Education and Information Technologies
https://doi.org/10.1007/s10639-026-13979-1

A STEM dictionary for quantitative text analysis in online education: development and validation

Michael Heilemann[1] · **Heidrun Stoeger**[1]

Received: 8 September 2025 / Accepted: 19 March 2026

Abstract
Effective online STEM education programs depend on participants' engagement in discussions about STEM-related topics. The lack of efficient and validated measurement instruments for systematically analyzing such communication, however, limits the collection of behavioral data in this area. This article presents the development and assessment of the main psychometric properties of a LIWC-based STEM dictionary designed to measure STEM-related communication. A series of 10 studies examined the dictionary using classical test theory criteria, including objectivity (particularly objectivity of scoring and interpretation), reliability (parallel-test, split-half, and retest reliability), and validity (construct and criterion validity, including both concurrent and predictive validity). These studies were conducted within the context of a nationwide online mentoring program for girls, focusing on participants' STEM communication on a secure mentoring platform. Results indicate that the STEM dictionary provides a reliable and valid instrument for measuring STEM-related communication in online educational contexts. We discuss the potential of the dictionary-based approach for evaluating computer-based educational measures in STEM and its use as an efficient method for analyzing communication processes in digital learning environments.

Keywords STEM education · Quantitative text analysis · Digital educational contexts · Dictionary development · Computer-based education · STEM communication · Educational assessment · Mentoring

✉ Michael Heilemann
michael.heilemann@ur.de

[1] Department of Educational Sciences, University of Regensburg, Regensburg, Germany

Published online: 23 April 2026

1 Introduction

STEM (science, technology, engineering, and mathematics) education plays a key role in addressing global challenges, fostering innovation, and ensuring a highly skilled workforce. It prepares the next generation for the demands of an increasingly technology-driven world. STEM education offers children and adolescents a wide range of learning opportunities in both school and out-of-school contexts (Baran et al., 2019; Sun et al., 2023; Young et al., 2016). Even before the COVID-19 pandemic, numerous educational initiatives had already been implemented online (Chiang et al., 2022; Kefalis & Drigas, 2019; Watson, 2023). While online learning offers significant advantages, such as flexibility (Müller & Mildenberger, 2021) and personalized learning experiences (Chaku et al., 2021), it also presents challenges that must be addressed to maintain the quality of these offerings, particularly with regard to understanding and evaluating participants' communication and engagement in online learning environments.

One important factor in the success of online STEM education programs is the ability of program managers and educators to focus participants' communication on STEM topics. For example, research has shown that STEM mentoring is more effective when participants' discussions remain centered on STEM-related topics rather than shifting toward more general or organizational conversations (Stoeger et al., 2016, 2021). Furthermore, studies indicate that online learning is more effective when students actively engage in discussions directly related to the course content or learning objectives (Peng & Xu, 2020), or when discussions focus on core areas of study (Liu et al., 2023a). Understanding the extent to which participants' communication in online learning environments focuses on STEM-related topics is therefore an important task for program evaluation and educational research.

To investigate language and communication patterns, particularly in offline mentoring contexts, surveys, checklists, and interviews are commonly used. In comparison, documenting and analyzing participants' actual communication behavior can be challenging (Alwani et al., 2023; DuBois & Neville, 1997; Finkelstein et al., 2012; Parra et al., 2002). One advantage of online learning environments is that they enable researchers to observe participants' real communication patterns unobtrusively. However, analyzing these communication data systematically can be challenging. Manual analysis of such communication data often requires extensive effort and resources, especially in long-term online learning programs, which can generate large volumes of textual data. Bozkurt et al. (2019) emphasized the potential of using data mining approaches to analyze communication in online learning environments, thereby enhancing the efficiency and effectiveness of STEM education research. However, a systematic review of 258 articles on STEM education research revealed that data mining methods such as text mining or sentiment analysis have rarely been applied in this field (Bozkurt et al., 2019). One reason may be the lack of suitable measurement instruments tailored to the context of STEM education.

To address these challenges, we adopted a dictionary-based approach to efficiently analyze STEM content in large amounts of online communication data in educational contexts. In this form of text analysis, software compares predefined or self-developed word lists – also referred to as lexicons or dictionaries – with the texts to be ana-

lyzed, counting the frequency of specific words. This enables the automatic encoding of large text corpora with minimal effort. Specifically, the aim of this study was to develop and systematically validate a German-language LIWC-based dictionary for measuring STEM-related communication in computer-based educational contexts.

LIWC (Linguistic Inquiry and Word Count) is a widely used software program that provides built-in dictionaries for analyzing text across various domains, including education and psychology. It distinguishes between function words (e.g., pronouns, articles, prepositions) and content words (e.g., nouns, verbs, adjectives) to identify linguistic patterns and gain insights into communication. While LIWC has been extensively applied in the social sciences to explore text characteristics, language use, and behavioral insights (Boyd & Schwartz, 2021; Donohue et al., 2014; Tausczik & Pennebaker, 2010), a dictionary specifically designed to capture STEM-related communication in educational contexts has been lacking.

In the following, we first discuss the methodological potential of dictionary-based approaches for analyzing communication data in online learning environments. Next, we describe the development and validation of the STEM dictionary, an approach that can also be applied to the development and validation of other dictionaries. Finally, we report the results of the empirical studies conducted within the context of a nationwide online mentoring program for girls in STEM in Germany, demonstrating that the STEM dictionary provides a reliable and valid instrument for measuring STEM communication in computer-based educational contexts.

2 Application and usefulness of the dictionary-based approach in computer-based educational contexts

LIWC is a computer-based text analysis tool that systematically counts words in texts and assigns them to predefined linguistic and psychological categories (Boyd et al., 2022). Initially developed in the early 1990s by James W. Pennebaker as part of his research on the health benefits of expressive writing (Pennebaker, 1997), LIWC quickly gained widespread use in psychological research. It has since been applied to investigate the relationship between language use and various psychological and social processes, including social relationships, thinking styles, group processes, and individual differences such as age, gender, personality traits, and mental health (Tausczik & Pennebaker, 2010).

Dictionary-based text analysis has been widely used in psychological and social science research (Tausczik & Pennebaker, 2010). It has been applied to the study of emotional processes (Tov et al., 2013), personality differences (Pennebaker & King, 1999), social interaction patterns (Niederhoffer & Pennebaker, 2002), and online communication (Kramer et al., 2014). Taken together, these studies show that dictionary-based approaches can capture meaningful linguistic indicators of psychological processes and communication patterns in large text corpora. In addition to these applications, several methodological studies have emphasized the importance of systematically developing and validating dictionaries when they are used as measurement instruments for theoretically meaningful constructs (Donohue et al., 2014; Short et al., 2010; Tausczik & Pennebaker, 2010). However, despite these advances,

comparatively little attention has been devoted to the systematic validation of newly developed dictionaries in specific research domains.

The value of LIWC has also been widely recognized in educational research contexts. For instance, it has been applied with its general-purpose dictionaries to analyze textbooks (Lucy et al., 2020; Sell & Farreras, 2017), students' reflective essays (Lin et al., 2016; Savicki & Price, 2021), and school-based consultation (Newman et al., 2015). Additionally, LIWC's main dictionaries, including psychological categories, such as cognitive or emotional word lists, as well as function word categories (e.g., pronouns, conjunctions, and prepositions), have been used to examine language use in application essays and its relation to students' subsequent course and study performance (Pennebaker et al., 2014; Robinson et al., 2013). Other applications include sentiment analysis of educational Twitter data (Borchers et al., 2021), analyses of mentoring partner communication (Scielzo et al., 2011), and assessments of learners' cognitive engagement in online course discussions (Liu et al., 2023b).

However, all of these applications relied on LIWC's general-purpose dictionaries. None of them included a dictionary specifically tailored to STEM-related vocabulary. Building on LIWC's wide application in educational contexts, the development of a STEM-specific dictionary is therefore particularly compelling, as it allows researchers to move beyond general linguistic features and capture the domain-specific content of STEM discussions. Furthermore, research evidence suggests that focusing on program-relevant content is closely related to the effectiveness and success of educational and mentoring programs, making LIWC a promising tool for analyzing such discussions in online learning environments when equipped with a domain-specific dictionary. We know from mentoring research that the discussions of program-relevant content by mentoring pairs are closely related to the effectiveness of the mentoring program (Alwani et al., 2023; Parra et al., 2002; Scielzo et al., 2011). Similarly, studies of online learning environments show that participants achieve greater benefits when they engage in discussions specifically related to course content or learning objectives (Liu et al., 2023a; Peng & Xu, 2020). For instance, in the context of English language learning in higher education, Liu et al. (2023a) analyzed forum discussions on an online learning platform and found that students who concentrated on topics directly related to their learning content, such as major-specific concepts and domain-specific knowledge, achieved higher learning outcomes. In contrast, those who primarily discussed general learning methods or focused on social interactions exhibited comparatively lower academic performance. These findings highlight the importance of guiding participants' discussions toward content that closely aligns with their learning goals and subject matter. Accordingly, online STEM education programs may be more successful when participants engage in STEM-related discussions.

In this study, STEM-related communication refers to the presence of domain-specific STEM terminology in participants' written communication within online educational environments. The dictionary-based measure used in this study therefore captures manifest topical markers of STEM-related discourse rather than deeper aspects of communication. It identifies the occurrence of STEM-related terminology in text but does not assess conceptual understanding, reasoning quality, misconceptions, epistemic stance, or emotional tone.

To investigate this assumption, measurement instruments are needed that enable researchers to analyze large amounts of textual communication data, ideally, a suitable dictionary that can be used to identify the frequency of STEM-related terms and thus determine whether the communication relates to the relevant STEM content.

In recent years, advances in natural language processing (NLP), including embedding-based and transformer-based language models, have enabled increasingly sophisticated analyses of textual data (Incitti et al., 2023; Lin et al., 2022; Qiu et al., 2020). These approaches can capture contextual meaning, implicit references, and complex semantic relationships between words. However, for the research objective of the present study, namely the identification of STEM-related topical content in participants' written communication, a dictionary-based approach offers several important advantages. Dictionary methods provide transparent and interpretable coding procedures, allowing researchers and educational practitioners to clearly understand how textual features are identified and quantified. This transparency facilitates reproducibility and comparability of results, which are particularly important in program evaluation contexts. In addition, dictionary-based approaches require relatively limited computational resources and can be applied without large annotated training datasets, making them suitable for analyzing communication data in smaller or privacy-sensitive educational environments. At the same time, more advanced NLP methods may complement dictionary-based approaches in future research by capturing contextual meaning, implicit references to STEM topics, or deeper aspects of reasoning that go beyond the identification of explicit topical markers in text.

Therefore, our goal was to develop a STEM dictionary for quantitative text analysis with LIWC, tailored explicitly to STEM education, to provide an efficient method for determining the extent to which participants in online STEM learning environments engage in STEM-related discussions. This dictionary could be used in program evaluations and research on STEM education. For example, the dictionary could be combined with existing LIWC emotion dictionaries (Tov et al., 2013) to identify positively and negatively connoted STEM discussion posts and analyze how such emotional tones influence successful participation in computer-based STEM educational interventions.

