annual basis. This involved (a) a review of existing text-parsing tools based on NLP models, (b) a review of the use of these tools in mixed methods research, particularly in university course evaluations, and (c) a review of the extent to which mixed-methods investigations have attempted to validate information garnered from quantitative and qualitative questions concerning a similar topic or domain in the R programming environment. Even a brief review of mixed methods research

illustrates that though qualitative and quantitative methods are often used simultaneously, there has been little to no attempt to validate one against the other (Chang et al., 2021; Maxwell, 2016; Onwuegbuzie et al., 2022). From a psychometric perspective, information addressing a similar domain of investigation (e.g., participant satisfaction) should exhibit a degree of concurrent validity, regardless of differences in collection methods (Campbell & Fiske, 1959)

Review of Existing Text-Parsing Tools

In this section, I review the extensive literature on text-parsing approaches, including thematic analysis, saturation analysis, NVivo, and NLP. Qualitative data analysis refers to the exploration of non-numerical data, such as human language and phenomena, and is a leading approach to investigating open-ended reports of lived experience within numerous fields of research, especially across social science disciplines (Ayre & McCaffery, 2022; Oliveira et al., 2015). Qualitative methods encompass several analytical approaches to assessing open-ended information, including traditional thematic analysis, saturation, and contemporary natural language processing (NLP).

Thematic Analysis. Thematic analysis is a standard qualitative procedure consisting of *content*, *reflexive*, and *codebook techniques*, all of which vary in theme production, researcher subjectivity, and coding structure. Thematic *content* analysis is considered the most widespread approach within this research domain, obtaining a general frequency of ideas, terms, thoughts, and explanations reported by participants (Ayre & McCaffery, 2022; Braun & Clarke, 2006). This approach predates all NLP-based analytic methods that have emerged in recent years. Thematic *content* analysis is reliant upon pre-determined categories defined by theory and is executed through a systematic coding structure (Ayre & McCaffery, 2022; Braun & Clarke,

2006). Hence, thematic *content* analysis can be understood as a theory-driven qualitative approach.

Conversely, *reflexive* thematic analysis extracts themes through “patterns of shared meanings” that emerge from either explicit (e.g., “I feel X emotion when Y occurs”) or implicit ideas, thoughts, and concepts reported by participants (Ayre & McCaffery, 2022, p. 78). This approach can be implemented through a fluid inductive, deductive, or hybrid coding structure, thus providing more depth than content analysis (Ayre & McCaffery, 2022; Braun & Clarke, 2019). Accordingly, *reflexive* thematic analysis can be understood as a researcher-driven qualitative approach. Moreover, *codebook* thematic analysis offers a bridge between content and reflexive approaches, producing themes that the researcher must map into a representative ‘codebook’ framework with detailed descriptions of uncovered themes and coding guidelines (Ayre & McCaffery, 2022; Braun et al., 2019). Yet, unlike content and reflexive methods, *codebook* thematic analysis can be understood as a collaborative qualitative approach, having a semi-structured coding system developed by multiple researchers throughout the analytical process (Ayre & McCaffery, 2022; Braun et al., 2019).

Saturation. Saturation refers to “a criterion for discontinuing data collection and/or analysis” once additional themes, issues, or insights cease to emerge from data (Saunders et al., 2018, p. 1894). This approach also assumes that qualitative research on a particular phenomenon can conclude once “all relevant conceptual categories have been identified, explored, and exhausted,” consequently indicating the phenomenon of interest has been subject to sufficient inquiry (Hennink et al., 2017, p. 592). Like thematic analysis, saturation comprises several approaches, namely *theoretical*, *data*, *code or thematic*, and *meaning* techniques.

First, *theoretical* saturation focuses on sampling adequacy and conceptual development, specifically the event in which collecting additional data on a theoretical construct or category becomes redundant as further analyses do not produce novel insights (Hennink et al., 2017; Saunders et al., 2018). Achieving *theoretical* saturation suggests all relevant concepts, categories, or constructs are thoroughly explored, thus further research on a strong, valid, and supported theory is no longer necessary (Hennink et al., 2017). Second, *data* saturation emphasizes data collection sufficiency (Saunders et al., 2018). This method suggests data repeats findings once collection fails to yield supplementary concerns (Hennink et al., 2017; Saunders et al., 2018). Compared to theoretical saturation, *data* saturation can infer whether sample sizes are adequate within a given study's methodology (Hennink et al., 2017). Third, *code or thematic* saturation refers to the point at which no new themes or issues emerge from the qualitative dataset, indicating the codebook has reached a state of stabilization (Hennink et al., 2017). Codebook stabilization is demonstrated once the codebook has consistently and comprehensively captured all thematic content (Hennink et al., 2017).

Thematic saturation also operates on two levels – inductive and a priori – the former on the emergence of novel themes or topics in a data set and the latter on the representation of pre-defined themes or codes in the data set (Saunders et al., 2018). Resembling theoretical saturation, inductive *thematic* saturation is achieved once further themes or topics cannot be identified in the data set, leading to codebook redundancy (Hennink et al., 2017; Saunders et al., 2018). In contrast, a priori *thematic* saturation is accomplished at the point where a given data set exemplifies all pre-determined themes, topics, or codes of concern adequately, indicating thematic patterns are well-established within the study (Saunders et al., 2018).

NVivo. Many researchers argue that instead of conducting qualitative research manually, performing these methods in Computer Assisted Qualitative Data Analysis Software (CAQDAS), such as NVivo, can enhance analytical quality and efficiency by reducing the time-consuming labour of traditional qualitative methods (Dhakal, 2022; Zamawe, 2015). Accordingly, NVivo has become a dominant software used by researchers to assess textual data (Dhakal, 2022; Zamawe, 2015). NVivo is computer software used to manage, analyze, and report qualitative data (Dhakal, 2022). Additionally, NVivo can be used for mixed methods approaches, with qualitative results often exported and processed in statistical software, including SAS, IBM SPSS, and Microsoft Excel (Chang et al., 2021; Dhakal, 2022; Oliveira et al., 2015; Zamawe, 2015).

The goal of this software application is to assist researchers in coding for themes, topics, and concepts of interest, as well as sort, label, and classify their data sets for simplicity before utilizing NVivo's tools for thematic mapping (Dhakal, 2022; Oliveira et al., 2015). Alongside thematic mappings, NVivo offers a few basic machine learning text-parsing functions, such as sentiment and relationship analyses (Dhakal, 2022). Yet, these analyses are manual, with users having to tag data they consider 'positive' or 'negative' and specify whether relationships are associative, unidirectional, or bidirectional based on subjective interpretations of textual data (Dhakal, 2022). Thus, like traditional thematic and saturation techniques, conducting qualitative research in NVivo requires researchers to be deeply involved with data processing and assessment due to its point-and-click and coding procedures (Dhakal, 2022; Zamawe, 2015). This software is not open-source, thereby preventing researchers from fully understanding or vetting the methods (Dhakal, 2022). NVivo is also proprietary, costing a minimum of \$1,249.00

for academic licensure, making it relatively inaccessible to the public, students, and researchers with limited funding (Dhakal, 2022).

Natural Language Processing (NLP). Natural language processing (NLP) refers to a set of machine learning techniques that systematically process, analyze, and modify human writing and/or speech (Chowdhary, 2020; Cranton & Santor, 2024; Kang et al., 2020). The goal of NLP-based methods is to provide computers with the ability to understand human language to the extent that another human can (Khurana et al., 2023). These methods are an emerging approach in qualitative research, offering several advantages, including the rapid analysis of large volumes of unstructured text, methodological flexibility, versatility, accessibility, reduced rater bias, and interdisciplinary applicability (Chowdhary, 2020; Cranton & Santor, 2024; Kang et al., 2020; Sawicki et al., 2023). While NVivo and SAS are considered leading analytical software supporting the utilization of NLP models and subsequent statistical analyses, R provides a favourable programming environment for this approach's integration in mixed-methods research designs (R Core Team, 2020; RStudio Team, 2020).

Unlike the proprietary nature of NVivo, R is an open-access, cost-effective program that is highly efficient with large data sets (Chowdhary, 2020; Kang et al., 2020; Sawicki et al., 2023; Welbers et al., 2017). Open-access packages in R contain a range of advanced machine learning functions, such as frequency, sentiment, and latent Dirichlet allocation analyses, allowing users to execute NLP techniques alongside unique visual analytical tools, such as word clouds (Cranton & Santor, 2024). Furthermore, R can advance transparency and reproducibility of research findings by conducting both qualitative and quantitative analyses in a single environment, reducing the necessity of exporting qualitative data into external statistical software like SAS, thus offering potential to advance mixed-methods approaches (Chowdhary, 2020;

Kumar & Paul, 2016; Sawicki et al., 2023; Welbers et al., 2017). With a substantial quantity of contributors, R provides an extensive library of NLP packages, rendering its automation capabilities, script customization, text pre-processing control, and compatibility with machine learning text-parsing tools better than its competitors (Kang et al., 2020; Kannan et al., 2014; Sawicki et al., 2023; Welbers et al., 2017).

Review of NLP Text-Parsing Tools in Mixed Methods Research

In the last five years, social and scientific disciplines have increasingly sought practical and expeditious mixed methods research for analyzing extensive textual data from large sample sizes; NLP techniques provide a methodological lens to address this demand (Chang et al., 2021; Griffith et al., 2024). Mixed methods research has long been strenuous and time-consuming, particularly with the involvement of large-scale textual datasets (Chang et al., 2021; Griffith et al., 2024; Onwuegbuzie et al., 2022). Yet, despite its persistent time and resource expenditure, this research convention is essential in addressing adversity to inform theory, method, and policy, improving quality of life (Chang et al., 2021; Fàbregues et al., 2021). As a quick and robust alternative to traditional qualitative methods, NLP techniques are increasingly recognized as a novel and evolving solution to the challenges of mixed methods procedures as an expeditious and “inherently mixed analysis approach” producing qualitative and quantitative material (Chang et al., 2021, p. 401). This methodological shift is evident in higher education, where NLP can be explored for analyzing student feedback.

Course evaluations are essential to understand student experiences in educational settings, alongside the utilization of student feedback to improve data-driven instructional and administrative change (Kastrati et al., 2021; Yuan & Hu, 2014). Despite routine administration, course instructors often prioritize quantitative survey findings over open-ended comments from

students, leading to negligence of textual feedback and the assumption that few responses, if read, are representative and generalizable to the broader student population (Yuan & Hu, 2024). In response to a lack of attention to qualitative data collected from course evaluations NLP models have recently emerged as a vital tool to improve this assessment process by way of automating the analysis of several thousands of words in a matter of seconds (Alqahtani et al., 2023; Shaik et al., 2022; Yuan & Hu, 2024; Zaki et al., 2023). Although NLP is a prominent tool in today's education system and is used for various assessment protocols such as grading short-answer examinations, academic papers, and peer reviewing, they are rarely applied to course evaluations (Alqahtani et al., 2023; Kastrati et al., 2021; Shaik et al., 2022). Nevertheless, NLP can reduce the laborious process of manual feedback assessment that contributes to instructors' inattention to textual responses collected through course evaluations (Kastrati et al., 2021; Yuan & Hu, 2024; Zaki et al., 2023).

Recent studies have demonstrated NLP's capacity to transform how course evaluations are interpreted in higher education. For instance, findings from Yuan and Hu's (2024) pioneering study using machine learning text-parsing tools to interpret 100 course evaluations at a Chinese university indicated that NLP techniques are not only highly efficient and effective when processing large datasets, but capable of producing valuable insights from extensive textual responses that would otherwise be discarded by instructors. Additionally, Kastrati et al.'s (2021) systematic mapping study on 92 articles that applied NLP-based sentiment analysis for 'opinion mining' in educational domains revealed that merely 13% of studies investigated students' textual perspectives on course instruction and content. Hence, existing literature suggests NLP techniques in R can advance how student experiences in academic institutions are measured, understood, and attended to (Kastrati et al., 2021; Yuan & Hu, 2024). Further research

demonstrating the utility of NLP in R for assessing course evaluations is imperative. To date, no study has directly compared NLP-based approaches with traditional quantitative methods, such as Likert-scale ratings, in evaluating aspects like course satisfaction or perceived benefit.

Gradually populating the realm of mixed methods clinical research, NLP techniques demonstrate promise beyond academic feedback by advancing how intervention outcomes are understood. NLP-based text parsing, thematic mapping, sentiment analysis, and topic modelling procedures are shown to supplement Likert-scale findings obtained from self-report measures (Griffith et al., 2024). For example, a mixed methods study conducted by Griffith et al. (2024) on 107 exit interviews from a clinical intervention utilized LDA topic modelling and sentiment analysis in R to interpret Likert-scale ratings from an exit survey. Their NLP-derived metrics identified topics from awareness to feedback, alongside emotionality, entrenched in participant experiences (Griffith et al., 2024). Compounded with quantitative scores, these NLP-based methods aided them to understand the nuances of intervention helpfulness and its impact on mental health and wellness by pinpointing the most and least effective aspects reported by participants (Griffith et al., 2024). This study exemplified how NLP-based approaches to integrating qualitative and quantitative data are advancing mixed methods research, offering a more holistic understanding of human experiences. However, rather than explicitly testing the relationship between Likert-scale ratings and NLP-derived qualitative metrics, validation was implied through patterns observed in information obtained from both analytical methods (Griffith et al., 2024). Therefore, despite its rapid adoption, the application of NLP in effectively combining numerical and textual data necessitates psychometric validation to ensure accuracy (Chang et al., 2021).

Review of Mixed Methods Validation of NLP Tools in R

While limited research has evaluated the validity of NLP methods in R within the scope of the current study, including LDA, frequency analysis, and NRC clustering, there is some statistical evidence supporting the construct validity of basic sentiment analysis in R (Weissman et al., 2019). However, a study assessing the validity of the *syuzhet* package in R could not be identified, indicating a gap in the literature that should be addressed.

Convergent validity examines the degree of association between instruments measuring related constructs and is obtained through strong correlations (Weissman et al., 2019). Weissman et al. (2019) assessed the convergent validity of the same sentiment analysis packages by conducting pairwise Pearson's correlations and calculating internal reliability using Cronbach's alpha. Their findings showed modest positive correlations between packages, all of which were statistically significant: *AFINN* and *sentimentr* ($r = 0.26, p < 0.001$), *AFINN* and *CoreNLP* ($r = 0.23, p < 0.001$), *sentimentr* and *CoreNLP* ($r = 0.09, p < 0.001$) (p. 118). These correlations demonstrate an unacceptable degree of convergent validity. Furthermore, the reported Cronbach's alpha of 0.65 (95% CI 0.64-0.65) indicates insufficient internal consistency among the sentiment packages (p. 117). Together, these findings illustrate that the convergent validity of certain sentiment analysis packages in R remains limited. Thus, exploring the psychometric properties of an alternative sentiment analysis package in R, such as *syuzhet*, could provide a more consistent and valid option for future mixed methods research.

Objectives of the Current Study

The current study used self-report data from a sample of student course evaluations on their experience with an undergraduate course at the University of Ottawa to examine and compare traditional quantitative methods (i.e., Likert-scale scores) to NLP-based methods (i.e., frequency analysis, sentiment analysis, sentiment shift analysis, NRC clustering, and latent

Dirichlet allocation). Course evaluations included thematically similar quantitative (e.g., “How much did you enjoy the course?”) and qualitative (e.g., “What were the five things that you liked most about this course (and why did you like them)?”) questions. These questionnaires produced textual and numerical information concerning the domains of course satisfaction and course helpfulness, subsequently facilitating the comparison of both qualitative and quantitative approaches within each domain. Data processing and qualitative analysis of text-based feedback from participants were completed in R, incorporating several existing natural language processing (NLP) functions, all of which produce a variety of metrics that can subsequently be used to evaluate the convergence of NLP metrics with more traditional quantitative metrics. Quantitative metrics generated from qualitative text parsing functions were correlated with quantitative metrics from Likert-scale ratings.

NLP-metrics can be characterized as a hybrid of qualitative and quantitative methods, as they produce a metric that summarizes, for example, the frequency of a theme, which is routinely done with existing CAQDAS, such as NVivo (Dhakal, 2022; Oliveira et al., 2015). There are several NLP methods, such as sentiment analysis, that will quantify qualitative themes in various ways, producing different types of results. The degree of quantification of qualitative themes needs to be examined carefully and validated whenever possible. One goal of this study was to assess the unique benefits of these different NLP-analytic methods, not only in comparison to each other but also against more traditional quantitative methods, such as Likert-scale ratings. The NLP-based analytic methods used in the current study will be described in detail throughout the following section. All R programming developed and employed in this study has been deposited in a GitHub repository and is freely available for researchers to utilize and scrutinize as desired.

Theoretical Framework

The primary goal of the current study was to validate the extent to which information garnered from qualitative questions would be related to information obtained from similar quantitative questions. Psychometric theory of validity posits that information assessing a shared domain, regardless of its collection method, should nonetheless be correlated (Campbell & Fiske, 1959; Weissman et al., 2019). This is often achieved through a multi-trait multidimensional matrix (Campbell & Fiske, 1959). Accordingly, two methods assessing the same trait should be correlated, just as two quantitative measures of the same trait should be associated (Campbell & Fiske, 1959). These processes are theoretically predicted to result in an acceptable degree of convergent validity (Weissman et al., 2019). However, as previously discussed, there is little research examining the convergent validity between qualitative and quantitative methods, especially with NLP models (Weissman et al., 2019).

In contrast, information typically produced by qualitative methods is used to enrich or interpret information garnered from quantitative methods without establishing a direct comparison or attempting to evaluate the validity of two differing data sources tapping into similar domains. While it is expected to obtain additional insights from qualitative data with respect to quantitative findings, it is also reasonable to anticipate a degree of association reinforcing this theoretical relationship. In the current study, the degree of that association is estimated through a correlation coefficient involving two metrics – first from quantitative data and the second through textual analysis embedded in NLP. The following hypotheses were stipulated prior to analyses:

Hypothesis 1

H_0 : No correlation exists between quantitative Likert-scale scores and NLP-derived frequency counts ($H_0: \rho = 0$).

H_1 : A correlation exists between quantitative Likert-scale scores and NLP-derived frequency counts ($H_1: \rho \neq 0$).

Hypothesis 2

H_0 : No correlation exists between quantitative Likert-scale scores and NLP-derived sentiment counts ($H_0: \rho = 0$).

H_2 : A correlation exists between quantitative Likert-scale scores and NLP-derived sentiment counts ($H_2: \rho \neq 0$).

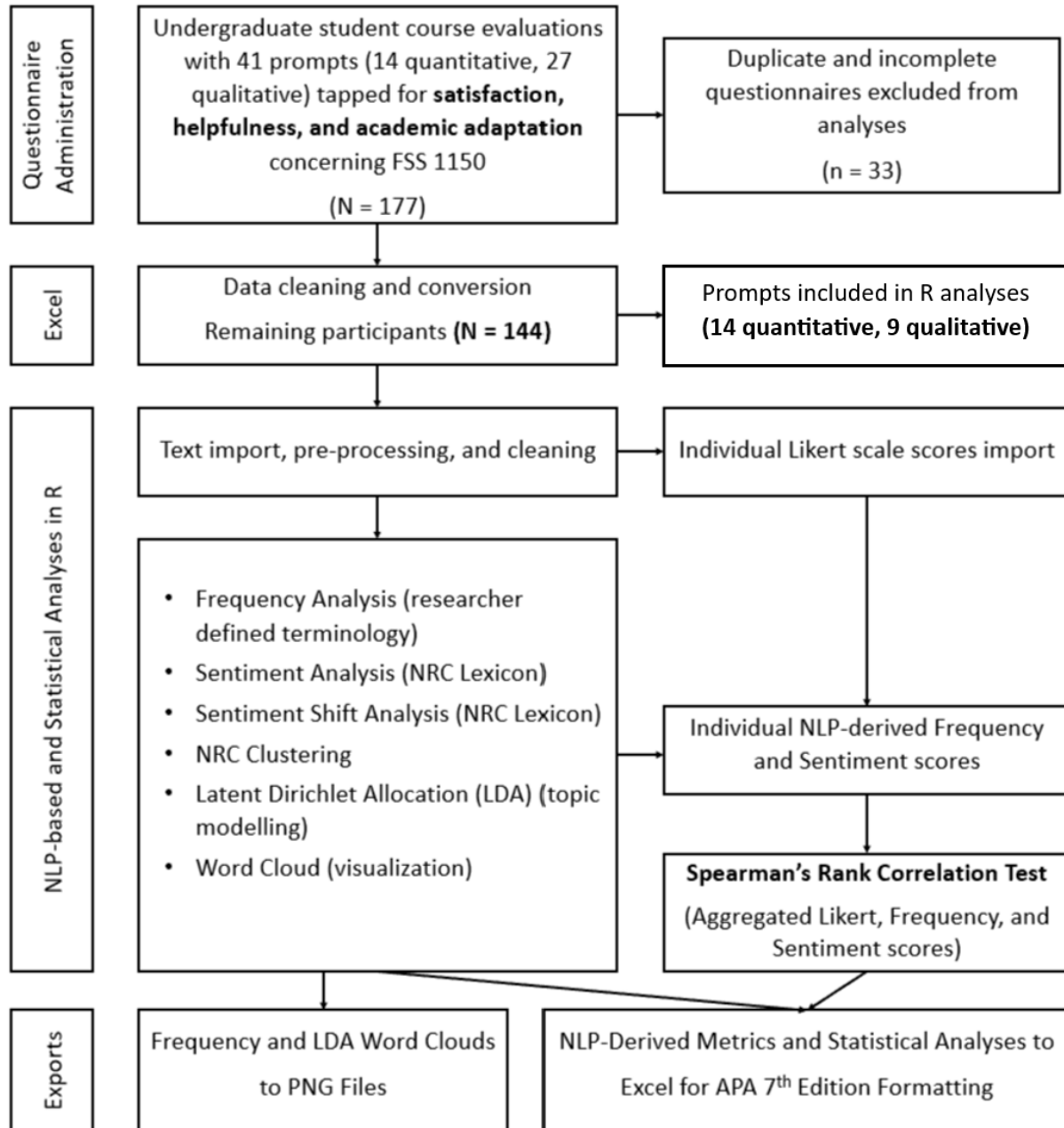
Hypothesis 3

H_0 : NLP-derived insights provide no incremental utility beyond quantitative Likert-scale scores.

H_3 : NLP-derived insights provide incremental utility beyond quantitative Likert-scale scores.

Methodology

The purpose of this section is to concretely describe both the data collection procedure alongside the NLP methods used to parse and analyze qualitative text responses in undergraduate course evaluations. An overview of the current study's mixed methods design is shown in Figure 1.

Figure 1*Overview of the Current Study's Mixed Methods Design*

Note. NLP refers to natural language processing. NRC refers to the sentiment lexicon embedded in the *syuzhet* R package.

Questionnaire Administration

Course evaluations were administered through Qualtrics XM to 177 ($N = 144$) undergraduate students enrolled in *Skills for Success and Well-Being in the Social Sciences* (FSS 1150) at the University of Ottawa. Students provided quantitative ratings and qualitative feedback on the course. The entire course evaluation questionnaire used in this study is shown in Appendix A.

Quantitative Questions

Students answered 14 questions about the course (e.g., how helpful they found the course) on a 10-point Likert scale ranging from 0 (“Not at all”) to 10 (“Almost all the time”). The current study examined only a few of the questions, namely (a) “How much did you enjoy the course?”, “How meaningful was this course to you?”, “How engaging was the material for you?”. And (b) “How helpful did you find the course?”.

Qualitative Questions

Students were also provided with open-ended qualitative questions through which they provided detailed textual elaborations on each prompt. First, in the domain of course satisfaction, students were asked, “What were the five things that you liked most about this course (and why did you like them)?”. Second, in the domain of course helpfulness, students were asked the following questions: “I would like to know how this course personally helped you (or didn’t help you). This is separate from how much you enjoyed or did not enjoy the course. Please be as detailed as possible. Try to think of at least five things (or more) that you feel were helpful to you”, “What other kinds of help would have helped you do better this semester at university?”, “What were the five most important skills that were helpful to you and would like to continue to

use? For each skill, I would like to know (a) what the skill was and (b) how it helped you or if it did not help you.” As previously stated, when responding to the qualitative prompt on the most important skills from the course, students were required to complete the two Likert-scale rating counterparts for each skill concerning skill helpfulness and usage.

Data Cleaning

All quantitative Likert-scale scores and qualitative responses were stored in a Microsoft Excel spreadsheet and required several modifications before being imported into the R programming environment. To ensure the formatting of the raw dataset was compatible with the structure and integrity of the R script, all rows containing missing values were excluded. Duplicate rows and pilot responses were also eliminated to ensure each participant included in the study was represented only once. Incomplete and duplicate questionnaires ($n = 33$) were excluded from the study to preserve statistical precision, qualitative depth, and accurate representation of student perspectives, thus producing a final sample size of 144 students. Columns containing special characters, titles, irrelevant labels, and sensitive information, such as student email and IP addresses, were deleted. As a result, the cleaned version of the dataset only included column labels, row labels, and cells with information relevant to quantitative ratings, qualitative text, and response tracking. Finally, all cells containing text were subject to alignment reformatting to ensure text did not spill over into subsequent cells and create errors in the analytical process. Before being imported into R, the spreadsheet was converted to CSV.UTF-8 format for efficiency in the programming environment.

RStudio

Once converted, the dataset was imported into the RStudio programming interface for further cleaning and pre-processing using the *readr* and *stringr* packages in R (Wickham, 2023; Wickham et al., 2024). A custom ‘Clean_String’ function was defined to apply sequential cleaning logic to the textual data, allowing standardization in formatting for subsequent NLP text parsing analyses. This function checked over all rows, columns, and cells to ensure any missing data overlooked by the manual cleaning process in Microsoft Excel was flagged and returned an empty string (i.e., chain of characters) to evade programming errors. ‘Clean_String’ also removed non-alphanumeric characters, stop-words, and replaced punctuations with spaces across all strings. All text was converted to lowercase, multiple spaces were collapsed, and a clean string was returned for each cell. Additionally, all names in the first column, representing each participant, were replaced with anonymous ID labels consistent with row numbers to preserve response confidentiality.

Analytical Methods

All NLP-derived metrics and statistical analyses were coded and conducted directly in R syntax using open-access packages within the RStudio programming environment. Given that the code comprises 923 lines, the entire R script developed for the present study is available in a GitHub repository for open-access use and alteration in subsequent research:

<https://github.com/AngelineCranton/NLP-Mixed-Methods.git>.

All NLP metrics and statistical analyses were conducted using open-access NLP packages for R. These findings were exported from RStudio to Microsoft Excel using the *openxlsx* package in R to ensure table formatting adhered to APA 7th edition guidelines (Schauberger & Walker, 2024).

NLP Metrics in R

Frequency Analysis. Frequency analysis was used to count the occurrence of pre-defined terms in the cleaned text. This method resembles thematic *content* analysis, allowing the identification of common terms reported by participants and providing insights into key themes, topics, or patterns within textual responses. Frequency analysis required manually coding of umbrella topics and corresponding terms, also known as ‘consolidated terms.’ 18 ‘consolidated terms’ were pre-defined for subsequent analysis through a customized ‘Count_Targets’ function (see Figure 2). Frequency analysis was conducted in RStudio using the open-access *stringr* and *dplyr* packages in R (Wickham, 2023; Wickham et al., 2023). All individual responses to the nine qualitative questions were subject to frequency analysis. This analysis was also conducted on aggregated responses for the five most helpful skills reported by participants in the course helpfulness domain. Results of frequency analyses were plotted on a bar chart using the *ggplot2* package and visualized as a word cloud using the *wordcloud* package (Fellows, 2018; Wickham, 2016).