The development of the STEM dictionary and the subsequent validation studies were conducted in the STEM education context of CyberMentor, a Germany-wide online mentoring program designed to support girls with an interest in STEM subjects. Each year, CyberMentor provides up to 800 female secondary school students with at least one year of one-on-one mentoring by a female STEM professional. Participants also have the opportunity to interact with up to 1,600 other program participants on the CyberMentor online platform, which includes communication channels such as email, chat, and forum. Mentors support their mentees by discussing STEM-related and school-related topics, exploring college opportunities and career prospects in STEM fields, addressing work-life balance, and collaborating on STEM projects. In this context, we developed a STEM word list to measure the proportion of STEM-related communication in an automated manner. This word list aimed to provide a robust tool for supporting the analysis of the program's effectiveness. The following section describes the development of the STEM word list and the steps taken to ensure the content validity of the final STEM dictionary.

3 Development of the STEM dictionary

As the present study aims to identify manifest STEM-related topical markers in participants' written communication, a dictionary-based approach provides a transparent and reproducible method for systematically analyzing large volumes of textual communication data. Short et al. (2010) recommend combining deductive and inductive approaches to dictionary development. In the deductive approach, a comprehensive list of potentially relevant words is compiled independently of the texts to be analyzed, for example by consulting thesauri and synonym dictionaries. In the inductive approach, the texts to be analyzed are systematically examined for relevant words, ensuring that words not identified deductively are nonetheless captured if they are of potential relevance to the dictionary. Domain experts then review the resulting word lists and decide which words to include in the final dictionary. Their agreement is assessed with interrater reliability (Holsti, 1969). In this way, the content validity of the dictionary is ensured during its development.

Sources For the development of the STEM dictionary, three word lists were generated from different sources: (1) dictionary of synonyms (deductive), (2) technical dictionaries (deductive), and (3) forum posts (inductive).

Deductive approach The first word list was created from a dictionary of synonyms. Central STEM terms (e.g., *STEM*, *science*, *technology*, *engineering*, *mathematics*, *informatics*, *biology*, *chemistry*, and *physics*) were identified, and synonyms and hyponyms for these terms (e.g., *astrophysics*, *microbiology*) were extracted using the CanooNet meaning dictionary, which integrates GermaNet, a lexical-semantic network mapping semantic relations between German nouns, verbs, and adjectives (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010; Kunze & Lemnitzer, 2007). The second word list was created from technical dictionaries by consulting Wiktionary glossaries to identify STEM terms in categories such as astronomy, geosciences, computer science, mathematics, physics, biology, and chemistry (e.g., *electrode*, *atomic*, *endoplasmic*).

Inductive approach The third word list was created inductively from forum posts written by CyberMentor mentees and mentors between 2009/10 and 2011/12. Using AntConc (Anthony, 2014), a frequency-based word list was created. Common but irrelevant words (e.g., *the*, *and*, *or*) were excluded using a stopword list, and words occuring fewer than three times were removed, as recommended by Short et al. (2010).

Content validation The three word lists were evaluated by two independent STEM domain experts. Words were included in the dictionary only if both raters agreed they were STEM-related. After joint discussions, additional terms suggested by one expert were either added or excluded. The validated words were then formatted for use in LIWC by truncating them with an asterisk (*) to enable recognition of inflected forms. After removing duplicates across the three word lists, the final STEM dictionary contained 1,926 STEM words.

Interrater reliability Interrater reliability was calculated using Holsti's (1969) formula ($PA_O = 2A/n_A + n_B$), where A is the number of matches between raters, and n_A and n_B are the total words coded by each rater. Following established thresholds (Lombard et al., 2002; Neuendorf, 2002), coefficients of $r \geq .90$ indicate excellent agreement, while $r \geq .80$ indicate good agreement. For the three lists, reliability was .90 (dictionary of synonyms), .82 (technical dictionaries), and .80 (forum posts). For the entire dictionary, reliability was .82 with duplicates included and .80 without duplicates, indicating good interrater reliability overall. A digital version of the dictionary, including the mapping of entries to the three sources – dictionary of synonyms, technical dictionaries, and forum posts – is available at https://doi.org/10.17605/OSF.IO/3JC9X (Heilemann, 2015).

4 Examination of the main psychometric properties of the STEM dictionary

It is widely recognized that the psychometric properties (i.e., objectivity, reliability, and validity) of newly developed dictionaries for text analysis with LIWC must be verified and established (Donohue et al., 2014; Marszałek et al., 2023; McDonnell et al., 2020; Short et al., 2010; Tausczik & Pennebaker, 2010). In dictionary-based text analysis, dictionaries function as measurement instruments that operationalize theoretical constructs through predefined lexical categories. Similar to psychometric scales, such instruments require systematic evaluation to ensure that they reliably and validly capture the intended construct. Methodological work on dictionary-based analyses has therefore emphasized the importance of evaluating dictionaries with regard to reliability, construct validity, and criterion validity (Donohue et al., 2014; Short et al., 2010; Tausczik & Pennebaker, 2010). Although many applied studies use existing dictionaries without extensive validation, more recent methodological contributions have argued that newly developed dictionaries should be validated systematically, particularly when they are used as indicators of theoretically meaningful constructs. However, there is no unified procedure for ensuring the quality of dictionary-based text analysis, and scholars have emphasized the need for systematic validation strategies, particularly as LIWC is increasingly applied to online communication and other text-based phenomena (Boyd & Schwartz, 2021; Donohue et al., 2014; McDonnell et al., 2020).

Our goal is twofold: to establish the STEM dictionary as a widely accepted instrument for measuring STEM communication in computer-based educational settings with established psychometric properties, and to provide an adaptable validation framework for newly developed dictionaries. To achieve this, we propose a systematic strategy to test the main quality criteria of classical test theory: objectivity (particularly objectivity of scoring and interpretation), reliability (particularly parallel-test reliability, split-half reliability, and retest reliability), and validity (construct validity and criterion validity, including both concurrent and predictive validity).

To systematically examine these quality criteria, the validation of the STEM dictionary was conducted through a series of complementary empirical studies, each addressing a specific psychometric criterion of the dictionary.

Table 1 Overview of the validation studies and their respective validity objectives

Study	Validity aspect	Data source/communication channel	Key metric
Study 1	Objectivity of scoring	Forum and chat posts	Consistency of automated coding
Study 2	Objectivity of interpretation	Forum, chat, and email posts (norm tables)	Distributional comparison
Study 3	Reliability (parallel-test)	Forum, chat, and email posts	Correlation between dictionary components
Study 4	Reliability (split-half)	Forum, chat, and email posts	Split-half coefficient
Study 5	Reliability (retest)	Forum, chat, and email posts across time	Test–retest correlation
Study 6	Construct validity (manual coding comparison)	Forum posts and manual coding	Correlation with human ratings
Study 7	Construct validity (self-report comparison)	Forum, chat, and email posts and self-report survey	Correlation with self-reported STEM communication
Study 8	Construct validity (known-groups comparison)	Forum posts from STEM and non-STEM discussion areas	Group comparison
Study 9	Criterion validity (concurrent)	Forum, chat, and email posts and self-reported mentoring outcomes	Regression analysis
Study 10	Criterion validity (predictive)	Forum, chat, and email posts and long-term program participation	Regression analysis

The studies follow the main quality criteria of classical test theory: objectivity, reliability, and validity.

The following section reports the studies conducted in the STEM education setting of CyberMentor, which were designed to systematically establish the psychometric properties of the STEM dictionary. Table 1 provides an overview of the 10 studies, linking each study to its primary validity objective, the corresponding data source or communication channel, and the key metric used to evaluate the respective psychometric property. Because the STEM dictionary is intended to function as a measurement instrument, its evaluation required examining multiple aspects of measurement quality. The validation strategy therefore followed a multi-study design in which each study addressed a specific psychometric criterion derived from classical test theory. Rather than representing independent empirical projects, the studies should be understood as complementary validation steps within an overall validation framework that jointly establish the reliability and validity of the dictionary.

4.1 Objectivity

Objectivity refers to the requirement that the outcome of a measurement is independent of the person conducting it and not influenced by the examiner. Three forms of

objectivity can be distinguished, corresponding to different phases of the assessment process: procedural objectivity, objectivity of scoring, and objectivity of interpretation (Ingenkamp & Lissmann, 2008; Lienert & Raatz, 1998).

4.1.1 Procedural objectivity

Procedural objectivity refers to the independence of measurement results from variations in examiner behavior during test administration (Lienert & Raatz, 1998). In dictionary-based text analyses, procedural objectivity is inherently perfect, as all texts are processed identically through fully automated coding. For this reason, procedural objectivity was not examined in a separate empirical study.

4.1.2 Scoring objectivity

Scoring objectivity refers to the independence of the scoring of test results from the examiner, that is, the numerical or categorical assignment of responses or observed behavior according to predefined rules (Lienert & Raatz, 1998). Although dictionary-based text analyses rely on automated coding, scoring objectivity can still be affected by factors such as the version of the dictionary, the software implementation, or the preprocessing of texts (e.g., correction of spelling errors). To examine the scoring objectivity of the STEM dictionary, we conducted a study examining how spelling errors in texts influence the results (Study 1).

4.1.3 Objectivity of interpretation

Objectivity of interpretation refers to the independence of test score interpretation from the person conducting the interpretation. When applied to dictionary-based text analyses, a challenge arises in classifying test values: For instance, does a text with 1.3% positive emotion words reflect a high versus a low level of emotional expression? The interpretation of such values is challenging because people generally lack intuitive reference points for such values, and suitable norms are often unavailable. To improve the objectivity of interpretation, we established norms for students' STEM communication on the CyberMentor platform, allowing individual students' STEM communication to be compared with that of their peers (Study 2).

4.2 Reliability

Reliability refers to the degree of precision with which a particular construct is measured, irrespective of whether the instrument actually measures the construct it is intented to capture (Lienert & Raatz, 1998). Several approaches have been established to estimate the precision of measurement instruments, with the most common being parallel-test reliability, split-half reliability, and retest reliability (Ingenkamp & Lissmann, 2008; Lienert & Raatz, 1998).

4.2.1 Parallel-test reliability

Parallel-test reliability is determined by having the same sample complete two tests that are as similar in content as possible; a high correlation between the results of both tests is expected (Lienert & Raatz, 1998). Applied to dictionary-based text analyses, this involves applying two highly similar dictionaries to the same texts and correlating their results. Because developing an additional dictionary is typically resource-intensive, it is usually not feasible to create two separate dictionaries for this purpose. However, parallel-test reliability can be estimated using dictionary components that were developed independently from different sources (e.g., dictionary of synonyms, technical dictionaries, forum posts; see Sect. 3 for details on the development of the STEM dictionary) and later combined into a final dictionary. The STEM dictionary was assessed using its components to establish parallel-test reliability (Study 3).

4.2.2 Split-half reliability

Split-half reliability, as a measure of the internal consistency of a test, is determined by dividing the items of a test administered to a sample into two equivalent halves; a high correlation between the two halves is expected (Lienert & Raatz, 1998; Ingenkamp & Lissmann, 2008). Applied to dictionary-based text analyses, this involves splitting the words of a dictionary into two equivalent halves, applying both halves to the same text material, and correlating the results. The internal consistency of the STEM dictionary was examined by dividing it into two halves and calculating split-half coefficients across different communication channels on the CyberMentor platform (Study 4).