Figure 2*Consolidated Term Mappings for NLP Frequency Analysis in R*

```

# Define consolidated terms for mapping - defined by researcher using terminology from the course and participant reflection responses
# Alternative to researcher-defined terms = Latent Dirichlet Allocation (LDA) which derives relevant terminology from the data based on frequency.
consolidated_terms <- list(
  "acquiring grit" = c("acquiring grit", "grit", "courage", "drive", "driven"),
  "help-seeking" = c("asking for help", "help seeking", "seeking help"),
  "backwards scheduling" = c("backwards scheduling", "schedules", "schedule", "scheduling", "time management"),
  "burn out" = c("burn out", "burnt out", "exhausted", "exhaustion", "exhausting", "fatigue", "fatigued", "tired", "tiring", "overworked", "overworking", "depleting", "depleted"),
  "coping with setbacks" = c("setback", "setbacks", "failed", "failing", "failure", "academic setbacks", "growth", "growth mindset", "resilience", "resilient"),
  "cornell note-taking" = c("cornell note", "cornell note-taking", "cornell notes"),
  "curbing procrastination" = c("curb procrastination", "curbing procrastination", "procrastination", "procrastinating", "procrastinate"),
  "goal-setting" = c("goal setting", "setting goals", "goals", "making goals", "creating goals"),
  "learning" = c("learning", "learned", "understand", "understanding", "comprehend", "comprehending", "comprehension", "deep learning", "retention", "memory"),
  "motivation" = c("motivation", "motivate", "motivates", "motivated", "motivating", "carrots", "carrot", "motivational factors", "rewards", "reward", "rewarding"),
  "sleep hygiene" = c("sleep hygiene", "sleep", "sleeping"),
  "stress management" = c("stress", "stress management", "stress as load", "as load", "and load", "course load", "overwhelmed", "overwhelm", "stress as worry", "as worry", "worried", "worrying", "and worry", "anxious", "anxiety", "negative thinking", "negative thoughts", "rumination", "ruminate", "ruminative thinking", "spiral", "spiralling", "catastrophize", "catastrophizing", "relaxation breathing", "breathing", "mindfulness", "fact-checking", "fact checking"),
  "study habits" = c("study", "studying", "study", "study strategies", "study skills"),
  "study skills" = c("pomodoro technique", "pomodoros", "breaks", "taking breaks", "spaced repetition", "interleaved practice", "deliberate practice", "active recall", "cue cards", "flashcards", "flash cards", "que cards"),
  "to-do lists" = c("to-do lists", "to do list", "to-do list", "to do lists", "todo list", "todo lists"),
  "well-being" = c("well-being", "well being", "happy", "happiness", "satisfaction", "satisfied", "satisfying", "perma", "PERMA", "perma theory", "positive emotions", "engagement", "relationships", "meaning", "achievement", "mental health", "self-esteem", "depression", "depressed", "mental illness", "mental health condition"),
  "course satisfaction" = c("satisfaction", "satisfied", "satisfying", "enjoyed", "enjoy", "positive", "liked", "appreciated", "appreciate", "positive experience"),
  "course helpfulness" = c("helpful", "helped", "helps", "useful", "benefitted", "benefit", "beneficial", "valued", "value", "valuable", "practical", "practicality", "applicable", "application", "effective", "works", "worked", "impactful", "impact", "support", "supportive", "supporting")
)

```

Note: This figure illustrates the code snippet from R syntax wherein NLP-based frequency analysis terminology is being manually defined, or ‘hard-coded’. Umbrella topics (i.e., consolidated terms) were bolded for interpretability. “c(“...”, “...”)” refers to the terms being listed for each topic of interest.

Sentiment Analysis (NRC Lexicon). Sentiment analysis determined the underlying emotional tone and polarity behind cleaned text by assigning ‘sentiment counts’ (Kastrati et al., 2021; Saini et al., 2019). This approach enabled the categorization of overall sentiment content as positive, negative, or neutral. By applying R’s pre-defined NRC word-emotion association lexicon, sentiments were categorized as the following: joy, fear, trust, anger, disgust, positive, negative, sadness, surprise, and anticipation (Mohammad & Turney, 2013). Sentiment analysis was performed in RStudio using the *syuzhet* package in R (Jockers, 2015). All individual responses to qualitative questions were subject to sentiment analysis. Sentiment analysis was also conducted on the combined five most helpful skills reported by participants in the course helpfulness domain. Sentiment results were plotted on a bar chart using the *ggplot2* package (Wickham, 2016).

Sentiment Shift Analysis (NRC Lexicon). Sentiment shift analysis is built upon basic sentiment analysis, tracking changes in sentiment counts over different segments of the questionnaire. This method permitted sentiment profiles to be identified for an increased, decreased, or zero change as the text progressed across participant responses. In short, this method compared sentiment counts at different response intervals throughout the questionnaire. Like sentiment analysis, sentiment shift was implemented using the *syuzhet* package and utilized the NRC lexicon in R (Jockers, 2015). Sentiment shift was calculated using individual sentiment results for the course helpfulness domain. Delta (Δ) values for sentiment profiles were calculated by subtracting the pre-sentiment profile from the post-sentiment profile for each participant. The following qualitative prompts were paired to calculate their delta values for all sentiments in a participant’s profile: “How did this course help you?” versus “How did this course not help you?” These specific prompts were selected as they provide an opportunity to understand how

participants' perspectives changed when asked to critically reflect on course helpfulness. Once the sentiment shift results were exported to Microsoft Excel, a randomized subset was selected for every 10th participant ($n = 14$) to generate a customized condensed heat map table to disseminate individual findings.

NRC Clustering. NRC clustering is another sentiment technique through which partition-based, non-hierarchical k-means clustering algorithms were used to group words based on their designated emotional categories derived from R's NRC sentiment lexicon (Shaik et al., 2022). As with sentiment shift analysis, this process required sentiment analysis to be conducted before its inclusion in R syntax. This approach mapped the emotional landscape of the dataset to determine which emotions were most prevalent in a nuanced context of interest and was performed in RStudio using the *syuzhet* and *cluster* packages in R (Jockers, 2015; Maechler et al., 2014). To perform NRC clustering, individual sentiment results for each domain response set were first aggregated to depict macro-level patterns. NRC clustering was also conducted on a sum of combined sentiment counts for the five most helpful skills reported by participants in the course helpfulness domain. Five clusters were pre-set to obtain a greater understanding of emotional variation embedded in sentiment profiles throughout each qualitative prompt.

Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation (LDA) is a rigorous probabilistic topic modeling technique that uncovers hidden themes or topics within large sets of text (Blei et al., 2003; Shaik et al., 2022). Resembling reflexive thematic analysis, LDA analyzed co-occurrence patterns of words used by participants to identify topics without pre-defined labels (Blei et al., 2003). LDA was especially useful for complementing findings from frequency analysis for post-hoc comparative purposes. While frequency analysis required researcher-defined topics containing a particular set of terms (i.e., consolidated terms), LDA generated a

predetermined number of topics based on how a specific number of terms were identified and grouped in the textual dataset. 18 topics were pre-set, each with five terms, through a customized ‘Perform_LDA’ function. LDA was conducted in RStudio using the *tm* and *topicmodels* packages in R (Feinerer & Hornik, 2024; Feinerer et al., 2008; Grün & Hornik, 2011; Grün & Hornik, 2024). All individual responses to qualitative prompts were subject to LDA.

Word Cloud. A word cloud is a data visualization tool that emerged from NLP. This technique generated a visual representation of textual data, where word size and colour were proportional to frequency in the qualitative dataset (Saini et al., 2019). Several word clouds were generated using NLP-derived metrics yielded from frequency analysis and LDA topic modeling for both domains of interest. Notably, the word cloud of LDA results in domain 2, based on participants’ reports of the five most important course skills related to course helpfulness, was filtered using a frequency threshold of 150 counts to facilitate comparison with the frequency analysis of the same question. Unlike bar plots or spreadsheets, this tool provided an immediate summary of frequency analysis results, making it digestible to the viewer in a straightforward manner. Word clouds were conducted in RStudio using the *wordcloud* and *RColorBrewer* packages in R and were exported as a PNG file using the *png* package in R (Fellows, 2018; Neuwirth, 2022; Urbanek, 2022).

Statistical Analyses in R

Given the non-normal and ordinal distribution of the dataset, non-parametric Spearman’s rank order correlations were computed to assess convergent validity between quantitative Likert-scale scores and aggregated NLP-derived frequency and sentiment metrics across both domains of interest. Spearman’s rho and *p*-values were yielded using the *cor.test* base package in R without specifying an alpha level (R Core Team, 2020; RStudio Team, 2020). Descriptive

statistics, including median, interquartile range (IQR), alongside minimum and maximum values, were generated for all variables using a customized ‘get_desc_stats’ function to describe central tendency and dispersion without assuming normality. Analyses were stratified by questionnaire sections to evaluate domain-specific patterns. Data manipulation and wrangling through ‘get_desc_stats’ was conducted using the *dplyr*, *ggplot2*, and *stringr* packages in R (Henry & Vaughan, 2023; Wickham, 2016; Wickham, 2023).

Results

The results are presented in four parts. Findings from the analysis of quantitative ratings and NLP-based analytic methods are presented separately in Parts 1 and 2. Bivariate associations between quantitative and NLP-based analytics are presented in Part 3. Selective and illustrative indicators of incremental utility offered by NLP metrics are presented in Part 4. Tables and figures have been inserted directly into the text to facilitate the presentation of findings.

Part 1: Results for Quantitative Metrics

Quantitative results for all Likert-scale questions assessing course satisfaction and helpfulness, including acquired skill helpfulness and usage, are presented as descriptive statistics in Table 1.

Table 1*Descriptive Statistics for Quantitative Ratings*

Variable	<i>Mdn</i>	<i>IQR</i>	<i>Min</i>	<i>Max</i>
Domain 1: Course Satisfaction				
Course Enjoyment Rating	8	2	2	10
Course Meaning Rating	8	3	1	10
Course Engagement Rating	7	2.25	1	10
Domain 2: Course Helpfulness				
Course Helpfulness Rating	8	2	1	10
Skill 1 Helpfulness Rating	9	2	2	10
Skill 1 Usage Rating	8	3	2	10
Skill 2 Helpfulness Rating	9	2	0	10
Skill 2 Usage Rating	8	3	0	10
Skill 3 Helpfulness Rating	9	2	0	10
Skill 3 Usage Rating	8	4	0	10
Skill 4 Helpfulness Rating	9	2	0	10
Skill 4 Usage Rating	8	4	0	10
Skill 5 Helpfulness Rating	9	2	0	10
Skill 5 Usage Rating	8	4	0	10

Note. The table shown illustrates descriptive statistics of quantitative Likert-scale scores calculated in R. *Mdn* = median, *IQR* = interquartile range. All medians and interquartile ranges are reported due to non-normal distributions. *Min* and *Max* values indicate response heterogeneity.

Results demonstrated that Likert-scale ratings for course satisfaction varied widely for each question, indicating diverse experiences of enjoyment (*Mdn* = 8 (*IQR* = 2); *Min* = 2, *Max* = 10), meaning (*Mdn* = 8 (*IQR* = 3); *Min* = 1, *Max* = 10), and engagement (*Mdn* = 7 (*IQR* = 2.25); *Min* = 0, *Max* = 73) among participants. Despite response variation, ratings were generally high for this domain, suggesting that students overall found the course to be a positive and meaningful experience that cultivated a satisfying learning environment.

Furthermore, Likert-scale ratings for general course helpfulness varied greatly ($Mdn = 8$ ($IQR = 2$); $Min = 1$, $Max = 10$), suggesting that while perceptions were generally positive, few participants had less favourable views concerning the degree to which the course was beneficial. When rating the helpfulness of their top five acquired skills in the course helpfulness domain, participants reported these skills were consistently helpful ($Mdn = 9$ ($IQR = 2$)). However, the wide range of helpfulness scores across these five skills ($Min = 0$, $Max = 10$) suggests students likely experienced perceptual differences in utility depending on their individual-level learning needs, preferences, prior knowledge, and future applications of these skills. In contrast, usage ratings for all five skills were consistently lower than skill helpfulness ($Mdn = 8$ ($IQR = 3-4$); $Min = 0$, $Max = 10$), signifying that while participants recognized the value of these skills, their implementation varied. These discrepancies between skill helpfulness and usage may indicate barriers linked to a perceived lack of opportunity, confidence, or relevance in applying these acquired skills beyond the course context.

In summary, findings revealed that while course satisfaction and helpfulness were rated highly among students, responses varied widely across individual experiences. Despite rating acquired skills as consistently helpful, lower usage ratings suggest external factors, such as confidence in application, may have influenced students' practical implementation.

Part 2: Results for NLP Metrics

Results for the various NLP-derived metrics are presented in this section, including measures of consolidated term frequency and sentiment analysis. Findings garnered from latent Dirichlet allocation (LDA), sentiment shift analysis, and NRC clustering will be presented in section four.

Frequency Analysis

Frequency analysis calculated the occurrence of consolidated terms present in textual responses. Consolidated terms refer to the list of 18 topics containing relevant terms defined by the researcher, such as “backwards scheduling,” “stress management,” and “study habits,” before conducting analyses (see Figure 2). Descriptive statistics for frequency analysis are shown in Table 2.

Table 2

Descriptive Statistics for Consolidated Term Frequency Analysis

Variable	<i>Mdn</i>	<i>IQR</i>	<i>Min</i>	<i>Max</i>
Domain 1: Course Satisfaction				
Course Satisfaction Term Frequency	18	22	0	73
Domain 2: Course Helpfulness				
Course Helpfulness Term Frequency	14	14	0	66
Course Unhelpfulness Term Frequency	2	5	0	21
Skill 1 Term Frequency	4	5	0	30
Skill 2 Term Frequency	4	5	0	20
Skill 3 Term Frequency	4	4	0	17
Skill 4 Term Frequency	4	4	0	21
Skill 5 Term Frequency	4	4	0	21

Note. The table shown illustrates descriptive statistics of frequency analysis calculated in R.

Term Frequency” refers to the number of NLP-derived consolidated terms identified in textual responses; *Mdn* = median, *IQR* = interquartile range. All medians and interquartile ranges are reported due to non-normal distributions. *Min* and *Max* values indicate response heterogeneity.

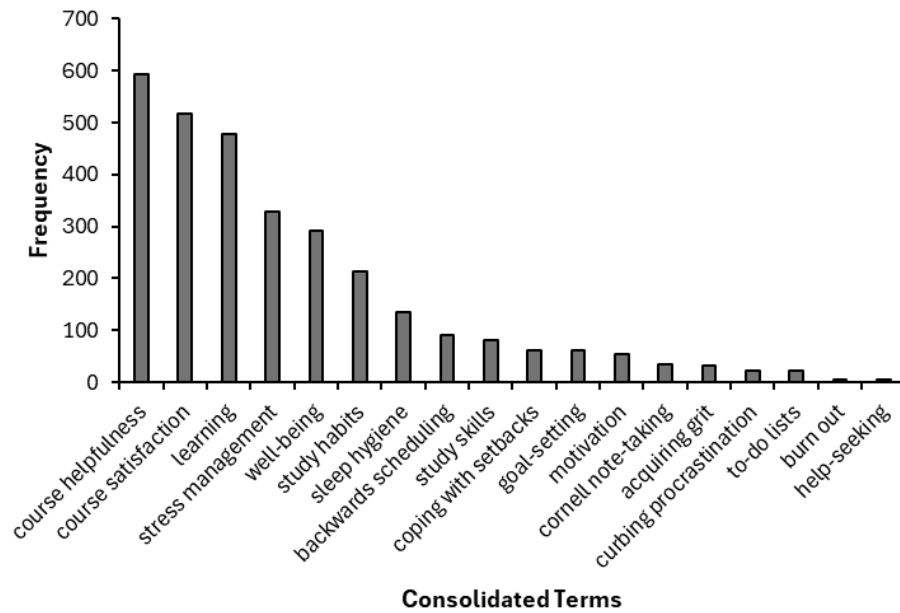
Course Satisfaction. Frequency analysis of participant responses concerning course satisfaction revealed moderate consolidated term occurrence (*Mdn* = 18 (*IQR* = 22); *Min* = 0, *Max* = 73) and a positively skewed frequency distribution. While some textual responses

contained no mentions, others included extensive mentions, suggesting that students' language choices deviated from the anticipated terminology when expressing their enjoyment, meaning, and engagement with the course. Overall, participants exhibited substantial variability in the presence of consolidated terms when articulating their retrospective satisfaction with instructional material and design.

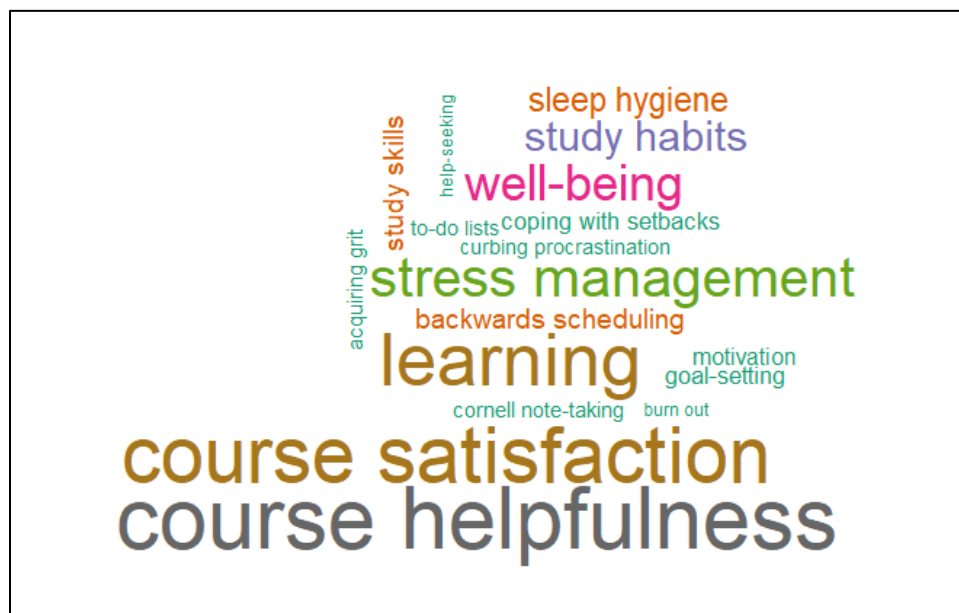
Consolidated term frequency analysis conducted on course satisfaction prompts from the questionnaire was visualized on a bar plot (see Figure 3) and an NLP-derived word cloud (see Figure 4). The five most frequently reported consolidated terms were “course helpfulness,” “course satisfaction,” “learning,” “stress management,” and “well-being,” suggesting participants consistently associated terms coded for these topics with their general course experience. These terms likely represent central aspects of student satisfaction, indicating that instructional content, alongside its practical benefits, was valued for personal development and subjective well-being. Variation in consolidated term frequency counts also demonstrated diversity in students' reflections of the course, illustrating the individualized nature of academic experiences.

Figure 3

Frequency of Consolidated Terms Reported for Course Satisfaction

**Figure 4**

NLP-Derived Word Cloud of Consolidated Terms Reported for Course Satisfaction

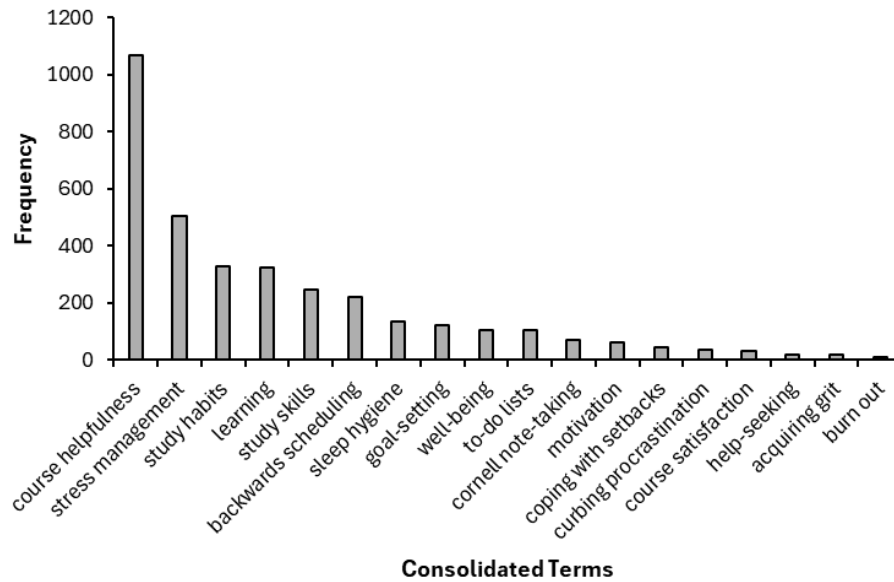


Course Helpfulness. NLP-derived frequency analysis on textual responses on general course helpfulness revealed moderate consolidated term occurrence ($Mdn = 14$ ($IQR = 14$); $Min = 0$, $Max = 66$). However, responses on general course unhelpfulness generated lower consolidated term frequency ($Mdn = 2$ ($IQR = 5$); $Min = 0$, $Max = 21$), indicating that while some participants expressed concerns within the bounds of predefined terminology, these instances were less frequent and more varied in distribution. Nevertheless, all distributions were positively skewed. When discussing helpfulness regarding the five skills they considered most important, participants exhibited consistently sparse consolidated term occurrences ($Mdn = 4$ ($IQR = 4-5$); $Min = 0$, $Max = 30$), suggesting students may have relied more on context-specific language over predefined terminology when expressing how these skills were personally or academic advantageous both within and outside the course setting. Like course satisfaction, the course helpfulness frequency distributions were positively skewed, wherein consolidated term count variability persisted.

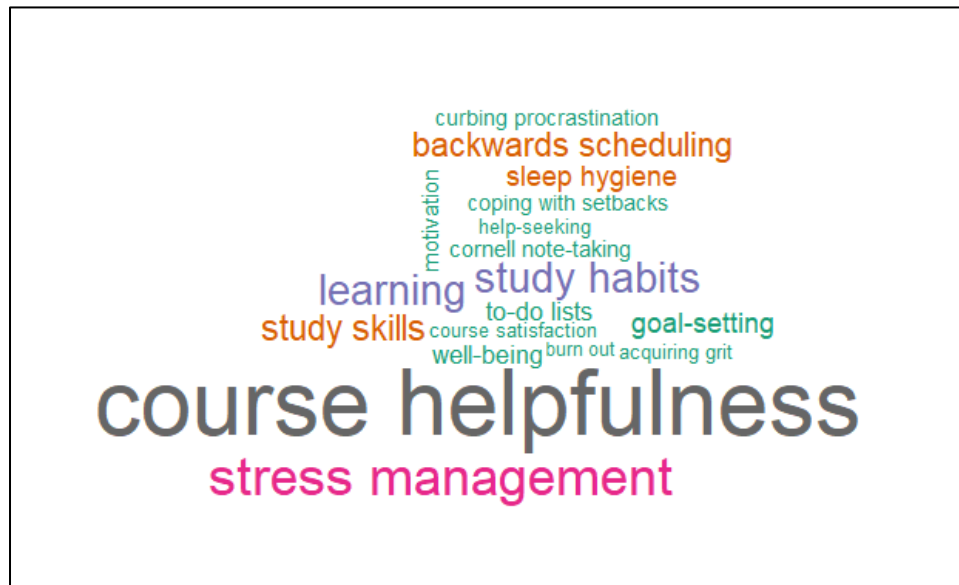
Consolidated term frequency analysis conducted on acquired skill helpfulness responses from the questionnaire was visualized on a bar plot (see Figure 5) and an NLP-derived word cloud (see Figure 6). The five most frequently reported consolidated terms were “course helpfulness,” “stress management,” “study habits,” “learning,” and “study skills,” suggesting participants associated terms coded for these topics with the acquired skills they considered most helpful. Following explicit mentions of terminology coded for helpfulness (e.g., “helped” and “helpful”), the remaining topics likely represent the skills commonly reported as advantageous among students. This distribution suggests students both recognized these skills as beneficial and linked their reflection to broader themes of academic success and personal development characteristic of a satisfying and helpful learning environment.

Figure 5

Frequency of Consolidated Terms Reported for Acquired Skills Helpfulness

**Figure 6**

NLP-Derived Word Cloud of Consolidated Terms Reported for Acquired Skills Helpfulness



In summary, consolidated term frequency analysis revealed considerable variability in how students articulated course satisfaction and helpfulness, with responses ranging from zero to extensive mentions of predefined terminology. The most frequently reported terms in both domains suggest that students valued instructional content and its practical benefits for academic and personal growth. Furthermore, responses concerning the helpfulness of skills acquired from the course indicated that while students consistently recognized specific skills as valuable, they often relied on individualized or context-specific language apart from consolidated terminology. Overall, these findings elucidate the heterogeneous manner students evaluated the course and its ability to foster a pleasant environment in which useful knowledge can be disseminated.

Sentiment Analysis

Sentiment analysis detected emotional tones in textual responses to generate sentiment profiles comprised of positive and negative emotive dimensions, including feelings of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Descriptive statistics for sentiment analysis are shown in Table 3.

Table 3*Descriptive Statistics for Sentiment Analysis*

Variable	<i>Mdn</i>	<i>IQR</i>	<i>Min</i>	<i>Max</i>
Domain 1: Course Satisfaction				
Course Satisfaction Sentiment Count	72	54.25	4	212
Domain 2: Course Helpfulness				
Course Helpfulness Sentiment Count	40.5	31	5	192
Course Unhelpfulness Sentiment Count	20	15	4	68
Skill 1 Sentiment Count	25	12	8	69
Skill 2 Sentiment Count	25	12	0	58
Skill 3 Sentiment Count	25	13.25	0	52
Skill 4 Sentiment Count	25	12	0	53
Skill 5 Sentiment Count	25	12	0	74

Note. The table shown illustrates descriptive statistics of sentiment analysis calculated in R.

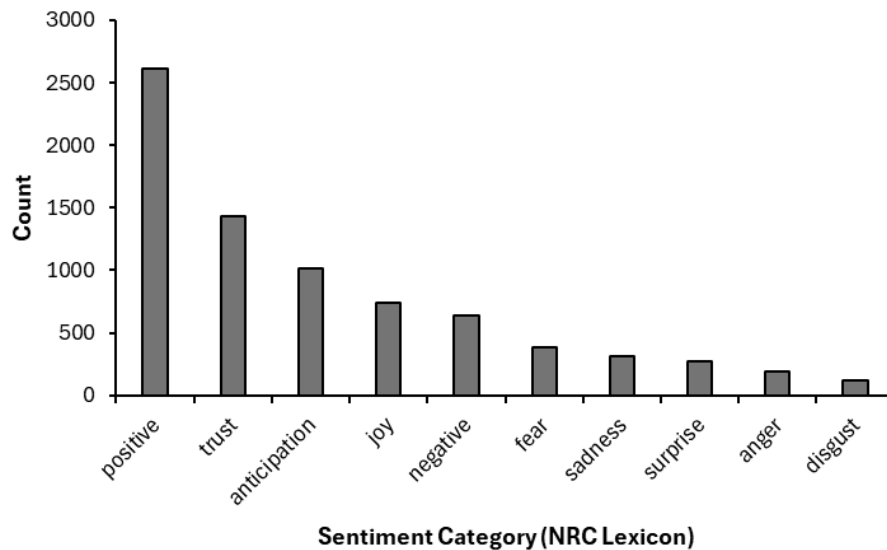
“Sentiment Count” refers to the number of NLP-derived, NRC lexicon sentiments identified in textual responses; *Mdn* = median, *IQR* = interquartile range. All medians and interquartile ranges are reported due to non-normal distributions. *Min* and *Max* values indicate response heterogeneity.

Course Satisfaction. NLP-derived sentiment analysis of textual responses concerning course satisfaction revealed moderate sentiment expression (*Mdn* = 72 (*IQR* = 54.25); *Min* = 4, *Max* = 212) with a positively skewed distribution. These findings indicated considerable variability in emotional tone, as some participants used minimal affective language while others conveyed strong sentiments about their enjoyment, meaning, and engagement with the course. While most responses exhibited a moderate level of sentimentality, a subset of participants demonstrated either low or highly pronounced emotional expression, likely contributing to the distribution’s skew.

Sentiment analysis conducted on the course satisfaction prompt from the questionnaire was visualized on a bar plot (see Figure 7). Responses were overwhelmingly characterized by positive sentiments among participants, followed by trust and anticipation. These findings suggest that students generally responded favourably toward the course, as frequent sentiments reflected feelings of confidence and enthusiasm for its impact on their learning experiences. Alongside the dominance of positive emotions, the degree of trust and anticipation identified in course evaluations insinuates that most students adopted optimistic perspectives towards the course.