4.2.3 Retest reliability

Retest reliability is determined by administering the same test to the same sample at a later point in time, with the interval between the two measurements typically spanning several weeks (Bortz & Döring, 2006; Lienert & Raatz, 1998). A high correlation between the results of both test administrations is expected. Applied to dictionary text analyses, this involves collecting two sets of text samples from the same participants at different points in time under comparable conditions, analyzing them with the dictionary, and correlating the results. The retest reliability of the STEM dictionary was examined using text samples from CyberMentor participants written several weeks apart (Study 5).

4.3 Validity

Validity is considered the most important quality criterion of a measurement instrument (Ingenkamp & Lissmann, 2008) and refers to the extent to which an instrument actually measures what it is intented to measure. Three types of validity are typically distinguished: content validity, construct validity, and criterion validity (Lienert & Raatz, 1998).

4.3.1 Content validity

Content validity is established when the content of a test adequately represents the construct to be measured in all its essential aspects (Bortz & Döring, 2006). This is typically assessed through expert judgments, with a high degree of agreement among experts being expected (Lienert & Raatz, 1998). Applied to dictionary-based text analyses, content validity implies that, ideally during the development of a new dictionary, multiple experts evaluate and decide which words should be included in the final version. The content validity of the STEM dictionary was ensured through expert ratings during its development (see Sect. 3).

4.3.2 Construct validity

Construct validity refers to the extent to which test results can be interpreted as measuring the theoretically defined construct of interest (Bühner, 2011; Lienert & Raatz, 1998). Whether a test captures a construct is determined on the basis of theoretical considerations and subsequent empirical investigations (Lienert & Raatz, 1998). In dictionary-based text analyses, this involves empirically examining whether words included in a dictionary actually represent the intended construct. Although content validity established through expert ratings during dictionary development provides an initial indication, additional empirical evidence is required to ensure construct validity. Accordingly, the construct validity of the STEM dictionary was examined in three ways: by comparing its automated coding results with those obtained through manual coding (Study 6), by relating its scores to students' self-reports of STEM communication (Study 7), and by comparing forum posts assumed to contain average versus extreme levels of STEM-related content (Study 8).

4.3.3 Criterion validity

Criterion validity is established when test results allow conclusions about a practically relevant criterion outside the testing situation (Hartig et al., 2012). Such criteria may include outcomes (e.g., grades), behaviors (e.g., teachers' feedback), or traits (e.g., motivation; Blömeke, 2013). When a criterion is measured concurrently with the test, this is referred to as concurrent validity; when it is assessed at a later point in time, this is referred to as predictive validity (Ingenkamp & Lissmann, 2008). In dictionary-based text analyses, criterion validity is examined by correlating dictionary scores with independently measured criteria. The criterion validity of the STEM dictionary was assessed in two studies: its concurrent validity was tested by relating STEM dictionary scores to self-reported program outcomes collected alongside program participation (Study 9), and its predictive validity by relating STEM dictionary scores to long-term program participation (Study 10).

5 Study 1: Objectivity of scoring of the STEM dictionary

Our first study aimed to evaluate the scoring objectivity of the STEM dictionary. In dictionary-based text analysis, scoring objectivity depends on how the text material is prepared. For instance, typing or spelling errors may cause certain words to be missed during the automatic coding process. This study examined how text corrections, such as correcting errors and standardizing language, affect the results of the STEM dictionary's text analysis for two communication channels: forums and chats.

Previous research by Wolf et al. (2008) showed that correcting text material had minimal impact on the results of LIWC analyses for email communication. However, their findings cannot be directly extrapolated to other forms of computer-mediated communication. In this study, we assumed that while the correction process would lead to the detection of more STEM-related words, the overall scores of uncorrected texts would be highly comparable to those of corrected texts in both forum and chat contributions.

5.1 Method

Participants and text material This study was conducted in the context of the CyberMentor program during the 2009–2010 mentoring year. A total of 789 students (M=14.56 years, SD=2.29) participated in the program for the first time. Of these, 295 students were actively engaged in the forum, and 231 students were actively participating in the chat rooms. For the analyses, we focused on 69 students (M=14.12 years, SD=2.16) who had written at least 100 words in both the forum and the chat. In total, these students produced 3,403 forum posts and 21,525 chat posts. We randomly sampled up to 50 forum posts (M=25.30, SD=18.86) and up to 300 chat posts (M=148.13, SD=109.15) per student. The final sample included 1,746 forum posts and 10,221 chat posts. On average, a forum post consisted of 47 words (M=46.86, SD=59.36), while a chat post averaged five words (M=5.35, SD=6.51).

Data preparation The text samples were corrected by the first author based on predefined criteria (Pennebaker et al., 2007a; Wolf et al., 2008): correcting typographical, spelling, and punctuation errors, standardizing abbreviations and acronyms, adapting colloquial language (e.g., youth slang), removing hyperlinks and graphical artifacts, and editing unique features of Internet language (e.g., elongated words such as *heeelllooo*).

Data analysis Both corrected and uncorrected texts were analyzed with the STEM dictionary. Statistical analysis included difference scores, mean differences, Spearman's rank correlations, and intraclass correlations (ICCs) to compare corrected and uncorrected text scores. The nonparametric Wilcoxon signed-rank test was used to analyze mean differences, while intraclass correlations (ICCs) were calculated to assess agreement in scoring. Following Shrout & Fleiss (1979), an ICC (3, 2) model was applied (a two-way mixed model with absolute agreement as the definition of agreement).

5.2 Results and discussion

Supplementary Table A.1 provides descriptive statistics and additional relevant metrics for the corrected and uncorrected forum and chat posts. The results showed that, with the exception of the proportion of STEM words in forum posts, significantly more STEM words were detected in corrected texts compared to uncorrected texts: in the proportion of STEM words in chat posts ($Z = -3.04$, $p < .01$), in the total number of STEM words in forum posts ($Z = -3.58$, $p < .01$), and in the total number of STEM words in chat posts ($Z = -3.03$, $p < .01$). For forum posts, 747 STEM words were identified in corrected texts compared to 691 STEM words in uncorrected texts, corresponding to 92.5% agreement. For chat posts, 363 STEM words were identified in corrected texts compared to 332 in uncorrected texts, corresponding to 91.5% agreement. The correlations between corrected and uncorrected text scores were very high for both forum ($r = .98$) and chat ($r = .99$). The intraclass correlations (ICCs), which assess absolute agreement, were also high, ranging from .92 to 1.00.

The analyses further revealed that forum posts were more suitable for analysis with the STEM dictionary than chat posts, primarily because fewer words required correction. In forum posts, 26% of words required correction, whereas in chat posts, this proportion was nearly double at 46.7%. STEM words in chat posts were also more likely to require correction (28.7%) compared to STEM words in forum posts (17.5%).

Taken together, these findings highlight that although text correction significantly increases the number of detected STEM words, it does not substantially affect the overall agreement or stability of the results. The very high ICCs suggest that even uncorrected texts provide results that are consistent and comparable for dictionary-based text analysis. This indicates that in large-scale studies, where manual text correction is often impractical due to time and resource constraints, uncorrected text data can still yield accurate insights into participants' STEM communication. Overall, these findings underscore the robustness of the STEM dictionary in capturing STEM communication across different types of computer-mediated communication and provide evidence for its high scoring objectivity.

6 Study 2: Objectivity of interpretation of the STEM dictionary

The interpretation of results from dictionary-based text analysis can be challenging when there is no representative comparison group for contextualizing the findings. This study aimed to enhance the objectivity of interpretation for the STEM dictionary by developing comparison measures or norms. Specifically, we established norms for students' STEM communication on the CyberMentor platform, allowing the comparison of an individual student's STEM communication with that of her peers. These norms enable the determination of whether a student's STEM communication is below, average, or above average, regardless of the interpreter. Given the different characteristics of the communication channels on the CyberMentor platform (forum, chat, and email), we assumed that the percentage of STEM words would

vary between them. Accordingly, separate norms were developed for each channel, as outlined below.

6.1 Method

Participants and text material To develop the norms, we analyzed 11,918 forum posts, 242,493 chat posts, and 40,833 emails from students (N=2,158; M=14.36 years, SD=2.19) who participated in the CyberMentor program during the three mentoring years between 2009 and 10 and 2011–12. Only text contributions from the students' first mentoring year were included in the analysis. Additionally, students were only considered if they had written at least 100 words in the respective communication channel.

Measures and data analysis The STEM dictionary was used to calculate the percentage of STEM words in the students' text contributions. To account for the variability in text length across students, the raw test scores were based on the percentage of STEM words rather than the absolute number of STEM words. The norming procedure followed Bühner's (2011) guidelines for norm-oriented test evaluation. Raw test scores were first assessed to determine whether the data followed a normal distribution, using histograms and statistical tests, which revealed a right-skewed distribution of STEM word proportions in all three communication channels.

For non-normally distributed data, percentile ranks and stanine scores are recommended (Bühner, 2011). Percentile ranks describe the percentage of students with equal or lower scores (e.g., a percentile rank of 60 indicates that 60% of students have the same or fewer STEM words). Stanine scores group percentile ranks into a nine-point scale (M=5, SD=2), providing a normalized metric that is useful for calculating confidence intervals. Percentile ranks and stanine scores were calculated separately for forum, chat, and email data and presented as norm tables (see Supplementary Table A.2).

6.2 Results and discussion

Forum norms For the forum data, we analyzed 11,391 forum posts from 533 students. On average, students wrote 8.21 STEM words (SD=16.87) in their forum posts during the mentoring year. Notably, 24.8% of the students did not use any STEM words in their forum contributions. Supplementary Table A.2 shows that 60% of students had a STEM word proportion of 0.8% or less. Students who wrote 2.5% or more STEM words were classified as above average in their forum communication.

Chat norms For the chat data, 239,605 posts from 613 students were analyzed. On average, students wrote 11.21 STEM words (SD=46.91) in chat posts during the mentoring year. No STEM words were found in the chat posts of 21.4% of students. Supplementary Table A.2 indicates that 60% of students had a STEM word proportion of 0.6% or less. Students with 1.6% or more STEM words were categorized as above average in chat communication.

Email norms For the email data, we analyzed 40,438 emails from 1,531 students. On average, students wrote 24.36 STEM words ($SD = 48.58$) in their email contributions during the mentoring year. Only 7.2% of students did not use STEM words in their email communication. Supplementary Table A.2 shows that 60% of students had a STEM word proportion of 1.2% or less, while those with a proportion of 2.4% or more were classified as above average.

Developing norm tables enhances the objectivity of interpretation for the STEM dictionary, enabling more precise classification of individual students' proportion of STEM communication, especially in online mentoring programs for girls. Depending on the STEM program and the research focus, additional norms can be developed for specific subgroups, such as age groups or grade levels. Establishing norms for dictionaries is particularly valuable in educational settings, as it improves the objectivity of interpretation and facilitates more meaningful comparisons.