Figure 7

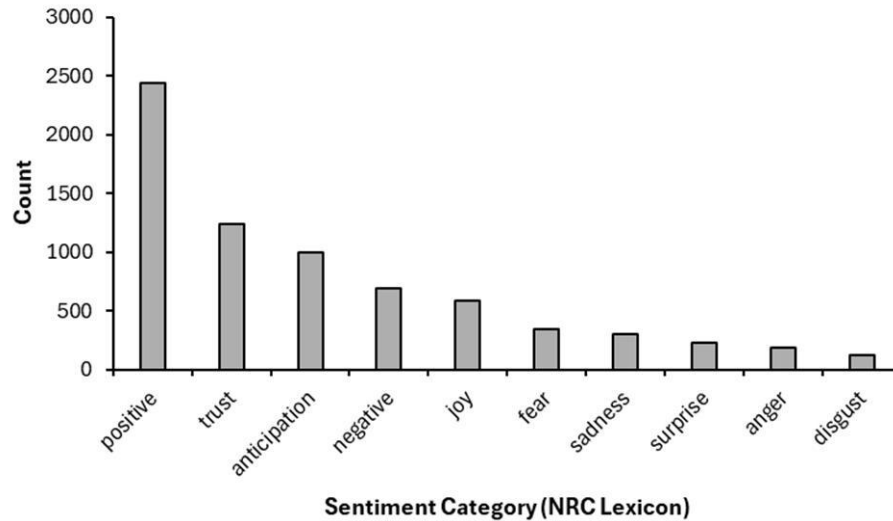
Sentiment Analysis of Course Satisfaction Responses



Course Helpfulness. NLP-derived sentiment analysis of textual responses on general course helpfulness revealed moderate sentiment expression ($Mdn = 40.5$ ($IQR = 31$); $Min = 5$, $Max = 192$). Conversely, responses on general course unhelpfulness demonstrated the lowest sentimentality overall ($Mdn = 20$ ($IQR = 15$); $Min = 4$, $Max = 68$), suggesting participants

expressed relatively restrained emotionality compared to their helpfulness evaluations. Furthermore, when discussing the helpfulness of their top five acquired skills from the course, participants exhibited variability in their emotive tone ($Mdn = 25$ ($IQR = 12-13.25$); $Min = 0$, $Max = 74$). Overall, sentiment distributions in this domain were positively skewed, indicating students typically conveyed stronger emotional engagement when discussing course helpfulness than when evaluating unhelpfulness or skill-specific utility.

Sentiment analysis conducted on acquired skill helpfulness responses from the questionnaire was visualized on a bar plot (see Figure 8). Like the course satisfaction domain, responses exhibited predominantly positive sentiments among participants, followed by trust and anticipation. These findings reaffirm quantitative indicators that students typically held favourable views of the course and the knowledge they acquired, as frequent sentiments illustrated themes of hope and excitement for its ability to assist in their personal and academic growth. Combined with sentiment patterns observed in the course satisfaction domain, these results suggest students maintained an overall optimistic perspective toward both the course and its potential benefits in future academic semesters.

Figure 8*Sentiment Analysis of Acquired Skill Helpfulness Responses*

In summary, sentiment analysis of evaluations on course satisfaction and helpfulness revealed students expressed moderate emotional engagement when articulating their responses. Positive sentiments, alongside trust, anticipation, and joy, characterized most reflections. While students conveyed enthusiasm and confidence in the course's ability to support their personal and academic development, skewed sentiment distributions displayed variability in emotional expression. Although students largely perceived the course positively, these findings suggest that individual interpretation, degrees of emotional investment, and comfort with expressing vulnerability may have influenced how they used emotive language in their textual reflections.

Part 3: Bivariate Associations Between Quantitative Metrics and NLP Metrics

Spearman's rank order correlation test was executed between quantitative and NLP-derived metrics to examine the relationship between quantitative Likert metrics and NLP analytics. Given our associations involve frequency and sentiment counts that are likely to be

positively skewed, being bound by zero on the low end and having no upper limit on the high end, Spearman's correlations were deemed appropriate than Pearson's correlations. Bivariate associations were used to assess the degree of convergent validity offered by frequency and sentiment counts with respect to Likert-scale scores. Two Spearman's rho tables are provided to illustrate correlation coefficients (ρ) and p -values between Likert-scale scores and frequency counts (see Table 4) as well as sentiment counts (see Table 5).

Table 4

Spearman's Rank Order Correlation Test Between Quantitative Likert-scale Scores and NLP-Derived Frequency Counts

Correlation	ρ	p
Domain 1: Course Satisfaction		
Course Enjoyment Rating- Course Satisfaction Term Frequency	0.2	0.017*
Course Meaning Rating- Course Satisfaction Term Frequency	0.22	0.009**
Course Engagement Rating- Course Satisfaction Term Frequency	0.15	0.084
Domain 2: Course Helpfulness		
Course Helpfulness Rating- Course Helpfulness Term Frequency	0.31	< .001***
Course Unhelpfulness Rating- Course Helpfulness Term Frequency	0.11	0.186
Skill 1 Helpfulness Rating- Skill 1 Term Frequency	0.19	0.021*
Skill 1 Usage Rating- Skill 1 Term Frequency	0.06	0.497
Skill 2 Helpfulness Rating- Skill 2 Term Frequency	0.16	0.052
Skill 2 Usage Rating- Skill 2 Term Frequency	0.19	0.026*
Skill 3 Helpfulness Rating- Skill 3 Term Frequency	0.23	0.006**
Skill 3 Usage Rating- Skill 3 Term Frequency	0.27	0.001**
Skill 4 Helpfulness Rating- Skill 4 Term Frequency	0.07	0.409
Skill 4 Usage Rating- Skill 4 Term Frequency	0.12	0.17
Skill 5 Helpfulness Rating- Skill 5 Term Frequency	0.11	0.186
Skill 5 Usage Rating- Skill 5 Term Frequency	0.12	0.151

Note. ρ = Spearman's rho correlation coefficient. * indicates statistical significance at $p < .05$. ** indicates statistical significance at $p < .01$. *** indicates statistical significance at $p < .001$.

Table 5

Spearman's Rank Order Correlation Test Between Quantitative Likert-scale Scores and NLP-Derived Sentiment Counts

Correlation	ρ	p
Domain 1: Course Satisfaction		
Course Enjoyment Rating- Course Satisfaction Sentiment Count	0.35	<.001***
Course Meaning Rating- Course Satisfaction Sentiment Count	0.37	<.001***
Course Engagement Rating- Course Satisfaction Sentiment Count	0.35	<.001***
Domain 2: Course Helpfulness		
Course Helpfulness Rating- Course Helpfulness Sentiment Count	0.32	<.001***
Course Unhelpfulness Rating- Course Helpfulness Sentiment Count	0.27	0.001**
Skill 1 Helpfulness Rating- Skill 1 Sentiment Count	0.34	<.001***
Skill 1 Usage Rating- Skill 1 Sentiment Count	0.21	0.013*
Skill 2 Helpfulness Rating- Skill 2 Sentiment Count	0.57	<.001***
Skill 2 Usage Rating- Skill 2 Sentiment Count	0.52	<.001***
Skill 3 Helpfulness Rating- Skill 3 Sentiment Count	0.56	<.001***
Skill 3 Usage Rating- Skill 3 Sentiment Count	0.69	<.001***
Skill 4 Helpfulness Rating- Skill 4 Sentiment Count	0.33	<.001***
Skill 4 Usage Rating- Skill 4 Sentiment Count	0.39	<.001***
Skill 5 Helpfulness Rating- Skill 5 Sentiment Count	0.43	<.001***
Skill 5 Usage Rating- Skill 5 Sentiment Count	0.51	<.001***

Note. ρ = Spearman's rho correlation coefficient. * indicates statistical significance at $p < .05$. ** indicates statistical significance at $p < .01$. *** indicates statistical significance at $p < .001$.

Quantitative Metrics and NLP Frequency Metrics

Results yielded from Spearman's rank order correlation test between quantitative Likert-scale metrics and NLP frequency metrics, obtained from qualitative data, will be discussed in the following sections. Evidence for or against convergent validity will also be considered.

Course Satisfaction. Spearman's rank order correlation test revealed weak statistically significant associations between course satisfaction Likert-scale ratings and corresponding

consolidated term frequency counts. A small correlation was yielded between course enjoyment ratings and consolidated term counts ($\rho = .20, p < .05$), indicating that higher enjoyment ratings were weakly associated with frequent consolidated term use in course satisfaction responses. A small but slightly stronger correlation was obtained between course meaning ratings and consolidated term counts ($\rho = .22, p < .01$). The correlation between course engagement ratings and consolidated term counts was very weak and failed to reach statistical significance ($\rho = .15, p = .084$), providing insufficient evidence for an association between these metrics. These findings indicate only two associations in the course satisfaction domain rejected the null hypothesis ($H_0: \rho = 0$), whereas the association between course engagement ratings and consolidated term frequency counts failed to reject the null hypothesis ($H_0: \rho = 0$). Thus, the alternative hypothesis ($H_1: \rho \neq 0$) was only partially supported in the domain of course satisfaction.

Course Helpfulness. Spearman's rank order correlation test revealed a significant weak association between course helpfulness Likert-scale ratings and consolidated term frequency counts ($\rho = .31, p < .001$), indicating participants who rated the course as highly helpful tended to use consolidated terms more frequently in their responses. Significant correlations were also found between consolidated term counts for skill 1 helpfulness ratings ($\rho = .19, p < .05$), skill 3 helpfulness ratings ($\rho = .23, p < .01$), and skill 3 usage ratings ($\rho = .27, p < .01$), suggesting that higher perceived value and usage of specific skills acquired from the course were moderately associated with increased mention of consolidated terms. However, most usage ratings for skills 1, 2, 4, and 5, alongside skill 4 and 5 helpfulness ratings, showed no significant correlations with consolidated term frequencies ($p > .05$). These findings demonstrate that these dimensions were not consistently reflected in textual responses, at least through the lens of predefined

terminology. The course unhelpfulness ratings also showed no significant relationship with consolidated term counts ($\rho = .11, p = .186$), potentially due to limited vocabular encapsulation in the consolidated terms list. Like the course satisfaction domain, the alternative hypothesis ($H_1: \rho \neq 0$) was only partially supported as Spearman's rho for relationships between consolidated term frequencies, course helpfulness, and three skill-specific ratings rejected the null hypothesis ($H_0: \rho = 0$). Nevertheless, relationships with course unhelpfulness and most skill usage ratings failed to reject the null hypothesis ($H_0: \rho = 0$).

Convergent Validity. Bivariate associations between quantitative Likert-scale scores and NLP frequency metrics provided only partial evidence in support of convergent validity. Overall, Spearman's rank order correlations revealed statistically significant yet weak associations in several dimensions, suggesting limited alignment between self-reported ratings and consolidated term frequency counts in textual responses tapping the same domains of interest, even when quantitative and qualitative questions were administered simultaneously.

In the course satisfaction domain, small but significant correlations were found between consolidated term counts and Likert-scale ratings of course enjoyment ($\rho = .20, p < .05$) and course meaning ($\rho = .22, p < .01$), whereas the association with course engagement was non-significant ($\rho = .15, p = .084$). In the course helpfulness domain, a moderate correlation emerged between general course helpfulness Likert-scale ratings and consolidated term frequency counts ($\rho = .31, p < .001$). However, skill-specific ratings demonstrated mixed results, with only few yielding statistically significant but weak-to-moderate associations: Skill 1 helpfulness rating and term frequency ($\rho = .19, p < .05$), Skill 2 usage rating and term frequency ($\rho = .19, p < .05$), Skill 3 helpfulness rating and term frequency ($\rho = .23, p < .01$), and Skill 3 usage rating and term

frequency ($\rho = .27, p < .01$). Other skill-related metrics and general course unhelpfulness ratings ($\rho = .11, p > .05$) failed to demonstrate significant associations.

Collectively, these findings indicate that NLP frequency metrics provided insufficient evidence for convergent validity with quantitative Likert-scale metrics. The limited validity observed may be due, in part, to how constructs were operationalized in the consolidated

Quantitative Metrics and NLP Sentiment Metrics

Results yielded from Spearman's rank order correlation test between quantitative Likert-scale metrics and NLP sentiment metrics, obtained from qualitative data, will be discussed in the following sections. Evidence for or against convergent validity will also be considered.

Course Satisfaction. Spearman's rank order correlation test revealed significant moderate associations between all Likert-scale course satisfaction ratings and corresponding sentiment counts. Course enjoyment ($\rho = .35, p < .001$), course meaning ($\rho = .37, p < .001$), and course engagement ($\rho = .35, p < .001$) were all significantly correlated with sentiment expression, suggesting that higher ratings of enjoyment, meaning, and engagement coincide with greater sentiment expression in textual responses. These findings indicate that all associations between quantitative Likert-scale metrics and NLP sentiment metrics rejected the null hypothesis ($H_0: \rho = 0$). Thus, the alternative hypothesis ($H_2: \rho \neq 0$) was supported.

Course Helpfulness. Spearman's rank order correlation test revealed significant associations between all course helpfulness Likert-scale ratings and sentiment counts. Course helpfulness ($\rho = .32, p < .001$) and course unhelpfulness ($\rho = .27, p < .001$) were moderately correlated with sentiment expression, suggesting that stronger perceptions of course helpfulness or unhelpfulness correspond with greater emotional expression in textual responses. Similarly,

significant correlations emerged between skill-specific ratings and sentiment counts. Skill 1 helpfulness ($\rho = .34, p < .001$) and usage ($\rho = .21, p < .001$) showed moderate and weak associations, respectively, indicating that perceived usefulness of skills influenced emotional expression. Stronger correlations were observed for skill 2 helpfulness ($\rho = .57, p < .001$) and usage ($\rho = .52, p < .001$), alongside skill 3 helpfulness ($\rho = .56, p < .001$) and usage ($\rho = .69, p < .001$), suggesting higher sentiment expression among participants who considered these skills more helpful and applicable. Furthermore, skill 4 helpfulness ($\rho = .33, p < .001$) and usage ($\rho = .39, p < .001$) exhibited moderate correlations, while skill 5 helpfulness ($\rho = .43, p < .001$) and usage ($\rho = .51, p < .001$) were also significantly associated with sentiment counts. These findings indicate that higher skill helpfulness and usage ratings consistently correspond with greater sentiment expression in qualitative responses, reinforcing the degree to which skill acquisition shapes participants' textual reflections on course effectiveness. Like the course satisfaction domain, all associations between quantitative Likert-scale metrics and NLP sentiment metrics supported the alternative hypothesis ($H_2: \rho \neq 0$), as they rejected the null hypothesis ($H_0: \rho = 0$).