7 Study 3: Parallel-test reliability of the STEM dictionary

The goal of this study was to evaluate the parallel-test reliability of the STEM dictionary. In dictionary-based text analysis, developing multiple dictionaries to capture the same construct is uncommon due to the substantial effort required. As a result, parallel-test reliability is rarely assessed. However, the STEM dictionary, composed of three partially overlapping components, provided a unique opportunity to evaluate parallel-test reliability by comparing scores across these components. The study examined correlations between the two deductively created dictionary components and the inductively created component. Additionally, the components' scores were compared with the overall STEM dictionary to validate their reliability further.

7.1 Method

Participants and text material The study analyzed text contributions from students participating in the CyberMentor program during the 2009–10 mentoring year. Of the 665 students ($M = 14.50$ years, $SD = 2.29$) who wrote on the platform, 295 students were active in the forum, 231 students used the chat, and 641 students used the email function. Across all channels, the students produced 4,961 forum posts, 35,459 chat posts, and 12,603 emails.

Measures and data analysis The STEM dictionary and its components were used to analyze the proportion of STEM words in students' text contributions. The deductive components (DC 1 and DC 2) focused on key STEM terms and their synonyms, as well as technical terms sourced from specialized dictionaries. The inductive component (DC 3) captured frequent STEM words identified from prior analyses of mentor and mentee forum posts. The percentage of STEM words for each component was calculated, and Spearman correlation coefficients were used to assess parallel-test reliability between the dictionary components across all communication channels.

7.2 Results and discussion

In forum posts, the mean proportion of STEM words was 1.31% (*SD*=2.20) with DC 3, 1.10% (*SD*=2.03) with DC 1, and 1.06% (*SD*=1.93) with DC 2. In chat posts, the mean proportion of STEM words was 0.31% (*SD*=0.81) with DC 3, 0.30% (*SD*=0.81) with DC 2, and 0.26% (*SD*=0.74) with DC 1. In email posts, the mean proportion of STEM words was 1.13% (*SD*=1.15) with DC 3, 1.01% (*SD*=1.08) with DC 1, and 0.89% (*SD*=1.00) with DC 2.

Supplementary Table A.3 presents the parallel-test reliability coefficients, which revealed consistently high correlations between the dictionary components across all communication channels. For example, in the combined dataset from forum, chat, and email, correlations between dictionary components ranged from .95 to .99, demonstrating excellent parallel-test reliability. According to established thresholds, reliability coefficients above .70 are considered acceptable, above .80 as good, and above .90 as very good (Bortz & Döring, 2006; Rammstedt, 2004; Schermelleh-Engel & Werner, 2012).

The high correlations indicate that the different components of the STEM dictionary measure STEM communication with comparable accuracy, regardless of the communication channel. This study also highlights the advantages of combining multiple dictionary components. While a single dictionary may be sufficient for some applications, using multiple components enhances the detection of less frequent STEM-related terms, improving the overall measurement accuracy. This is particularly important for capturing specialized or niche STEM topics in participants' communication. Overall, the results demonstrate that assessing the parallel-test reliability of dictionary components is a feasible approach to ensuring the parallel-test reliability of the final dictionary.

8 Study 4: Split-half reliability of the STEM dictionary

The purpose of this study was to evaluate the split-half reliability of the STEM dictionary. In dictionary-based text analyses, text material is typically split in half for reliability testing (Slatcher et al., 2007; Yarkoni, 2010). However, since dictionaries themselves serve as the measurement instruments in these analyses, halving the dictionary provides a more relevant and meaningful approach to assessing their internal consistency (Lienert & Raatz, 1998). To do so, the words in the STEM dictionary were alphabetically sorted and alternately assigned to two halves, each of which was used to assess text contributions. Split-half coefficients and Cronbach's alphas were calculated to assess the internal consistency of the dictionary across different communication channels on the CyberMentor platform.

8.1 Method

Participants and text material The study was conducted during the 2010–11 mentoring year and included 816 students (*M*=14.14 years, *SD*=2.07) who contributed text posts on the CyberMentor platform. In total, these students produced 144,003 text

posts across the various communication channels, including the forum, chat, and email. Of these students, 390 were active in the forum (6,351 posts), 491 students wrote in chat (118,812 chat posts), and 760 students used the email function (18,840 posts).

Measures and data analysis The STEM dictionary, consisting of 1,926 words, was divided into two halves (DH 1 and DH 2) by alternating the assignment of alphabetically sorted words. Each half contained 963 words. Students' text posts were analyzed using the two dictionary halves to calculate three variables: the relative frequency of STEM words as a percentage of total words, the absolute frequency of STEM words, and whether a STEM word was used at least once (binary coding).

Split-half coefficients were calculated using Guttman's formula, which provides a more conservative estimate of reliability when standard deviations differ between dictionary halves (Bühner, 2011). Cronbach's alphas were also computed as an additional measure of internal consistency. Analyses were conducted separately for each communication channel.

8.2 Results and discussion

In forum posts, the mean proportion of STEM words was 0.47% ($SD = 0.96$) with DH 2 and 0.31% ($SD = 0.82$) with DH 1. The average number of STEM words was 3.96 ($SD = 10.22$) with DH 2 and 2.46 ($SD = 7.72$) with DH 1.

In chat posts, the mean proportion of STEM words was 0.34% ($SD = 0.75$) with DH 2 and 0.15% ($SD = 0.39$) with DH 1. The average number of STEM words was 4.73 ($SD = 14.06$) with DH 2 and 2.06 ($SD = 5.57$) with DH 1.

In email posts, the mean proportion of STEM words was 0.68% ($SD = 0.75$) with DH 2 and 0.42% ($SD = 0.57$) with DH 1. The average number of STEM words was 13.29 ($SD = 23.31$) with DH 2 and 7.62 ($SD = 14.16$) with DH 1.

Supplementary Table A.4 presents the split-half reliability coefficients and Cronbach's alphas for forum, chat, and email. Split-half coefficients based on relative word frequencies were unsatisfactory ($\leq .70$), whereas those based on binary coding were consistently excellent ($r = .91$–$.94$). However, binary coding likely overestimates reliability (Pennebaker et al., 2007b). Using absolute word frequencies provided a balanced and robust measure of reliability, with coefficients ranging from .75 to .93.

The overall split-half coefficient across all communication channels was .82, indicating good reliability. However, differences emerged among the various communication channels. Forum posts exhibited the highest split-half reliability ($r = .93$), followed by email posts ($r = .83$). In contrast, chat posts showed the lowest reliability ($r = .75$). The lower reliability for chat posts likely reflects their brevity and informal nature, which may limit the consistent use of technical STEM terms compared to the more structured and asynchronous communication in forums and email.

Overall, these results demonstrate the satisfactory split-half reliability of the STEM dictionary and the usefulness of dividing the dictionary into two halves when assessing internal consistency in dictionary-based text analyses. The discrepancy between reliability estimates based on proportions, absolute counts, and binary cod-

ing primarily reflects the distributional characteristics of natural language and the respective calculation methods rather than differences in the discriminatory power of individual dictionary entries. Reliability estimates based on relative word frequencies tend to underestimate internal consistency because base rates of word usage vary considerably across texts, whereas binary coding tends to overestimate reliability because it only captures whether a word appears at least once in a text (Pennebaker et al., 2007b). Calculating reliability based on absolute word frequencies therefore provides a more balanced estimate. Accordingly, future applications of the STEM dictionary should primarily rely on absolute word frequencies when assessing split-half reliability, while proportion-based measures should be interpreted with caution, particularly in datasets consisting of short textual contributions.

9 Study 5: Retest reliability of the STEM dictionary

The purpose of this study was to evaluate the retest reliability of the STEM dictionary. Previous research has shown that good retest reliability can be expected under experimental conditions, whereas lower reliability is often observed in more ecological settings and everyday language use (Mehl & Pennebaker, 2003; Schnurr et al., 1986). Thus, it is likely to find lower retest reliability in educational settings as well. In educational settings, the retest reliability of domain-specific communication is influenced not only by the time interval between measurements but also by many other aspects, such as the content discussed at a specific time.

To examine the retest reliability of the STEM dictionary, we analyzed forum, chat, and email posts written by students on the CyberMentor platform with a two-week interval between contributions. Since the study was conducted in a non-experimental and very open educational setting where students could discuss both STEM and non-STEM topics, we anticipated retest coefficients to be low to satisfactory.

9.1 Method

Participants and text material The study was conducted during the 2011–12 mentoring year and included 715 students ($M = 14.06$ years, $SD = 1.96$) who contributed text posts on the CyberMentor platform. Text samples were collected from posts written during the second and third weeks of September, October, and November 2011. Across these periods, 195–295 mentees sent 2,028 email messages, 52–77 mentees wrote 7,193 chat posts, and 52–69 mentees contributed 410 forum posts.

Measures and data analysis The STEM dictionary was applied to determine the proportion of STEM words and the absolute frequency of STEM words in the text contributions. Retest reliability coefficients were calculated between the measurement points: the mentees' text contributions from September were compared with their contributions from October, and the contributions from October were compared with those from November, using intraclass correlations (ICC). Following Shrout & Fleiss (1979), an ICC (3, 2) model was used (a two-way mixed model with consistency as

the definition of agreement). Intraclass correlations of .70 or higher were interpreted as good (Wirtz & Caspar, 2002).

9.2 Results and discussion

In forum posts, the mean proportion of STEM words was 0.50% (*SD* = 1.24) in September, 1.14% (*SD* = 2.34) in October, and 0.71% (*SD* = 1.48) in November. For the absolute number of STEM words, forum posts averaged 0.62 (*SD* = 1.52) in September, 1.09 (*SD* = 2.16) in October, and 1.26 (*SD* = 3.70) in November.

In chat posts, the mean proportion of STEM words was 0.60% (*SD* = 1.77) in September, 0.36% (*SD* = 1.42) in October, and 0.52% (*SD* = 1.06) in November. For the absolute number of STEM words, chat posts averaged 1.88 (*SD* = 5.75) in September, 1.12 (*SD* = 4.50) in October, and 2.52 (*SD* = 8.02) in November.

In email posts, the mean proportion of STEM words was 0.80% (*SD* = 1.26) in September, 1.00% (*SD* = 1.40) in October, and 0.72% (*SD* = 1.27) in November. For the absolute number of STEM words, email posts averaged 2.28 (*SD* = 4.62) in September, 2.79 (*SD* = 5.47) in October, and 1.77 (*SD* = 2.95) in November.

Supplementary Table A.5 presents the retest reliability coefficients. For the proportion of STEM words, coefficients ranged from .24 to .64, indicating low reliability. For the absolute frequency of STEM words, coefficients ranged from .51 to .87, suggesting low to good reliability.

The lower retest reliability observed for the proportion of STEM words is likely due to the open nature of the CyberMentor platform, which allows students to discuss both STEM and non-STEM topics. As a result, the proportion of STEM-related terms in participants' communication may vary depending on the specific discussion topics addressed at a given time. The higher reliability for the absolute frequency of STEM words indicates greater stability in the number of STEM-related terms used over time, even when the relative proportion of STEM words varied.