Convergent Validity. Bivariate associations between quantitative Likert-scale scores and NLP sentiment metrics provided consistent evidence supporting convergent validity. Spearman's rank order correlations revealed statistically significant and moderate to strong associations across all dimensions, indicating closer alignment between self-reported ratings and sentiment expression in textual responses compared to NLP frequency metrics.

In the course satisfaction domain, significant moderate correlations emerged between sentiment counts and Likert-scale ratings of course enjoyment ($\rho = .35, p < .001$), course meaning ($\rho = .37, p < .001$), and course engagement ($\rho = .35, p < .001$). These results suggest that participants who reported greater satisfaction were more likely to exhibit stronger sentiment

expression in their textual feedback. In the course helpfulness domain, overall Likert-scale helpfulness ratings were moderately correlated with sentiment counts ($\rho = .32, p < .001$) alongside unhelpfulness ratings ($\rho = .27, p < .001$), indicating that a stronger view of the course's impact was accompanied by higher emotive writing in qualitative responses. Skill-specific ratings yielded similarly moderate-to-strong results in skill 1 helpfulness ($\rho = .34, p < .001$) and usage ($\rho = .21, p < .05$), skill 2 helpfulness ($\rho = .57, p < .001$) and usage ($\rho = .52, p < .001$), skill 3 helpfulness ($\rho = .56, p < .001$) and usage ($\rho = .69, p < .001$), skill 4 helpfulness ($\rho = .33, p < .001$) and usage ($\rho = .39, p < .001$), and skill 5 helpfulness ($\rho = .43, p < .001$) and usage ($\rho = .51, p < .001$). These findings suggest that perceived advantages and application of skills acquired from the course correspond with the emotional tone embedded in students' qualitative responses.

Together, these results denote that NLP sentiment metrics, garnered from the NRC lexicon in the *syuzhet* package in R, provided sufficient evidence for convergent validity with quantitative Likert-scale metrics. Compared to NLP frequency metrics, sentiment counts are more likely to reliably align with self-reported ratings by way of effectively capturing emotive dimensions of academic experiences reported through course evaluations.

Part 4: Comparison of Quantitative Metrics and NLP Metrics

In this final part, results obtained from LDA topic modeling, sentiment shift analysis, and NRC clustering for the domains of interest will be selectively discussed to illustrate the incremental utility of these NLP-derived insights compared to traditional quantitative Likert-scale metrics.

Latent Dirichlet Allocation (LDA)

considered typical words linked to manners in which students may rate their degrees of enjoyment and engagement with a course. However, less frequent terms, such as “mental,” “health,” “life,” “sleep,” “stress,” “wellbeing,” “perfectionism,” and “relationship” also emerged, emphasizing deeper experiential dimensions of student learning that may not be explicitly captured through standard Likert-scale items. These NLP-derived insights suggest that students’ psychological and physical health, personal challenges, and life circumstances were often intertwined with their academic experiences, likely shaping how they perceived and evaluated course satisfaction. Notably, LDA also detected several French words, such as “pas,” “qui,” and “sur,” indicating that some students chose to express their reflections in French despite the course evaluation being delivered in English. This NLP-derived insight from LDA demonstrates that mixed-methods research implementing NLP metrics can be inclusive by capturing multilingual expression, even when researchers depend upon quantitative metrics developed in English.

Moreover, LDA offered additional insight compared to findings garnered from frequency analysis on acquired skill helpfulness responses. Although both word clouds illustrated similar terminology regarding helpfulness and skill-specific labels, such as “course helpfulness” versus “helped,” “stress management” versus “stress,” and “study habits” versus “studying,” LDA differentiated topics that were generalized by consolidated terms. For instance, the word cloud visualizing frequency analysis findings in this domain (see Figure 6) reported “stress management,” which in our consolidated terms list (see Figure 2), encompassed words such as “stress,” “relaxation breathing,” and “mindfulness.” However, the LDA generated word cloud disjointed “stress” from “breathing” and “mindfulness,” parsing them into distinct topics based on contextual co-occurrence across responses. Consequently, this thematic separation enabled a

more complex understanding of how students experienced and applied stress-management-related skills. LDA suggested that “mindfulness” was presumably discussed in the context of emotional regulation, whereas “breathing” may have been more frequently referred to in isolated stress episodes or testing anxiety. Therefore, LDA provided a more comprehensive description of acquired skill applications and perceived usefulness in students’ textual responses, offering additional probabilistic information that was otherwise obscured by the restrictive, predefined consolidated terminology utilized for frequency analysis.

In the domain of course helpfulness, frequently occurring terms such as “helped,” “skill,” “studying,” “breathing,” “notes,” and “recall” aligned with practical academic strategies and content taught in the course. However, more affective and experiential terms, including “feel,” “calm,” “allowed,” “easier,” and “life,” also appeared, indicating that students may have evaluated the course not only for academic utility but also its contribution to emotional regulation and everyday functioning. This pattern reflects the broader impact of instructional content on students’ personal lives, suggesting that perceived course helpfulness can include academic practicality alongside psychological and emotional value. In short, these findings demonstrated the unique value of LDA in uncovering contextual, experiential, and emotionally nuanced dimensions of student experience that traditional quantitative metrics alone are unlikely to capture. Thus, observations of LDA topic modeling results offer support for the alternative hypothesis (H_3) that NLP-derived insights provide incremental utility beyond quantitative Likert-scale scores by offering a more comprehensive understanding of how students engaged with and evaluated their learning.

Sentiment Shift Analysis

Sentiment shift analysis tracked changes in sentiment counts over different segments of the questionnaire. In the course helpfulness domain, sentiment shift scores (“what other kinds of help would have helped you better this semester at university?” minus “I would like to know how this course personally helped you, or didn't help you...?”) revealed predominantly negative Δ (delta) values across participants’ sentiment profiles, indicating increased negative emotionality when describing unmet helpfulness needs compared to helpful aspects of the course (see Table 6). The most substantial declines occurred for positive emotions ($\Delta = -43$), trust ($\Delta = -11$, -22), and anticipation ($\Delta = -20$). Four participants in the condensed subset demonstrated neutral shifts ($\Delta = 0$) in a minimum of five emotion categories, suggesting stable sentiment profiles across this helpfulness evaluation context. Isolated positive shifts emerged for joy ($\Delta = +4$), trust ($\Delta = +4$, $+2$). However, anger and disgust ranged from -3 to $+1$, showing minimal variation, indicating these emotions were less salient in differentiating between course helpfulness and unmet helpfulness needs.

Table 6*Sentiment Shift Scores Between Unmet Helpfulness Needs and Course Helpfulness Responses*

Participant ID	Anger Δ	Anticipation Δ	Disgust Δ	Fear Δ	Joy Δ	Sadness Δ	Surprise Δ	Trust Δ	Negative Δ	Positive Δ
10	-2	-9	-2	-5	-5	-4	-2	-9	-7	-26
20	0	-8	-1	-1	-11	0	-2	-16	-1	-31
30	0	-5	-1	0	-4	-1	-2	-9	-2	-9
40	0	-2	0	0	0	0	0	-1	0	-3
50	-3	-6	-1	-3	-2	-1	-1	-2	-3	-4
60	0	-1	-1	-2	-1	-2	-2	-2	0	-10
70	0	0	0	0	0	0	0	-1	-1	-3
80	-1	0	1	0	-2	0	0	0	0	0
90	1	1	-1	0	4	1	2	4	1	6
100	0	0	0	-1	1	0	1	0	0	-1
110	-1	-3	0	-1	-5	0	-2	-11	-5	-22
120	1	-4	0	2	-3	-2	-1	-6	-2	-14
130	-2	-20	-3	-7	-13	-4	-8	-22	-11	-43
140	1	0	0	1	0	0	0	2	1	1

Note. Every 10th participant ($n = 10$) was randomly selected from the total sample ($N = 144$) for a custom subset heatmap to convey clarity. Δ (delta) indicates a change or difference between two values. Red demonstrates a decrease ($-\Delta$) in sentiment, white demonstrates no change in sentiment, and green demonstrates an increase ($+\Delta$) in sentiment.

These findings provide support for the incremental value of sentiment shift analysis, offering deeper insight into students' experiences with the course that traditional quantitative Likert-scale metrics fail to capture by virtue of their design. While these ratings quantified general helpfulness concerning the course and acquired skills, sentiment shift analysis revealed the emotional contrast in textual responses as evaluative prompts shifted from help received to what was still needed or unmet by the course. Although the course was rated as highly helpful, NLP-derived sentiment shift analysis insights suggested that gaps in student support were likely

present. Overall, this NLP metric provided a dynamic understanding of students' affective experiences and unmet needs, thus providing additional support for the alternative hypothesis (H_3) by adding a valuable emotional dimension to reflections on course effectiveness.

NRC Clustering

NRC clustering is a partition-based, non-hierarchical k-means clustering algorithm utilized to group aggregated sentiment findings for both domains of interest to observe macro-level response patterns.

NRC cluster analysis of sentiment counts for course satisfaction responses revealed identical emotive patterns identified through sentiment analysis (see Table 7). Positive emotions dominated responses, with the highest sum across five clusters ($\Sigma = 2,612$, $k = 5$), followed by trust ($\Sigma = 1,427$, $k = 4$), anticipation ($\Sigma = 1,010$, $k = 3$), and joy ($\Sigma = 741$, $k = 3$). Negative emotions were less prevalent ($\Sigma = 639$, $k = 3$). However, fear ($\Sigma = 381$, $k = 2$) and sadness ($\Sigma = 312$, $k = 2$) were more prominent than anger ($\Sigma = 194$, $k = 1$) and disgust ($\Sigma = 116$, $k = 1$). These results indicate that satisfied students typically expressed optimism, confidence, and foresight regarding course satisfaction. Although negative evaluations were infrequent, students likely focused on passive emotions rather than active hostility when negative emotions were present.

Table 7*NRC Clustering of Course Satisfaction Responses*

Sentiment Category	Σ	k
positive	2612	5
trust	1427	4
anticipation	1010	3
joy	741	3
negative	639	3
fear	381	2
sadness	312	2
surprise	267	2
anger	194	1
disgust	116	1

Note. Σ refers to the sum of sentiment counts across all participants; k refers to the number of clusters per sentiment category.

NRC cluster analysis of sentiment counts revealed similar patterns in participants' responses concerning the helpfulness of skills acquired from the course (see Table 8). Like course satisfaction, positive emotions dominated responses, with the highest sum across five clusters ($\Sigma = 2,436$, $k = 5$), followed by trust ($\Sigma = 1,235$, $k = 4$), anticipation ($\Sigma = 996$, $k = 4$), and joy ($\Sigma = 587$, $k = 3$). Negative emotions were less prevalent ($\Sigma = 694$, $k = 3$), with fear ($\Sigma = 345$, $k = 2$), and sadness ($\Sigma = 295$, $k = 2$) being more prominent than anger ($\Sigma = 189$, $k = 1$) or disgust ($\Sigma = 126$, $k = 1$). These findings indicated that participants primarily associated skill acquisition with optimism, confidence, and anticipation. When present, negative emotions may have reflected passive concerns over dissatisfaction.

Table 8*NRC Clustering of Acquired Skill Helpfulness Responses*

Sentiment Category	Σ	k
positive	2436	5
trust	1235	4
anticipation	996	4
negative	694	3
joy	587	3
fear	345	2
sadness	295	2
surprise	225	1
anger	189	1
disgust	126	1

Note. Σ refers to the sum of sentiment counts across all participants; k refers to the number of clusters per sentiment category.

Overall, the NLP-derived insights from NRC clustering analysis revealed emotional patterns consistent with those identified through basic sentiment analysis, with positive emotions, trust, anticipation, and joy dominating students' course evaluations. Consequently, NRC clustering contributed minimal added value compared to previous NLP metrics, as it primarily reiterated emotional trends captured through sentiment counts discussed in section two. While this NLP technique may offer slight supplementary insight by structuring emotional patterns, it provided limited incremental utility beyond what was conveyed through NLP-derived sentiment analysis and quantitative Likert-scale insights, thus failing to reject the null hypothesis (H_0).

Discussion

The current study was conducted to assess the unique benefits of NLP-analytic methods, not only in comparison to each other but also against more traditional quantitative methods, such as Likert-scale ratings. Results of the study showed that quantitative and NLP methods provide important information concerning the extent to which course feedback indicated students were largely satisfied with the course and considered it extremely helpful. Likert-scale responses demonstrated consistently high ratings of satisfaction and helpfulness. Similarly, NLP-derived insights, including sentiment analysis and NRC clustering, revealed that students expressed predominantly positive emotions when discussing enjoyment and benefits attained from the course. These findings were reaffirmed by frequency analysis and LDA topic modeling, reinforcing students' positive evaluations while illustrating their use of explicit language reflecting themes of satisfaction and helpfulness in textual responses.

Comparison of NLP Methods

Although both analytical approaches suggested that the course was enjoyable and beneficial, results from the analysis of quantitative metrics and NLP-metrics produced slightly different findings. First, frequency analysis revealed that students often associated course satisfaction and helpfulness with predefined terms tapping for themes of practical benefits, academic support, and personal development offered by instructional content and design. However, variability in terminology usage suggests students likely relied upon personalized language to describe their experiences with the course. Second, sentiment analysis indicated that students had strong positive perceptions of course satisfaction and helpfulness, signifying enthusiasm and confidence in the course's ability to foster personal and academic growth.

Third, sentiment shift analysis suggested that gaps in student support were present as positive emotions declined between responses concerning how the course was beneficial and other forms of help that could have improved their academic experience. Fourth, NRC clustering yielded identical emotive patterns as sentiment analysis, demonstrating that students primarily associated course satisfaction and skill acquisition with optimism, trust, and anticipation. Finally, LDA topic modeling revealed that students associated course satisfaction and helpfulness with both expected academic terms (e.g., “helped,” “enjoyed,” “learning,” “skill,” “studying”) and unexpected experiential or emotional topics (e.g., “mental,” “health,” “life,” “feel,” “calm”), introducing subtle personal dimensions of learning.