These findings suggest that the dictionary-based measure should not necessarily be interpreted as capturing a stable individual-level trait in the sense of a fixed personal propensity to communicate about STEM topics. Rather, the measure reflects the extent to which participants' communication focuses on STEM-related content within a given interaction context. In this sense, the retest reliability of domain-specific communication primarily indicates how consistently participants discuss certain topics over time. Accordingly, the STEM dictionary may be particularly suitable for capturing context-dependent variations in STEM-related communication (i.e., state-like aspects of communication) within specific interactions or communication channels. At the same time, measures aggregated across larger sets of text contributions or longer time periods may provide meaningful indicators of overall engagement with STEM-related content at the individual, mentoring-pair, or program level.

10 Study 6: Construct validity of the STEM dictionary based on comparison with manual coding

While the content validity of the STEM dictionary had already been addressed during its development through expert judgments, additional evidence of construct validity is required to support its use as a measurement instrument. The purpose of this study was therefore to evaluate the construct validity of the STEM dictionary by comparing its automated coding results with those obtained through manual coding conducted by human raters. Forum posts written by students participating in CyberMentor were analyzed using both methods. High correlations between the automated scores and manual coding were expected to indicate good construct validity.

10.1 Method

Participants and text material During the 2011–12 mentoring year, 345 students participated in the forum. A random sample of forum posts was drawn, with a maximum of three posts per student included. The final sample consisted of 335 forum posts written by 125 students ($M = 14.14$ years, $SD = 1.94$).

Measures and data analysis The STEM dictionary was used to analyze the forum posts automatically. Automatic coding calculated the proportion of STEM words and the absolute number of STEM words in the posts.

For manual coding, two STEM experts evaluated the forum posts using a 5-point Likert scale to answer the question: "How much is this forum post about STEM?" Ratings ranged from (1) *not at all* to (5) *completely*. The coders assessed the overall topical focus of each post on STEM-related content without relying on the specific word list of the STEM dictionary. Interrater reliability was assessed using intraclass correlation coefficients (ICCs). Following Shrout & Fleiss (1979), an ICC (2, 2) model was used (a two-way mixed model with absolute agreement as the definition of agreement). The interrater reliability was excellent, with a value of .91.

To assess construct validity, Spearman correlation coefficients were calculated between the STEM dictionary scores and the mean manual coding scores.

10.2 Results and discussion

Automatic analysis using the STEM dictionary showed that mentees' forum posts contained, on average, 1.61% STEM words ($SD = 2.15$), with an absolute number of 2.76 STEM words ($SD = 3.76$). Manual coding revealed an average STEM content rating of 2.00 ($SD = 1.06$) for coder one and 1.91 ($SD = 1.02$) for coder two. The combined mean score across both coders was 1.95 ($SD = 1.02$).

The proportion of STEM words identified by the STEM dictionary was highly correlated with the manual coding results ($r = .80$). Similarly, the absolute number of STEM words identified by the STEM dictionary correlated strongly with the manual coding results ($r = .81$). According to Cohen's (1988, 1992) guidelines, these cor-

relations are considered high and provide strong evidence of the STEM dictionary's construct validity.

The high interrater reliability (ICC = .91) between the two human coders indicates that domain-specific communication content can be captured with excellent agreement through manual coding. However, manual coding required the use of a random sample due to the extensive effort needed to analyze all forum posts within a mentoring year. This approach nonetheless provided an economical and effective method for empirically evaluating the construct validity of the STEM dictionary.

The findings of this study show good construct validity of the STEM dictionary and provide empirical evidence complementing the content validity established during dictionary development. Because the manual coding was based on an independent qualitative assessment of the overall topical focus of each post rather than on the specific dictionary entries, the high correlations between manual and automated coding indicate that the STEM dictionary captures the intended construct of STEM-related communication. These results confirm that the STEM dictionary serves as a valid instrument for analyzing domain-specific communication in computer-based educational settings.

11 Study 7: Construct validity of the STEM dictionary based on comparison with self-reports

The purpose of this study was to assess the construct validity of the STEM dictionary by correlating its automated scores with participants' self-reports of STEM communication on the CyberMentor platform. Students' forum, chat, and email posts were analyzed using the STEM dictionary and correlated with self-reports collected immediately after platform visits. It was hypothesized that the actual STEM communication captured by the dictionary would positively correlate with the students' self-reported STEM communication.

11.1 Method

Participants and text material During a three-week period (November 17, 2011, to December 8, 2011) in the 2011–12 mentoring year, 256 students ($M = 13.77$ years, $SD = 1.84$) participated in the study. Students were prompted to self-report their STEM communication on the platform immediately after logging out. On average, students visited the platform 3.07 times ($SD = 4.46$) during the study period, with an average visit lasting 18.25 min ($SD = 31.86$). In total, the students produced 107 forum posts, 2,751 chat posts, and 347 emails across all visits.

Measures and data analysis The STEM dictionary was applied to analyze the proportion and absolute frequency of STEM words in the text contributions. Additionally, students completed a brief logout questionnaire asking whether they had communicated about STEM during their platform visit in the forum, chat, or via email. Responses (yes = 1, no = 0) were aggregated for each student across the three weeks to create sum variables for each communication channel. Spearman correlation coef-

ficients were calculated between the automatically coded STEM communication and the self-reports for each communication channel.

11.2 Results and discussion

The mean proportion of STEM words was 0.67% (SD = 0.98) in emails, 0.58% (SD = 1.31) in chat posts, and 0.30% (SD = 0.71) in forum posts. The average number of STEM words was 3.08 (SD = 10.35) in chat posts, 2.47 (SD = 5.31) in emails, and 0.70 (SD = 1.83) in forum posts. Self-reports indicated STEM communication scores of 0.49 (SD = 0.82) for email, 0.41 (SD = 1.56) for the forum, and 0.14 (SD = 0.47) for chat.

Supplementary Table A.6 presents the correlations between self-reported and dictionary-based STEM communication. Correlations ranged from small to large (r = .27–.51; Cohen, 1988, 1992), with the highest correlations observed for chat (r = .50–.51), followed by medium correlations for email (r = .39–.43), and the lowest for forum posts (r = .27–.28). The weaker correlations for forum posts may reflect differences in how students interpreted "communication about STEM" in self-reports. Students likely included not only writing but also reading forum posts as STEM communication in their self-reports.

The findings also indicate that students reported communicating about STEM content more frequently than the dictionary-based analysis detected. This discrepancy may reflect social desirability biases in the self-reports and underscores the need to carefully consider how participants perceive and define "communication" when formulating self-report items, strengthening the value of dictionary-based approaches even further. Despite these differences, self-reports remain a valuable tool for validating the construct validity of dictionary-based text analyses, especially for capturing domain-specific communication content in educational online settings.

12 Study 8: Construct validity of the STEM dictionary based on comparison of texts with average and extreme feature salience

This study aimed to test the construct validity of the STEM dictionary by comparing texts with varying feature salience levels, specifically those with average and extreme feature salience. Feature salience refers to the degree to which a particular characteristic is present in a text; a valid dictionary should be able to distinguish between texts with low versus high feature salience (Donohue et al., 2014; Kahn et al., 2007; Rude et al., 2004). Specifically, forum posts from the STEM-Talk subforum, designated for STEM-related discussions, were compared with posts from the Casual Discussion subforum, intended for non-STEM-related topics. It was hypothesized that posts in the STEM-Talk subforum would have a higher proportion and a higher absolute number of STEM words than posts in the Casual Discussion subforum.

12.1 Method

Participants and text material The forum posts of 75 students ($M = 14.07$ years, $SD = 2.00$) from the 2011–12 mentoring year were analyzed. These students contributed posts to both the STEM-Talk and Casual Discussion subforums. The text material consisted of 252 forum posts in the STEM-Talk and 508 posts in the Casual Discussion subforum.

Measures and data analysis The STEM dictionary was applied to determine the proportion and absolute number of STEM words in posts from both subforums. The nonparametric Wilcoxon signed-rank test was used to compare mean differences, as the data were not normally distributed (Bühner & Ziegler, 2010).

12.2 Results and discussion

Descriptive analyses and Wilcoxon signed-rank tests revealed significant differences between the two subforums. In the Casual Discussion subforum, students wrote significantly more posts ($M_{STEM\text{-}Talk} = 3.36$ posts, $SD = 3.23$; $M_{Casual\ Discussion} = 6.77$ posts, $SD = 6.88$; $Z = -4.52$, $p < .001$) and words ($M_{STEM\text{-}Talk} = 228.77$ words, $SD = 241.27$; $M_{Casual\ Discussion} = 521.53$ words, $SD = 629.84$; $Z = -4.42$, $p < .001$) than in STEM-Talk. However, as expected, posts in the STEM-Talk subforum contained significantly more STEM words than posts in the Casual Discussion subforum. The proportion of STEM words was also significantly higher in the STEM-Talk subforum. Supplementary Table A.7 summarizes the descriptive statistics and statistical test results.

These findings confirm the construct validity of the STEM dictionary by demonstrating its ability to distinguish between texts with different levels of domain-specific communication about STEM content. The results provide further empirical evidence supporting the dictionary's capacity to capture STEM communication in educational settings, such as CyberMentor.

13 Study 9: Concurrent validity of the STEM dictionary for positive mentoring outcomes

Criterion validity encompasses both concurrent validity and predictive validity. This study focused on the concurrent validity of the STEM dictionary by examining its relation to self-reported positive mentoring outcomes. Prior research, outside the context of STEM mentoring, has shown that mentoring communication focused on program-relevant content is closely linked to the effectiveness of mentoring programs (Alwani et al., 2023; Parra et al., 2002; Scielzo et al., 2011). In this study, STEM-related communication by mentees was analyzed using the STEM dictionary to examine associations with self-reported mentoring outcomes, such as STEM knowledge, interest, confidence, and elective intentions.

13.1 Method

Participants and text material The study included 425 students ($M = 14.15$ years, $SD = 1.92$) who participated in CyberMentor for the first time during the 2010/11 and 2011/12 mentoring years. In total, these students produced 4,398 forum posts, 120,222 chat posts, and 15,190 emails. For the regression analyses, only students who had written at least 100 words in each communication channel were included ($N = 126$).

Measures STEM communication was assessed as the percentage of STEM words in the students' forum, chat, and email posts using the STEM dictionary. Mentoring outcomes were assessed through questionnaires administered at the beginning and end of the mentoring year. These outcomes were measured using six scales – knowledge of STEM topics (Stoeger et al., 2013), knowledge about studies and occupations in STEM (Stoeger et al., 2013), STEM interest (Ziegler et al., 1998), confidence in one's own STEM abilities (Dweck, 1999; Dweck & Henderson, 1989), STEM activities (Stoeger et al., 2013), and STEM elective intentions (Ziegler & Stoeger, 2008) – each consisting of three to eight items. The internal consistency of these scales was high, with Cronbach's alphas ranging from .80 to .91. Further details about these scales are provided in Supplementary Table A.8.

Data analysis Stepwise regression analyses were conducted to examine whether STEM communication was associated with mentoring outcomes at the second measurement point, controlling for baseline values at the first measurement point. STEM communication variables were log-transformed (log + k) to normalize their distributions (Arguello et al., 2006), and all variables were z-standardized before inclusion in the models (Cohen et al., 2003).

13.2 Results and discussion

The mean proportion of STEM words was 1.12% ($SD = 1.04$) in mentees' emails, 0.87% ($SD = 1.31$) in forum posts, and 0.66% ($SD = 0.93$) in chat posts.

Supplementary Table A.9 presents the detailed regression results. The analysis showed that STEM communication in forum and email posts was significantly associated with five of the six mentoring outcomes at the second measurement point, including knowledge of STEM topics, STEM interest, confidence in STEM abilities, STEM activities, and STEM elective intentions (βs between .16 and .24; ΔR^2 between .02 and .05). STEM communication in chat posts showed no significant associations with any mentoring outcomes. Moreover, no relationship was found between STEM communication and mentees' knowledge about studies and occupations in STEM. This may be because knowledge about studies and occupations requires broader discussions beyond STEM, encompassing topics such as education, career paths, and professional opportunities.

These findings provide support for the concurrent validity of the STEM dictionary. However, the incremental variance explained by STEM-related communication was small, indicating modest effect sizes. Thus, STEM-related communication should

be regarded as only one of several factors that may contribute to positive mentoring outcomes in online mentoring programs. Forum and email posts, which often facilitate more structured and in-depth discussions, showed significant associations with mentoring outcomes. This pattern is consistent with prior research suggesting that communication about program-relevant content is related to mentoring effectiveness (Alwani et al., 2023; Parra et al., 2002; Scielzo et al., 2011; Stoeger et al., 2021). In contrast, chat communication, typically used for brief and informal exchanges, did not exhibit significant associations with mentoring outcomes.

Accordingly, the present findings should primarily be interpreted as evidence for the criterion validity of the STEM dictionary. Although the observed associations may still be meaningful in combination with other relevant factors, their practical implications should be interpreted with caution. At the same time, the effect sizes observed here are comparable to those typically reported in prior studies examining associations between linguistic indicators and self-reported psychological constructs (Bernard et al., 2016). Future research should examine additional elements, such as the emotional tone of STEM communication, to gain a deeper understanding of the conditions for successful online mentoring.

14 Study 10: Predictive validity of the STEM dictionary for long-term participation

This study addressed the second aspect of criterion validity, predictive validity, by examining whether the STEM dictionary predicts long-term participation in the CyberMentor online mentoring program. Long-term participation is a key indicator of successful mentoring, as previous research has demonstrated that the duration of mentoring positively influences outcomes (Grossman & Rhodes, 2002; Jekielek et al., 2002). Specifically, we investigated whether STEM communication was associated with mentees' re-registration for a subsequent mentoring year.

14.1 Method

Participants and text material The study included two cohorts of first-time CyberMentor participants: 747 students ($M = 14.22$ years, $SD = 2.15$) from the 2010–11 mentoring year, and 622 students ($M = 14.27$ years, $SD = 2.08$) from the 2011–12 mentoring year. In the 2010–11 cohort, mentees produced 4,278 forum posts, 103,237 chat posts, and 15,303 emails. In 2011–12, they wrote 2,679 forum posts, 103,797 chat posts, and 12,926 emails.

Measures The STEM dictionary was used to analyze the proportion and absolute number of STEM words in the mentees' forum, chat, and email posts. Long-term participation was measured by re-registration for another mentoring year, with participation coded as 1 (re-registered) or 0 (not re-registered).

Data analysis Stepwise logistic regressions (method: Forward Wald, P_{in}: .05, P_{out}: .10) were conducted to examine the relationship between STEM communication and

long-term participation. Separate models analyzed the proportion of STEM words and the absolute number of STEM words as predictors of the outcome. Variables were logarithmized (log + k) to normalize their distribution (Arguello et al., 2006) and z-standardized (Cohen et al., 2003).

14.2 Results and discussion

During the 2010–11 mentoring year, the mean proportion of STEM words was 1.14% (*SD* = 1.12) in emails, 0.70% (*SD* = 1.50) in forum posts, and 0.54% (*SD* = 1.03) in chat posts. In terms of the absolute number of STEM words, emails averaged 20.57 (*SD* = 34.60), chat posts 6.48 (*SD* = 16.35), and forum posts 4.39 (*SD* = 10.15).

During the 2011–12 mentoring year, the mean proportion of STEM words was 1.48% (*SD* = 2.80) in forum posts, 1.19% (*SD* = 1.13) in emails, and 0.58% (*SD* = 0.86) in chat posts. In terms of the absolute number of STEM words, emails averaged 24.20 (*SD* = 46.38), chat posts 10.35 (*SD* = 60.86), and forum posts 4.80 (*SD* = 9.94).

At the end of the 2010–11 mentoring year, 15.1% of mentees re-registered for another year, while 15.6% re-registered at the end of the 2011–12 mentoring year. Supplementary Table A.10 presents the logistic regression results for the proportion of STEM words. STEM communication in forum and chat posts significantly predicted re-registration (Nagelkerke's R^2 = .06), and the model correctly classified 84.1% of mentees into their respective groups. Supplementary Table A.11 presents the regression results for the absolute number of STEM words. The absolute number of STEM words in forum, chat, and email posts significantly predicted re-registration (Nagelkerke's R^2 = .18), and the model correctly classified 84.8% of mentees.

The findings demonstrate the predictive validity of the STEM dictionary in terms of long-term participation in the mentoring program. They show that the absolute number of STEM words is a stronger predictor of long-term participation compared to the proportion of STEM words. This may be because the absolute number better reflects the overall volume of STEM communication, capturing active engagement with STEM content regardless of additional non-STEM topics discussed. By contrast, the proportion of STEM words can underestimate STEM engagement if mentees frequently communicate about non-STEM topics alongside STEM. These results highlight the importance of promoting frequent STEM-related communication in online mentoring programs to encourage sustained participation.

15 General discussion

15.1 Validity evidence and measurement properties of the STEM dictionary

This study aimed to develop a STEM dictionary for analyzing communication in online STEM education and to systematically and comprehensively assess the psychometric properties of the STEM dictionary within an educational research setting, namely an online mentoring program for girls in STEM. By establishing its objectivity, reliability, and validity, the dictionary was intended to serve as a robust

measurement tool to support future research and the evaluation of computer-based educational programs in STEM.

The results of the 10 validation studies jointly provide strong evidence that the STEM dictionary is a valid measure of STEM-related communication in online educational contexts. Across the different analyses, consistent evidence was obtained for the main quality criteria of classical test theory, including objectivity, reliability, and validity. The automated coding procedure ensures a high level of scoring objectivity, while the reliability analyses indicate that the dictionary produces stable and reproducible measurements of STEM-related language use across different communication channels.

Furthermore, the construct and criterion validity analyses indicate that the dictionary-based measure captures meaningful aspects of participants' STEM-related communication. The observed associations with manual content ratings, self-reported communication measures, and mentoring outcomes provide converging evidence that the STEM dictionary identifies domain-specific communication patterns in online learning environments. Taken together, these findings suggest that the STEM dictionary provides a psychometrically sound instrument for analyzing STEM-related discourse in computer-based educational settings. However, these findings should be interpreted in light of the specific educational context in which the dictionary was developed, highlighting the importance of examining its external validity and transferability to other online learning environments and communication platforms.

Moreover, the norm tables developed in Study 2 provide practical guidance for interpreting the results of dictionary-based analyses in online educational contexts. Because STEM-related terminology typically occurs relatively infrequently in natural communication, the distribution of STEM word frequencies is often characterized by strong skewness and a large proportion of zero values. Consequently, individual messages or short text contributions frequently contain no STEM-related terms, even in STEM-focused communication environments. The norm tables therefore provide important reference values that allow researchers and program evaluators to contextualize observed STEM communication levels relative to typical communication patterns within the analyzed environment. For this reason, the interpretation of dictionary-based results is particularly meaningful when applied to aggregated text data, such as larger sets of messages or longer communication periods, rather than to individual short messages. In practical applications, these norms can help distinguish between communication that contains little or no STEM-related content and communication that reflects a stronger topical focus on STEM discussions.

15.2 Methodological implications of dictionary-based text analysis for educational research

Beyond the validation of the STEM dictionary itself, the findings of this study provide broader insights into the use of dictionary-based text analysis in computer-based educational settings. Using the developed STEM dictionary as an example, the results demonstrate that dictionary-based text analysis can be a robust and scalable method for examining domain-specific communication in educational online contexts. While the procedural objectivity of dictionary-based text analyses can generally be consid-

ered perfect, as all analyzed texts are processed identically through fully automated coding (see Sect. 4), our study suggests that their scoring objectivity is also high, even when applied to uncorrected text material. Their objectivity of interpretation can be enhanced through norming procedures that establish reference points or comparison standards for different participant groups, thereby improving the comparability of results and ensuring their applicability in educational research.

The measurement accuracy of dictionary-based text analyses was assessed using parallel-test reliability, comparing different dictionary components. The results indicate that the reliability of the STEM dictionary is very high to excellent. The split-half reliability, which evaluates internal consistency, was also found to be good. However, retest reliability was only moderate to satisfactory. This is likely due to the dynamic nature of domain-specific communication, which may fluctuate over time, especially in informal digital interactions where students can freely discuss both STEM and non-STEM topics.

The content validity of the STEM dictionary was established through expert ratings during its development. Its construct validity was further established using three methodological approaches: comparisons with manual coding, self-reports, and analyses of texts with average versus extreme levels of STEM-related content. Finally, criterion validity was established in two ways: concurrent validity through self-reported positive mentoring outcomes, and predictive validity through long-term program participation.

The findings highlight that dictionary-based text analysis is a valuable tool for assessing the effectiveness of computer-based educational measures in STEM. By automatically analyzing large volumes of text data generated in digital learning environments, this approach offers an objective, scalable, and non-reactive complement to traditional survey-based evaluations.

15.3 Dictionary-based text analysis in educational data analysis

Dictionary-based text analysis offers several advantages compared to other text analysis methods. One key benefit is its transparent and interpretable word-matching approach, which allows for a clear and systematic analysis of text and language data. In contrast to machine learning and AI-based methods, which often function as "black boxes" with opaque decision-making processes derived from training data, dictionary-based approaches provide explicit and comprehensible classifications. Additionally, dictionary-based text analysis ensures consistent and reliable results, as automated word-matching operates independently of text volume and remains unaffected by variations in human coders' training or attentiveness, which can affect the reliability of manual content analysis. A significant advantage of this method is its efficiency in processing large-scale text data within seconds, without requiring extensive data preparation or preformatting (Mehl, 2006).

At the same time, advances in natural language processing (NLP), including embedding-based and transformer-based language models, enable more context-sensitive analyses of textual data. Such approaches can capture implicit references, contextual meaning, and more complex semantic relationships between concepts. Future research could therefore combine dictionary-based approaches with these

more advanced NLP techniques. For example, the STEM dictionary could be used to identify explicit topical markers in large communication datasets that may serve as features or validation benchmarks for more complex machine learning models. In this way, dictionary-based and machine learning approaches can complement each other by combining transparent rule-based measurement with more context-sensitive language modeling (Dobbrick et al., 2022; van Atteveldt et al., 2021; Widmann & Wich, 2023).