These differences highlight numerous advantages of using NLP-based methods to evaluate qualitative feedback. While traditional quantitative metrics such as Likert-scale scores provide standardized summaries of student perceptions, NLP-metrics obtained through sentiment analysis, sentiment shift analysis, and LDA topic modeling capture contextual, experiential, and affective nuances, alongside desired academic themes, embedded in open-ended qualitative responses (Blei et al., 2003; Griffith et al., 2024; Mohammad & Turney, 2013). These methods facilitate the detection of unique emotional tones, sentimental shifts, and multidimensional themes that may otherwise be overlooked when relying solely on quantitative ratings (Yuan & Hu, 2024). Offering greater depth to course evaluations, NLP techniques enable researchers and instructors to accomplish a comprehensive understanding of student feedback, particularly regarding the influential interconnections of personal and academic experiences within learning environments. Therefore, observations supported our third hypothesis that NLP-derived insights provide incremental utility beyond quantitative Likert-scale scores, specifically metrics garnered from sentiment analysis, sentiment shift analysis, and LDA topic modeling.

The current study obtained several compelling findings demonstrating the advantages of incorporating machine learning text-parsing tools in mixed methods research designs. Although frequency analysis demonstrated weak correlations with Likert-scale metrics, with only half reaching statistical significance, it did provide an efficient manner to conduct qualitative inquiry mimicking traditional thematic *content* analysis (Ayre & McCaffery, 2022; Braun & Clarke, 2006). While predefined terminology was highly limited and restricted nuance detection in student course evaluations across the domains of course satisfaction and course helpfulness, these constraints are likely due to linguistic or syntax sensitivities. Specifically, it can be difficult to anticipate every word, synonym, and potential typo a student may write when describing a broad topic of interest. Accordingly, further research on methods whereby researcher-defined terms can be more extensive is worthwhile. For instance, subsequent studies may benefit from creating a series of terminology lists specific to each domain of interest, rather than using a single, general set of terms for all frequency analyses. Nevertheless, Spearman's rank order correlation test did capture weak associations between frequency counts and quantitative scores, half of which reached statistical significance, providing only partial support for our first hypothesis that a relationship exists between quantitative Likert-scale scores and NLP-derived frequency counts. Hence, it is not the case that frequency analysis itself is a useless NLP-based technique; rather, comprehensive brainstorming and potential collaboration with students could improve how terms are operationalized for each topic. Consequently, these efforts could improve the relationship between NLP-derived frequency counts and Likert-scale scores.

However, being capable of evaluating several languages, even those unanticipated by researchers and professors who administer open-ended questionnaires, such as course evaluations, LDA demonstrated promise as an alternative to frequency analysis, which assumed

all responses would be written in English. In the context of bilingual or multi-lingual academic institutions, this form of machine learning text parsing could facilitate nuanced understandings of student perspectives regardless of which language they are most proficient in. Furthermore, LDA did tap for similar terms as frequency analysis, suggesting the implementation of both text parsing methodologies could be advantageous when attempting to operationalize predefined terms and narrow the scope of mixed methods inquiry. Finally, LDA did demonstrate *thematic* saturation by way of codebook redundancy, as many variations of the same terms were present in results tables and exported word clouds, indicating that additional topics could not be identified (Hennink et al., 2017; Saunders et al., 2018). This could be due to the intentional coding of 18 topics, each with five terms, to compare alphanumeric depth with the 18 predefined terms required for frequency analysis. Thus, NLP-derived LDA findings provided additional support for our third hypothesis, demonstrating incremental utility beyond quantitative scores and, unexpectedly, frequency counts as well.

Convergent Validity

Another main objective of the present study was to examine the extent to which quantitative and NLP metrics converge. Although they are investigating a similar domain, very few studies have formally explored the validity of NLP-based metrics with respect to traditional quantitative metrics (Weissman et al., 2019). Given the growing popularity of NLP methods across numerous research disciplines, it is important from a psychometric perspective to validate the degree to which these analytic methods produce converging results. Weissman et al. (2019) conducted one of few studies examining the issue, assessing concurrent validity by comparing NLP-derived sentiment results in R to mortality risk within the same 24-hour period. They found evidence of convergence across sentiment analysis packages (e.g., *AFINN*, *sentimentr*, *CoreNLP*)

regarding a shared outcome (i.e., risk). Metrics yielded through both the *sentimentr* and *CoreNLP* R packages demonstrated an unacceptable degree of convergent validity.

In the current study, convergence was examined between NLP-metrics (i.e., frequency analysis and sentiment analysis) and traditional quantitative methods (i.e., Likert-scale ratings), rather than between two different NLP-metrics. Arguably, testing the degree of convergence between quantitative and qualitative analyses of a similar domain (i.e., helpfulness or satisfaction) lies at the heart of mixed-methods research. Psychometric theory stipulated that these two methods assessing the same content domain should be related, which is analogous to testing part of a multi-trait, multi-method matrix (Campbell & Fiske, 1959). Despite the value imparted by validation, there are very few, if any, existing investigations of the degree to which NLP and quantitative methods converge (Weissman et al., 2019). However, results from the current study provide evidence that some methods, namely sentiment analysis, show better evidence of convergence than others (i.e., frequency analysis). Specifically, our results show that sentiment analysis through the *syuzhet* R package yielded consistent and statistically significant, moderate to strong correlations with Likert-scale ratings when assessing the domains of course satisfaction and helpfulness. As a result, these findings demonstrated an acceptable degree of convergent validity, thus supporting our second hypothesis that a correlation exists between quantitative Likert-scale scores and NLP-derived sentiment counts.

Although frequency analysis fell short in demonstrating concurrent validity, sentiment analysis through R's *syuzhet* package provided meaningful evidence for convergent validity. These significant findings point to sentiment analysis, particularly through the NRC lexicon, as capable of evaluating the same construct(s) as well-established and validated quantitative Likert-scales, at least when administered simultaneously in the context of course evaluations. Capturing

emotional complexity in satisfaction and helpfulness response patterns, sentiment analysis is a vital tool in understanding the depth of human experiences conveyed in textual data (Mohammad & Turney, 2013). These implications extend beyond academic spheres, as NRC sentiment analysis could provide crucial insights into the broader population's experience within various institutions, interventions, and social contexts. As a step toward validating machine learning text-parsing tools in R syntax, sentiment analysis, and its shift or clustering offshoots, could be employed when interpreting diverse life experiences, thereby informing administrative change in health care, academia, rehabilitation, and anywhere in between (Biró et al., 2023; Kastrati et al., 2021; Saini et al., 2019).

Implications for Program Evaluations

Results have several implications and benefits for both researchers evaluating programs and courses that solicit written feedback from individuals participating in those programs, as well as for policymakers utilizing that information.

First, findings illustrate the value of NLP tools in processing large volumes of qualitative data in a quick and scalable manner. While the initial development and refinement stages of the R script were time-consuming and demanding, once finalized, it permitted automated qualitative data analysis in under a minute. Compared to traditional labour-intensive and tedious qualitative approaches, the efficiency of conducting NLP in R serves as a major methodological advantage (Dhakal, 2022; Zamawe, 2015). Capable of rapidly analyzing thousands of textual responses, researchers can incorporate qualitative inquiry into large-scale evaluations in a time-conservative manner (Alqahtani et al., 2023; Zaki et al., 2023). These implications reach policymakers, providing more immediate access to nuanced, data-informed perspectives on program

advantages, disadvantages, and themes in participant experiences. Consequently, influential decision-making processes can be based on both qualitative and quantitative information.

Within the context of course evaluations, professors could quickly and efficiently utilize sentiment analysis processing, alongside LDA topic modeling, to pay necessary attention to how students feel in their classrooms, their degrees of satisfaction with instructional and assessment strategies, and the extent to which courses are perceived as helpful in fostering personal and academic growth among students. Given that course evaluations are administered annually across many secondary and tertiary institutions, it is worthwhile to not only assess them comprehensively but also to understand reports of diverse lived experiences to better invest in student success (Kastrati et al., 2021; Shaik et al., 2022; Yuan & Hu, 2014).

Second, NLP approaches ensure the mitigation of rater bias and fatigue when evaluating and classifying feedback from numerous respondents (Cranton & Santor, 2024; Kang et al., 2020). Although some NLP metrics, such as consolidated term frequency counts, are more open to bias as these terms are stipulated by the researcher, sentiment analysis, LDA, and word clouds are not. These data-driven techniques offer more objective insights as they rely on natural language structure rather than researcher assumptions.

Third, once processed, these methods permit more efficient comparison across programs. Currently, qualitative feedback provided by my students attending many universities and colleges throughout North America is rarely used systematically and compared across courses or programs (Yuan & Hu, 2014). The lack of attention to student course evaluations is largely a consequence of the complexities of processing and analyzing these datasets (Parks & Peters, 2023). Results from this study showed that NLP-based qualitative analysis can be performed feasibly and with validity alongside other well-established quantitative methods. Moreover,

using the open-access GitHub repository of the present study's NLP script in R, professors can make simple alterations, import student data, and assess it extensively through machine learning text-parsing methods. As a result, student voices can be more actively centered in faculty-wide administrative decisions influencing instructional techniques, program policies, and course sequences.

Fourth, alongside its value in academic research and evaluation, NLP can inform institutional resources by increasing recognition of student needs concerning psychological, academic, and financial stressors. Using sentiment and topic modeling strategies to identify such patterns in open-ended feedback can prompt critical conversations, allowing individuals to both advocate for their necessities and have their voices heard across institutions worldwide.

Finally, NLP techniques offer several advantages compared to proprietary qualitative data analysis programs, such as NVivo. Although widely used by qualitative researchers worldwide, exact descriptions and evaluation of these methods are unavailable to researchers for review (Dhakal, 2022; Zamawe, 2015). Conversely, the current study's methodological design was based on open-source packages in R that have been developed, refined, and scrutinized by the broader research community (Welbers et al., 2017). In addition to being cost-effective, this approach is likely to be more transparent, reproducible, and open to innovation.

While implementing NLP in R can be a laborious learning curve, especially for those without a background in programming, there are several resources available (e.g., online forums, CRAN documentation, video tutorials, and artificial intelligence) to expedite the process. As a prospective solution to methodological underutilization and limitation, NLP in R can bridge the gap between qualitative depth and quantitative precision in mixed methods research, allowing

everyone with access to such findings the ability to become more cognizant of phenomena, issues, and experiences embedded in the world around them.

We do not wish to replace traditional qualitative approaches with NLP, as empathy is a vital component to understanding the aforementioned factors. Nevertheless, when confronted with the choice between discarding an individual's narrative due to time constraints and using machine learning text-parsing tools to expeditiously evaluate them, NLP can tip the balance scale toward honouring essential feedback. Furthermore, a computer can only perform what you instruct it to do, meaning the role of empathy remains involved in the process of developing NLP syntax in R. While NLP's automation capabilities can improve qualitative and mixed methods research, it cannot substitute the indispensable role of the researcher's insight and expertise.

Limitations

Although the current study yielded promising findings, it carries limitations. First, although our design greatly reduced rater bias, there remains room for it in the coding and interpretation processes. Defining consolidated terms for frequency analysis can present a degree of bias. Interpreting how topics obtained from LDA should be labelled based on their corresponding terms can sway reporting and dissemination. Similarly, deciphering NRC clustering and sentiment shift results, based on how the experimenter codes their functions and analytical pairs, can foster an element of bias. Second, the NLP syntax in R was developed without pre-existing programming knowledge prior to this project, meaning R was learned throughout the process. While the code ran smoothly, there may be errors in reasoning and organization influencing processing speed, analytical depth, and reliability. Subsequent research can benefit from collaborating with experts in computer science or qualitative research in R.

Third, during the Microsoft Excel exporting process, the code generated roughly 56 worksheets in the main workbook. This step rendered the process of interpreting findings relatively time-consuming compared to the previous methodological pace. Hence, it is useful to narrow how much data is worth reporting before exportation. However, exporting findings to Microsoft Excel may not be necessary, as additional time can be invested in coding APA 7th edition formatting when generating tables and figures in R. Future work may explore more automated data visualization strategies to reduce minimal formatting demands and improve efficiency. Fourth, to adhere to the scope of our study, we did not explore all potential domains offered by course reflections and only explored a fraction of qualitative responses. As a result, we likely overlooked important factors involved in retrospective accounts of academic experiences. Additionally, because we developed R syntax solely for the context of FSS 1150 course evaluations, the generalizability of our findings is limited to undergraduate students enrolled in this course at the University of Ottawa. Thus, future research should develop more universalized syntax to study human experiences in more diverse contexts beyond academia by incorporating and further validating the NLP-derived metrics presented in this study.

Fifth, coding rationale concerning stop-word removal and the number of topics, and related terms, chosen for LDA topic modeling may have influenced our findings. Although LDA demonstrated some thematic saturation, terms such as “way,” “put,” and “etc” were detected, suggesting that inessential words were poorly filtered out when defining the ‘Perform_LDA’ function in R. Accordingly, subsequently studies would benefit from manually coding additional filters before running analyses to improve topic coherence and thematic clarity. Although we realized our aim to establish the general utility and partial validity of NLP methods in R, these

machine learning text-parsing tools must be further investigated, refined, and compared to emerging R package publications to best capture human experiences across populations.

Conclusion

The present study demonstrated the potential of NLP to improve mixed methods research, exemplifying its ability to expeditiously analyze large qualitative datasets while complementing traditional quantitative methodologies. Despite limitations in frequency analysis, potential rater bias in coding, sentiment analysis through the NRC lexicon alongside LDA topic modeling provided valuable insight into student course evaluations that would otherwise be neglected. Our findings reinforced the feasibility and utility of conducting NLP in R across institutional and academic settings. By refining R script structure, expanding predefined terminology, and perfecting LDA filtering, future research can enhance the precision and interpretability of NLP-derived metrics. As a result, more nuanced understandings of lived experiences can be achieved. The integration of machine learning text-parsing tools in course evaluations and beyond holds promise for amplifying voices in decision-making processes, all while bridging the gap between qualitative depth and quantitative precision in mixed methods psychological research. While machine learning cannot replace researcher empathy and expertise, it is a powerful mechanism for capturing, analyzing, and advocating for the complexities of human experiences within institutional and societal contexts.