As big data and educational data mining gain increasing significance, and vast amounts of educationally relevant text data are continuously generated through social media and digital learning environments (Boyd & Pennebaker, 2016; Chung & Pennebaker, 2014; Ekin & Sabamehr, 2024; Peña-Ayala, 2014; Wiedemann, 2013), efficient methods for analyzing large text corpora are essential. Dictionary-based text analysis complements traditional survey methods by providing an additional, non-reactive source of data that does not rely on self-reports (Bantum & Owen, 2009; Hill et al., 2014). Unlike surveys, which may be subject to response biases, dictionary-based methods enable researchers to analyze naturally occurring language use in educational contexts. Furthermore, dictionary-based analyses are not limited to small text samples but can be applied to large and comprehensive corpora, making them particularly well-suited for longitudinal studies (Acerbi et al., 2013; Twenge et al., 2012). This scalability, along with its transparency and efficiency, highlights the method's value for analyzing communication in digital learning environments and assessing the effectiveness of computer-based educational interventions.

However, the quality assurance of dictionary-based text analyses remains an open question in the research literature (Donohue et al., 2014). In educational research, evaluating this method through classical test theory provides a suitable framework, as psychometric validation is essential for establishing new measurement instruments in this field (Ingenkamp & Lissmann, 2008). While previous studies have examined single quality aspects of dictionary-based analyses (Alpers et al., 2005; Bantum & Owen, 2009; Tov et al., 2013), a comprehensive validation of key psychometric criteria has been lacking. Moreover, prior research has primarily focused on evaluating dictionary-based analyses for function words (e.g., pronouns such as "I," "we," and "our") and domain-independent content words (e.g., emotion-related terms like "anxious," "excited," and "happy"), rather than on domain-specific vocabulary.

15.4 Limitations and directions for future research

Context insensitivity One limitation of dictionary-based text analyses is their inability to account for contextual meaning, as they analyze words in isolation. Mehl (2006) argues that for testing psychological hypotheses, it is sufficient to determine whether specific concepts appear in a text, even if grammatical relationships between words are ignored. This is particularly relevant for the analysis of domain-specific communication, where the presence of irony or negation does not fundamentally alter the subject matter (e.g., *I am not talking about computer science*). However, the inability of dictionary-based analyses to capture contextual meaning remains a methodological limitation, as it can introduce measurement inaccuracies. Consequently, this method is best suited for research questions focused on manifest textual features

rather than deeper semantic or contextual interpretations. Several approaches have been proposed to address this limitation. For example, dictionary-based text analyses can be combined with qualitative coding methods (Kuckartz, 2010) or supplemented with corpus-linguistic techniques (Oberlander & Gill, 2006; Pollach, 2012). Ultimately, the suitability of dictionary-based text analysis must be carefully evaluated for each educational research study to ensure that the selected analytical approach aligns with both the collected data and the research questions.

Limited generalizability beyond the CyberMentor context The findings of this study are not fully generalizable beyond the specific context of CyberMentor, as the STEM dictionary was optimized for this online mentoring program. Generalizability is further constrained by the fact that students who participate in CyberMentor tend to have a higher-than-average interest in STEM compared to their peers. Consequently, the developed norm tables for STEM communication among CyberMentor participants may not apply to broader student populations, such as those engaging in public discussion forums. Although the STEM dictionary was designed to capture STEM-related communication as broadly as possible, it should still be adapted for use in different contexts. Due to its modular structure, the dictionary allows not only for the substitution of its inductive component – originally derived from the most frequently used STEM-related words in CyberMentor discussions – with a new inductive module based on the most frequent STEM-related terms in the target context, but also for the addition of further inductive components. A promising avenue for future research involves tailoring the STEM dictionary to specific online learning environments, such as MOOCs, STEM discussion forums, or educational gaming platforms. However, any such application would require adjustments to the inductive dictionary component to align with the linguistic characteristics of the new setting. Nonetheless, the modular structure of the STEM dictionary offers a flexible foundation for adaptation, making it a reliable tool for capturing domain-specific STEM communication across diverse settings.

Sample characteristics The empirical analyses in this study were based on communication data from a mentoring program designed to support girls in STEM. As a result, the validation of the STEM dictionary relies exclusively on communication from female participants. Future studies could apply the STEM dictionary to datasets that include contributions from both male and female participants in order to examine potential gender differences in the frequency of STEM-related discussion in online learning environments (Kanze et al., 2018; Madera et al., 2009; Newman et al., 2008).

Translation and expansion for English-speaking contexts Currently, the STEM dictionary is designed for German-language texts. To increase its applicability, a crucial next step is to translate, expand, and validate the dictionary for English-language educational research. This adaptation would enhance its relevance for a broader range of studies and facilitate cross-linguistic comparisons in STEM education research. This should be particularly feasible given that STEM terminology is comparatively unambiguous and can thus be translated across languages with relative ease.

Combination with existing LIWC dictionaries In future research, dictionary-based text analysis could be further refined by combining the STEM dictionary with existing LIWC dictionaries (Boyd et al., 2022; Meier et al., 2019). For example, LIWC's emotion-word dictionaries could be used to identify positive or negative affect in STEM-related discussions and relate these patterns to self-reported mentoring outcomes or students' academic emotions in STEM subjects. Additionally, combining the STEM dictionary with LIWC dictionaries covering life domains such as school, career, and leisure could allow for a more nuanced investigation of students' STEM communication. By addressing these limitations and exploring future directions, dictionary-based text analysis can continue to evolve as a valuable methodological tool in educational research.

15.5 Conclusion

This article highlights the value of dictionary-based text analysis as an empirically validated extension to the methodological repertoire of educational research. By systematically evaluating its psychometric properties in an educational setting, this research has established the STEM dictionary as an objective, reliable, and valid instrument for analyzing domain-specific communication in online mentoring. The findings demonstrate that dictionary-based approaches can support research and evaluation of computer-based educational measures by providing scalable and non-reactive insights into STEM-related discourse. Additionally, this study contributes to the broader field of automated text analysis by applying classical test theory in a comprehensive and systematic way to validate dictionary-based methods, thereby enhancing their methodological transparency and reliability. Beyond online mentoring, the STEM dictionary can be adapted for other educational contexts to explore domain-specific learning processes further.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10639-026-13979-1.

Author contributions **Michael Heilemann:** Conceptualization, Methodology, Data Curation, Formal Analysis, Writing – original draft. **Heidrun Stoeger:** Conceptualization, Supervision, Writing – review & editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets used in the studies presented in this article are available from the first author upon request. The STEM dictionary is available at https://doi.org/10.17605/OSF.IO/3JC9X.

Declarations

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the author(s) used ChatGPT-4o in order to improve language and readability. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Competing interest None.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLoS ONE, 8*(3), Article e59030. https://doi.org/10.1371/journal.pone.0059030

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior, 21*(2), 361–376. https://doi.org/10.1016/j.chb.2004.02.008

Alwani, N. A., Lyons, M. D., & Edwards, K. D. (2023). Examining heterogeneity in mentoring: Associations between mentoring discussion topics and youth outcomes. *Journal of Community Psychology, 51*(3), 1233–1254. https://doi.org/10.1002/jcop.22938

Anthony, L. (2014). *AntConc (Version 3.4.3)*. Waseda University.

Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., & Wang, X. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries & G. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 959–968). New York, NY: ACM. https://doi.org/10.1145/1124772.1124916

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment, 21*(1), 79–88. https://doi.org/10.1037/a0014643

Baran, E., Canbazoglu Bilici, S., Mesutoglu, C., & Ocak, C. (2019). The impact of an out-of-school STEM education program on students' attitudes toward STEM and STEM careers. *School Science and Mathematics*, *119*(4), 223–235. https://doi.org/10.1111/ssm.12330

Bernard, J. D., Baddeley, J. L., Rodriguez, B. F., & Burke, P. A. (2016). Depression, language, and affect. *Journal of Language and Social Psychology, 35*(3), 317–326. https://doi.org/10.1177/0261927X15589186

Blömeke, S. (2013). *Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor [Validation as a task in the research program Competence modelling and competence assessment in the higher education sector]*. Humboldt-Universität.

Borchers, C., Rosenberg, J. M., Gibbons, B., Burchfield, M. A., & Fischer, C. (2021). To scale or not to scale: Comparing popular sentiment analysis dictionaries on educational Twitter data. *Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021*, 619–624.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* [Research methods and evaluation for human and social scientists]. Springer. https://doi.org/10.1007/978-3-540-33306-7

Boyd, R. L., & Pennebaker, J. W. (2016). A way with words: Using language for psychological science in modern era. In C. V. Dimofte, C. P. Haugtvedt, & R. F. Yalch (Eds.), *Consumer psychology in a social media world* (pp. 222–236). Routledge.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology, 40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22.* University of Texas at Austin. https://www.liwc.app

Bozkurt, A., Ucar, H., Durak, G., & Idin, S. (2019). The current state of the art in STEM research: A systematic review study. *Cypriot Journal of Educational Sciences*, *14*(3), 374–383. https://doi.org/10.18844/cjes.v14i3.3447

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. [Introduction to test and questionnaire construction]. Pearson.

Bühner, M., & Ziegler, M. (2010). *Statistik für Psychologen und Sozialwissenschaftler*. [Statistics for psychologists and social scientists]. Pearson.

Chaku, N., Kelly, D. P., & Beltz, A. M. (2021). Individualized learning potential in stressful times: How to leverage intensive longitudinal data to inform online learning. *Computers in Human Behavior, 121*, Article 106772. https://doi.org/10.1016/j.chb.2021.106772

Chiang, F.-K., Zhang, Y., Zhu, D., Shang, X., & Jiang, Z. (2022). The influence of online STEM education camps on students' self-efficacy, computational thinking, and task value. *Journal of Science Education and Technology, 31*(4), 461–472. https://doi.org/10.1007/s10956-022-09967-y

Chung, C. K., & Pennebaker, J. W. (2014). Counting little words in big data: The psychology of communities, culture, and history. In J. P. Forgas, O. Vincze, & J. László (Eds.), *Social cognition and communication* (pp. 25–42). Psychology.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Dobbrick, T., Jakob, J., Chan, C. -H., & Wessler, H. (2022). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures, 16*(4), 303–320. https://doi.org/10.1080/19312458.2021.1999913

Donohue, W. A., Liang, Y., & Druckman, D. (2014). Validating LIWC dictionaries: The Oslo I Accords. *Journal of Language and Social Psychology, 33*(3), 282–301. https://doi.org/10.1177/0261927X13512485

DuBois, D. L., & Neville, H. A. (1997). Youth mentoring: Investigation of relationship characteristics and perceived benefits. *Journal of Community Psychology, 25*(3), 227–234. https://doi.org/10.1002/(SICI)1520-6629(199705)25:3%3C;227::AID-JCOP1%3E;3.0.CO;2-T

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Taylor & Francis.

Dweck, C. S., & Henderson, V. L. (1989). *Theories of intelligence: Background and measures.* Paper presented at the Biennial Meeting of the Society for Research in Child Development, Kansas City, MO.

Ekin, C. Ç., & Sabamehr, M. (2024). Text mining and topic modeling in education: Revealing insights from educational textual data. In M. S. Khine (Ed.), *Text mining in educational research* (pp. 133–151). Springer. https://doi.org/10.1007/978-981-97-7858-4_8

Finkelstein, L. M., Allen, T. D., Ritchie, T. D., Lynch, J. E., & Montei, M. S. (2012). A dyadic examination of the role of relationship characteristics and age on relationship satisfaction in a formal mentoring programme. *European Journal of Work and Organizational Psychology, 21*(6), 803–827. https://doi.org/10.1080/1359432X.2011.594574

Grossman, J. B., & Rhodes, J. E. (2002). The test of time: Predictors and effects of duration in youth mentoring relationships. *American Journal of Community Psychology, 30*(2), 199–219. https://doi.org/10.1023/A:1014680827552

Hamp, B., & Feldweg, H. (1997). GermaNet – A lexical-semantic net for German. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo & Y. Wilks (Eds.), *Proceedings of the ACL/EACL-97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (S. 9–15). ACL.

Hartig, J., Frey, A., & Jude, N. (2012). Validität [Validity]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire construction] (pp. 143–171). Springer. https://doi.org/10.1007/978-3-642-20072-4_7

Heilemann, M. (2015). *MINT-Wörterbuch für die diktionärsbasierte Textanalyse mit LIWC* [STEM dictionary for dictionary-based text analysis with LIWC]. https://doi.org/10.17605/OSF.IO/3JC9X

Henrich, V., & Hinrichs, E. (2010). GernEdiT – The GermaNet editing tool. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (S. 2228–2235). ELRA.

Hill, A. D., White, M. A., & Wallace, J. C. (2014). Unobtrusive measurement of psychological constructs in organizational research. *Organizational Psychology Review, 4*(2), 148–174. https://doi.org/10.1177/2041386613505613

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion*, *89*, 418–436. https://doi.org/10.1016/j.inffus.2022.08.024

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. [Textbook of pedagogical diagnostics]. Beltz.

Jekielek, S. M., Moore, K. A., Hair, E. C., & Scarupa, H. J. (2002). Mentoring: A promising strategy for youth development. *Child Trends*. https://doi.org/10.1037/e479692006-001

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology, 120*(2), 263–286. https://doi.org/10.2307/20445398

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal, 61*(2), 586–614. https://doi.org/10.5465/amj.2016.1215

Kefalis, C., & Drigas, A. (2019). Web based and online applications in STEM education. *International Journal of Engineering Pedagogy*, *9*(4), 76–85. https://doi.org/10.3991/ijep.v9i4.10691

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kuckartz, U. (2010). *Einführung in die computergestützte Analyse qualitativer Daten [Introduction to computer-assisted analysis of qualitative data]*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92126-6

Kunze, C., & Lemnitzer, L. (2007). *Computerlexikographie [Computer lexicography]*. Gunter Narr.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. [Test construction and test analysis]. Beltz.

Lin, C.-W., Lin, M.-J., Wen, C.-C., & Chu, S.-Y. (2016). A word-count approach to analyze linguistic patterns in the reflective writings of medical students. *Medical Education Online, 21*(1), Article 29522. https://doi.org/10.3402/meo.v21.29522

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, S., Kang, L., Liu, Z., Zhao, L., Yang, Z., & Su, Z. (2023a). Exploring the relationships between students' network characteristics, discussion topics and learning outcomes in a course discussion forum. *Journal of Computing in Higher Education, 35*(3), 487–520. https://doi.org/10.1007/s12528-022-09335-0

Liu, Z., Kong, W., Peng, X., Yang, Z., Liu, S., Liu, S., & Wen, C. (2023b). Dual-feature-embeddings-based semi-supervised learning for cognitive engagement classification in online course discussions. *Knowledge-Based Systems, 259*, Article 110053. https://doi.org/10.1016/j.knosys.2022.110053

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604. https://doi.org/10.1093/hcr/28.4.587

Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D. (2020). Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *AERA Open*, *6*(3), 1–27. https://doi.org/10.1177/2332858420940312

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology, 94*(6), 1591–1599. https://doi.org/10.1037/a0016539

Marszałek, M., Miązek, A., & Roczniewska, M. (2023). Promotion and prevention regulatory focus LIWC dictionary. Polish adaptation and validation. *PLoS One, 18*(7), Article e0288726. https://doi.org/10.1371/journal.pone.0288726

McDonnell, M., Owen, J. E., & Bantum, E. O. C. (2020). Identification of emotional expression with cancer survivors: Validation of Linguistic Inquiry and Word Count. *JMIR Formative Research, 4*(10), Article e18246. https://doi.org/10.2196/18246

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141–156). APA. https://doi.org/10.1037/11383-011

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857–870. https://doi.org/10.1037/0022-3514.84.4.857

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). *LIWC auf Deutsch: The development, psychometrics, and introduction of DE- LIWC2015.* https://doi.org/10.31234/osf.io/uq8zt

Müller, C., & Mildenberger, T. (2021). Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education. *Educational Research Review, 34*, Article 100394. https://doi.org/10.1016/j.edurev.2021.100394

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*, 211–236. https://doi.org/10.1080/01638530802073712

Newman, D. S., Guiney, M. C., & Barrett, C. A. (2015). Language use in consultation: Can "we" help teachers and students? *Consulting Psychology Journal: Practice and Research, 67*(1), 48–64. https://doi.org/10.1037/cpb0000028

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*(4), 337–360. https://doi.org/10.1177/026192702237953

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42*(3), 239–270. https://doi.org/10.1207/s15326950dp4203_1

Parra, G. R., DuBois, D. L., Neville, H. A., Pugh-Lilly, A. O., & Povinelli, N. (2002). Mentoring relationships for youth: Investigation of a process-oriented model. *Journal of Community Psychology, 30*(4), 367–388. https://doi.org/10.1002/jcop.10016

Peña-Ayala, A. (Ed.). (2014). *Educational data mining. Applications and trends.* Springer. https://doi.org/10.1007/978-3-319-02738-8

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers and Education, 143*, Article 103673. https://doi.org/10.1016/j.compedu.2019.103673

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8*(3), 162–166.

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007a). *Linguistic Inquiry and Word Count: LIWC2007.* LIWC.net.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007b). *The development and psychometric properties of LIWC (2007).* LIWC.net.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE, 9*(12), 1–10. https://doi.org/10.1371/journal.pone.0115844

Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, *15*(2), 263–287. https://doi.org/10.1177/1094428111417451

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Rammstedt, B. (2004). *Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung* [Determination of the quality of multi-item scales: An introduction]. GESIS. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-201443

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology, 32*(4), 469–479. https://doi.org/10.1177/0261927X13476869

Rude, S. S., Gortner, E.-M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Savicki, V., & Price, M. V. (2021). Reflection in transformative learning: The challenge of measurement. *Journal of Transformative Education, 19*(4), 366–382. https://doi.org/10.1177/15413446211045161

Schermelleh-Engel, K., & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung [Methods of determining reliability]. In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]* (pp. 119–141). Springer. https://doi.org/10.1007/978-3-642-20072-4_6

Schnurr, P. P., Rosenberg, S. D., Oxman, T. E., & Tucker, G. J. (1986). A methodological note on content analysis: Estimates of reliability. *Journal of Personality Assessment, 50*(4), 601–609. https://doi.org/10.1207/s15327752jpa5004_7

Scielzo, S. A., Patel, A., & Smith-Jentsch, K. A. (2011). Academic mentoring relationship communication processes and participant-reported effectiveness. *Journal of Organizational Psychology, 11*(2), 81–93.

Sell, J., & Farreras, I. G. (2017). LIWC-ing at a century of introductory college textbooks: Have the sentiments changed? *Procedia Computer Science*, *118*, 108–112. https://doi.org/10.1016/j.procs.2017.11.151

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods, 13*(2), 320–347. https://doi.org/10.1177/1094428109335949

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*, 63–75. https://doi.org/10.1016/j.jrp.2006.01.006

Stoeger, H., Duan, X., Schirner, S., Greindl, T., & Ziegler, A. (2013). The effectiveness of a one-year online mentoring program for girls in STEM. *Computers & Education, 69*, 408–418. https://doi.org/10.1016/j.compedu.2013.07.032

Stoeger, H., Schirner, S., Laemmle, L., Obergriesser, S., Heilemann, M., & Ziegler, A. (2016). A contextual perspective on talented female participants and their development in extracurricular STEM programs. *Annals of the New York Academy of Sciences, 1377*(1), 53–66. https://doi.org/10.1111/nyas.13116

Stoeger, H., Heilemann, M., Debatin, T., Hopp, M. D. S., Schirner, S., & Ziegler, A. (2021). Nine years of online mentoring for secondary school girls in STEM: an empirical comparison of three mentoring formats. *Annals of the New York Academy of Sciences, 1483*, 153–173. https://doi.org/10.1111/nyas.14476

Sun, D., Zhan, Y., Wan, Z. H., Yang, Y., & Looi, C. K. (2023). Identifying the roles of technology: A systematic review of STEM education in primary and secondary schools from 2015 to 2023. *Research in Science & Technological Education*, 1–25. https://doi.org/10.1080/02635143.2023.2251902

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. *Psychological Assessment, 25*(4), 1069–1078. https://doi.org/10.1037/a0033007

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Male and female pronoun use in U.S. books reflects women's status, 1900–2008. *Sex Roles*, *67*(9–10), 488–493. https://doi.org/10.1007/s11199-012-0194-7

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Watson, C. (2023). An online STEM program for gifted students of color amidst COVID-19. *The Journal of STEM Outreach*, *6*(2), 1–12. https://doi.org/10.15695/jstem/v6i2.08

Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, *31*(4), 626–641. https://doi.org/10.1017/pan.2022.15

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research*, *38*(4), 332–357. http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität – Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystem und Ratingskalen [Rater agreement and rater reliability – Methods for determining and improving the reliability of assessments using category systems and rating scales]*. Hogrefe.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry und Word Count [Computer-assisted quantitative text analysis: Equivalence and robustness of the German version of the Linguistic Inquiry and Word Count.]. *Diagnostica, 54*(2), 85–98. https://doi.org/10.1026/0012-1924.54.2.85

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Young, J. R., Ortiz, N., & Young, J. L. (2016). STEMulating interest: A meta-analysis of the effects of out-of-school time on Student STEM Interest. *International Journal of Education in Mathematics Science and Technology*, *5*(1), 62–74. https://doi.org/10.18404/ijemst.61149

Ziegler, A., & Stoeger, H. (2008). Effects of role models from films on short-term ratings of intent, interest, and self-assessment of ability by high school youth: A study of gender-stereotyped academic subjects. *Psychological Reports, 102*(2), 509–531. https://doi.org/10.2466/PR0.102.2.509-531

Ziegler, A., Dresel, M., & Schober, B. (1998). *Messung motivationsbezogener Schüler(innen)merkmale* [Measurement of motivational characteristics of students]. LMU.