Declaration of Generative AI and AI-Assisted Technologies in the Coding Process

ChatGPT was used as an assistive tool in various stages of the coding process when developing the NLP script in R (OpenAI, 2025). AI-generated coding suggestions were sought for troubleshooting and resolving syntax errors, debugging code, verifying the logical structure

of functions, and obtaining alternative coding approaches to improve script efficiency and readability. Rather than generating the full NLP script in R, AI responses were used to inform and refine existing code written by the researcher.

Interactions with ChatGPT were iterative as the researcher provided code snippets along with specific questions, context clues, and errors for AI to offer adequate suggestions, corrections, or explanations on how to solve these issues (OpenAI, 2025). These outputs were critically assessed using prior knowledge of R programming, logic, and syntax obtained from coding experience and reputable sources, including the CRAN R-Project website, academic articles, R package publications, video tutorials, and online forums. Suggestions were tested and modified to ensure alignment with the intended script functionality and research objectives. A detailed example of a troubleshooting interaction with ChatGPT to correct a minor syntax error in R includes the issue in the R environment, the AI-generated response, and the code revision (see Figure 1B).

References

- Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., ... & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in social and administrative pharmacy*, 19(8), 1236–1242.
<https://doi.org/10.1016/j.sapharm.2023.05.016>
- Ayre, J., & McCaffery, K. J. (2022). Research Note: Thematic analysis in qualitative research. *Journal of Physiotherapy*, 68(1), 76–79. <https://doi.org/10.1016/j.jphys.2021.11.002>
- Biró, A., Cuesta-Vargas, A. I., & Szilágyi, L.. (2023). Precognition of mental health and neurogenerative disorders using AI-parsed text and sentiment analysis. *Acta Universitatis Sapientiae. Informatica*, 15(2), 359–403. <https://doi.org/10.2478/ausi-2023-0022>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5), 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597.
<https://doi.org/10.1080/2159676X.2019.1628806>
- Braun, V., Clarke, V., Hayfield, N., Terry, G. (2019). Thematic analysis in qualitative research. In P. Liamputtong (Ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 843–860). Springer. https://doi.org/10.1007/978-981-10-5251-4_103

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
<https://doi.org/10.1037/h0046016>
- Chang, T., DeJonckheere, M., Vydiswaran, V. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research*, 15(3), 398–412.
<https://doi.org/10.1177/15586898211021196>
- Chowdhary, K. R. (2020). Natural Language Processing. In *Fundamentals of Artificial Intelligence*. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19
- Cranton, C. A., & Santor, A. D. (2024, June 22). *Implementing natural language processing (NLP) in R to evaluate knowledge acquisition from a course on success and well-being* [Poster presentation]. CPA 2024 Convention, Ottawa, ON, Canada. https://cpa.ca/docs/File/Convention/2024/CPA_2024_Event_Program_9.pdf
- Dhakal, K. (2022). NVivo. *Journal of the Medical Library Association*, 110(2), 270–272.
<https://doi.org/10.5195/jmla.2022.1271>
- Fàbregues, S., Escalante-Barrios, E. L., Molina-Azorin, J. F., Hong, Q. N., Verd, J. M., & Perzynski, A. T. (2021). Taking a critical stance towards mixed methods research: A cross-disciplinary qualitative secondary analysis of researchers' views. *PloS One*, 16(7), e0252014. <https://doi.org/10.1371/journal.pone.0252014>
- Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54. <https://doi.org/10.18637/jss.v025.i05>

- Feinerer, I., & Hornik, K. (2024). *tm: Text mining package* (Version 0.7-15) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=tm>
- Fellows, I. (2018). *wordcloud: Word clouds* (Version 2.6) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=wordcloud>
- Griffith, F. J., Ash, G. I., Augustine, M., Latimer, L., Verne, N., Redeker, N. S., ... & Fucito, L. M. (2024). Natural language processing in mixed-methods evaluation of a digital sleep-alcohol intervention for young adults. *npj Digital Medicine*, 7(1), 1–12. <https://doi.org/10.1038/s41746-024-01321-3>
- Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1–30. <https://doi.org/10.18637/jss.v040.i13>
- Grün, B., & Hornik, K. (2024). *topicmodels: Topic models* (Version 0.2-17) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=topicmodels>
- Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research*, 27(4), 591–608. <https://doi.org/10.1177/1049732316665344>
- Jockers, M. L. (2015). *Syuzhet: Extract sentiment and plot arcs from text* [R package]. GitHub. <https://github.com/mjockers/syuzhet>
- Kabir, A. I., Ahmed, K., & Karim, R. (2020). Word cloud and sentiment analysis of Amazon earphones reviews with R programming language. *Informatica Economica*, 24(4), 55–71. <https://doi.org/10.24818/issn14531305/24.4.2020.05>

- Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 139–172. <https://doi.org/10.1080/23270012.2020.1756939>
- Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences*, 11(9), 3986. <https://doi.org/10.3390/app11093986>
- Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., & Nithya, M. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16.
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
- Kumar, A., & Paul, A. (2016). *Mastering text mining with R*. Packt Publishing Ltd.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2024). *cluster: Cluster analysis basics and extensions* (Version 2.1-8) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=cluster>
- Maxwell, J. A. (2016). Expanding the History and Range of Mixed Methods Research. *Journal of Mixed Methods Research*, 10(1), 12–27. <https://doi.org/10.1177/1558689815571132>
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>

- Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes* (Version 1.1-3) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=RColorBrewer>
- Oliveira, M., Bitencourt, C. C., Santos, A. C. M. Z. dos, & Teixeira, E. K. (2015). Thematic Content Analysis: Is There a Difference Between the Support Provided by the MAXQDA® and NVivo® Software Packages? *Revista de Administração Da UFSM*, 9(1), 72–82. <https://doi.org/10.5902/1983465911213>
- Onwuegbuzie, A. J., Mallette, M. H., & Mallette, K. M. (2022). A 41-year history of mixed methods research in education: A mixed methods bibliometric study of published works from 1980 to 2021. *Journal of Mixed Methods Studies*, 6, 7–56. <https://doi.org/10.59455/jomes.2022.6.2>
- OpenAI. (2025). *ChatGPT* (Jan 15 Version) [Large language model]. <https://chatgpt.com/>
- Parks, L., & Peters, W. (2023). Natural language processing in mixed-methods text analysis: A workflow approach. *International Journal of Social Research Methodology*, 26(4), 377–389. <https://doi.org/10.1080/13645579.2021.2018905>
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing* (Version 4.4.2). [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- RStudio Team. (2020). *RStudio: Integrated development environment for R* (Version 2024.12.0-467) [Computer software]. RStudio, PBC. <https://posit.co/>
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and

operationalization. *Quality & Quantity*, 52(4), 1893–1907.

<https://doi.org/10.1007/s11135-017-0574-8>

Saini, S., Punhani, R., Bathla, R., & Shukla, V. K. (2019, April). Sentiment analysis on twitter data using R. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 68–72). IEEE.

Sawicki, J., Ganzha, M., & Paprzycki, M. (2023). The state of the art of natural language processing—A systematic automated review of NLP literature using NLP techniques. *Data Intelligence*, 5(3), 707–749. https://doi.org/10.1162/dint_a_00213

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access*, 10, 56720–56739. <https://doi.org/10.1109/ACCESS.2022.3177752>

Schauberger, P., & Walker, A. (2024). *openxlsx: Read, write and edit xlsx files* (Version 4.2.7.1) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=openxlsx>

Urbanek, S. (2022). *png: Read and write PNG images* (Version 0.1-8) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=png>

Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R., & Halpern, S. D. (2019). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of biomedical informatics*, 89, 114–121. <https://doi.org/10.1016/j.jbi.2018.12.001>

- Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, 11(4), 245–265. <https://doi.org/10.1080/19312458.2017.1387238>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. [R package]. Springer-Verlag. <https://ggplot2.tidyverse.org>
- Wickham, H. (2023). *stringr: Simple, consistent wrappers for common string operations* (Version 1.5.1) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=stringr>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., Hester, J., & Bryan, J. (2024). *readr: Read rectangular text data* (Version 2.1.5) [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=readr>
- Yuan, B., & Hu, J. (2024). *An Exploration of Higher Education Course Evaluation by Large Language Models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2411.02455>
- Zaki, N., Turaev, S., Shuaib, K., Krishnan, A., & Mohamed, E. (2023). Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. *Education and Information Technologies*, 28(12), 16723–16742. <https://doi.org/10.1007/s10639-023-11877-4>

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis:

Evidence-based reflections. *Malawi Medical Journal*, 27(1), 13–15.

<https://doi.org/10.4314/mmj.v27i1.4>

Appendix A

Course Evaluation Questionnaire

This appendix presents the course evaluation questionnaire administered to participants. The questionnaire includes both quantitative Likert-scale questions, utilizing a slider ranging from zero to ten, and qualitative open-ended questions, which allow for detailed textual responses. Part one of the questionnaire addresses course satisfaction, while parts two and three focus on course helpfulness.

Intro FSS1150- End of term reflection (FALL 2024) Please provide your UofO Email Address Student ID and then copy and past your responses from your reflection worksheet into the appropriate form.



EMAIL

Email address: Please enter your University of Ottawa email address here. We use this to record that you have completed the survey.

Student_ID

Student ID: Please enter your student ID, so that we can ensure that you are awarded your homework points for completing the survey.

Satis_gen **Part 1: General Satisfaction. How much did you enjoy this course? (Please move the slider to indicate how much you enjoyed the course.)**

0 1 2 3 4 5 6 7 8 9 10

How much did you enjoy the course? 0 = Not at all; 10 = Very, very much ()	
How meaningful was this course to you? ()	
How engaging was the material for you? ()	

Satis_1 What were the five things that you liked most about this course (and why did you like them): #1

Satis_2 #2

Satis_3 #3

Satis_4 #4

Satis_5 #5

Helpfulness **Part 2.1: General Helpfulness.** How helpful did you find the course? (Please move the slider to indicate how helpful you found the course.)

0 1 2 3 4 5 6 7 8 9 10

How helpful did you find the course? 0 = Not at all; 10 = Very, very much ()	
--	--



How_helpful **Part 2.2-** I would like to know how this course personally helped you (or didn't help you). This is separate from how much you enjoyed or did not enjoy the course. Please be as detailed as possible. Try to think of at least five things (or more) that you feel were helpful to you.

Not_helpful **Part 2.3:** What other kinds of help would have helped you do better this semester at university?

Skill_1_desc **Part 3.1: Skills** What were the five most important skills that were helpful to you and would like to continue to use? For each skill, I would like to know (a) what the skill was and (b) how it helped you. If it did not help you. Skill #1

Skill_1 Please move the slider to indicate how helpful this skill was and how frequently you used it.



0 1 2 3 4 5 6 7 8 9 10

How helpful was this skill? (0=Not at all; 10 = Extremely helpful) ()	
How often did you use this skill (0=not at all; 10 = Almost all the time). ()	

Skill_2_desc **Skill #2**

Skill_2 Please move the slider to indicate how helpful this skill was and how frequently you used it.



0 1 2 3 4 5 6 7 8 9 10

How helpful was this skill? (0=Not at all; 10 = Extremely helpful) ()	
How often did you use this skill (0=not at all; 10 = Almost all the time). ()	

Skill_3_desc Skill #3

Skill_3 Please move the slider to indicate how helpful this skill was and how frequently you used it.



0 1 2 3 4 5 6 7 8 9 10

How helpful was this skill? (0=Not at all; 10 = Extremely helpful) ()	
How often did you use this skill (0=not at all; 10 = Almost all the time). ()	

Skill_4_desc Skill #4

Skill_4 Please move the slider to indicate how helpful this skill was and how frequently you used it.

0 1 2 3 4 5 6 7 8 9 10

How helpful was this skill? (0=Not at all; 10 = Extremely helpful) ()	
How often did you use this skill (0=not at all; 10 = Almost all the time). ()	

Skill_5_desc Skill #5

Skill_5 Please move the slider to indicate how helpful this skill was and how frequently you used it.

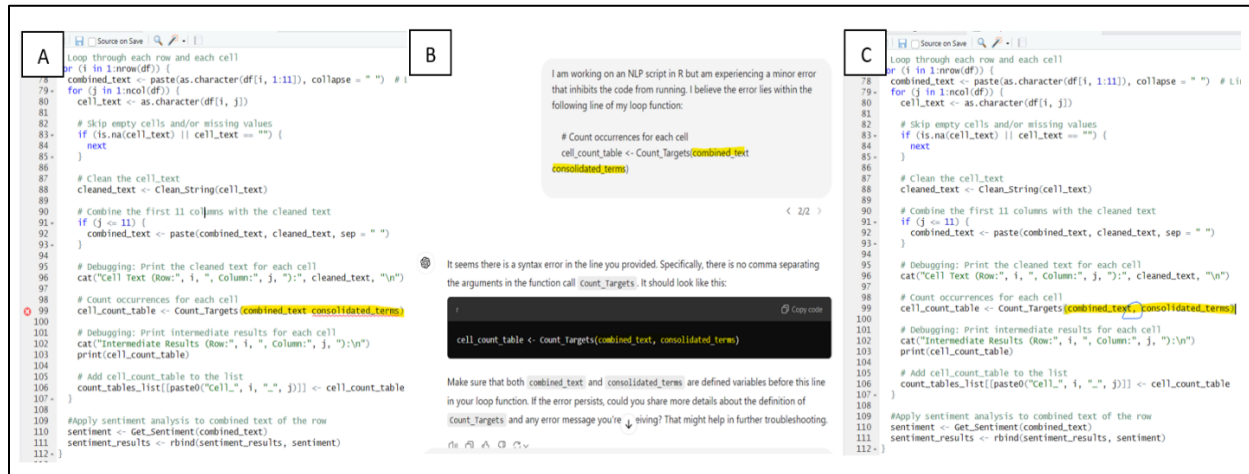
0 1 2 3 4 5 6 7 8 9 10

How helpful was this skill? (0=Not at all; 10 = Extremely helpful) ()	<div></div>
How often did you use this skill (0=not at all; 10 = Almost all the time). ()	<div></div>

Appendix B

Figure 1B

Demonstration of Generative AI and AI-Assisted Technologies in the Coding Process



Note. (A) illustrates a minor syntax error that arose while coding an advanced loop function in R, (B) demonstrates the troubleshooting interaction with ChatGPT that required the researcher's knowledge of R programming, and (C) demonstrates the syntax error in R being resolved.