Journal of Business Research 199 (2025) 115571

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Avoiding algorithm errors in textual analysis: A guide to selecting software, and a research agenda toward generative artificial intelligence

Janice Wobst [a], Rainer Lueg [a,b,c,*]

[a] Institute of Management, Accounting and Finance, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany
[b] International Institute of Management, Europa-Universität Flensburg, Auf dem Campus 1, 24939 Flensburg, Germany
[c] Department of Business and Management, University of Southern Denmark, Universitetsparken 1, 6000 Kolding, Denmark

ARTICLE INFO

JEL codes:
C80
C88
M10
M15
L86

Keywords:
Generative AI
Large language models
Textual analysis
Software selection
Algorithm error
Validity
Reliability
Value-based management

ABSTRACT

The use of textual analysis is expanding in organizational research, yet software packages vary in their compatibility with complex constructs. This study helps researchers select suitable tools by focusing on phrase-based dictionary methods. We empirically evaluate four software packages—LIWC, DICTION, CAT Scanner, and a custom Python tool—using the complex construct of value-based management as a test case. The analysis shows that software from the same methodological family produces highly consistent results, while popular but mismatched tools yield significant errors such as miscounted phrases. Based on this, we develop a structured selection guideline that links construct features with software capabilities. The framework enhances construct validity, supports methodological transparency, and is applicable across disciplines. Finally, we position the approach as a bridge to AI-enabled textual analysis, including prompt-based workflows, reinforcing the continued need for theory-grounded construct design.

1. Introduction

A growing number of studies in organizational and management research incorporate textual analysis to quantify textual information (e.g., Marshall et al., 2022; McKenny et al., 2018; Short et al., 2010; Short et al., 2018). Textual analysis is an automated content analysis approach that applies natural language processing (NLP) to extract and quantify textual information (Bochkay et al., 2023; Loughran & McDonald, 2016; Short & Palmer, 2008). One popular NLP method is the dictionary-based approach that counts words and phrases (collocations of words) in texts (Short et al., 2010). Within the families of (un)supervised machine learning and dictionary-based approaches, this study specifically focuses on targeted phrase-based dictionary methods to capture the complexity of value-based constructs. Many phenomena in the field of management are not directly or indirectly observable (i.e., constructs) and require representation through observable measures (Babbie, 2020; Bisbe et al., 2007; Godfrey & Hill, 1995). Textual analysis facilitates construct measurement, the replicability of studies, and the scalability of data collection (Duriau et al., 2007; Morris, 1994; Short & Palmer, 2008).

Scholars rely on pre-designed software packages to perform textual analysis (Hickman et al., 2020; Short et al., 2018), each of which has a specific design that may not match the features of the construct being investigated. Using a software package that is incapable of correctly processing the features of the construct constitutes a potential source of measurement error (i.e., algorithm error) (McKenny et al., 2018).

However, there is little guidance to aid scholars in selecting suitable software. Prior research has focused only on providing a compilation of different software packages to juxtapose their functionalities (e.g., Duriau & Reger, 2004; Klein, 2014; Neuendorf, 2016; Short et al., 2018), and the opacity of textual analysis algorithms impairs the matching of software with the features of the construct (Neuendorf, 2016). In particular, less apparent algorithmic functionalities such as word counting conventions or special character treatment require a deep examination of different software. To provide guidance for the software selection, we pose the question: *how can textual analysis software be selected in line with theory to mitigate algorithm error?*

* Corresponding author at: Institute of Management, Accounting and Finance, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany.
E-mail addresses: janice.wobst@leuphana.de (J. Wobst), rainer.lueg@leuphana.de (R. Lueg).

https://doi.org/10.1016/j.jbusres.2025.115571
Received 5 November 2024; Received in revised form 27 June 2025; Accepted 27 June 2025
Available online 28 June 2025

We compare the prevailing software packages in management research and assess the algorithmic fit between four different software packages (LIWC, DICTION, CAT Scanner, and custom Python-based algorithm) to measure value-based management (VBM) sophistication (Wobst et al., 2022). DICTION is designed for single-word matching and remains widely used in quantitatively oriented fields that rely on clearly defined constructs. We include it deliberately to assess how disciplinary acceptance does not ensure methodological fit with highly complex constructs used in less regulated or less quantities fields. This comparison eventually helps quantify how such misalignments can distort results when constructs require semantic precision. We generalize our findings and develop guidelines that aid scholars in selecting software packages that match the nature of their constructs.

A few software packages dominate the management discipline, such as Linguistic Inquiry and Word Count (LIWC) and DICTION. We find that software packages differ in their salient functionalities (e.g., level of analysis) and subtle functionalities (e.g., word counting conventions). A misfit between a software's functionalities and a construct's features is likely to produce an algorithm error. Textual analysis tools fall into different methodological families, such as word-based vs. phrase-based. Our findings suggest that once a suitable family is chosen, results are consistent and valid across different tools, without requiring advanced or customized solutions. Conversely, we also demonstrate the statistical magnitude to which choosing an inappropriate software – often based on its disciplinary popularity – may distort results and compromise construct validity.

This study contributes to research in several ways. It particularly enhances the seminal study by McKenny et al. (2018): We provide both a *decision-making* and a *reporting tool* for computer-aided text analysis (CATA) software selection, introduce a customizable dictionary with extended features that captures complex constructs in new fields of organizational research, and a guideline for practice how to process unstructured data. In detail, our study first structures the software selection process. We analyze the interrater reliability of different software packages and assess the economic materiality of their differences compared to the hand-collected VBM dataset of Firk et al. (2019). Second, it develops generalizable guidelines that may also serve as a reporting tool to transparently outline the reasons for (not) choosing a particular software package. Such transparent reporting facilitates a study's replicability. In doing so, our study complements the method-selection typology developed by Herhausen et al. (2025). While they provide strategic guidance on when to use which type of text analysis method, our work delivers the technical implementation logic for construct-valid software selection once a CATA-based approach is chosen. Third, the study contributes to VBM research in particular by demonstrating which software is capable of measuring the VBM sophistication construct reliably. Fourth, it has implications for practitioners – who are increasingly using textual analysis (Chan et al., 2020) – the study's proposed software selection guidelines may guide practitioners in finding a suitable software package to prevent erroneous decisions. Finally, we position this study as a bridge between current dictionary-based CATA practices and emerging artificial intelligence (AI) and large language models (LLMs)-supported textual analysis workflows beyond our test case: VBM has features such as conceptual complexity, low term prevalence, and strong theoretical anchoring. These characteristics make the proposed guideline applicable to other domain-specific constructs and measurement practices in organizational, strategic, and governance research (Gaur & Kumar, 2018).

2. Software usage in textual analysis research

CATA can be broadly categorized into three main families (Marshall et al., 2022; McKenny et al., 2018; Short et al., 2010) (1) supervised machine learning methods, (2) unsupervised machine learning methods, and (3) dictionary-based approaches. Supervised learning methods require annotated training data to develop classification models, while unsupervised methods infer latent structures such as topics or clusters without labeled inputs. Dictionary-based approaches have two main methods. (3a) *Word-list* methods rely on dictionaries containing single words or stems (*lemmatization*), which are matched against the text corpus to capture relevant constructs (Loughran & McDonald, 2016). These are particularly common in sentiment analysis, where individual, established terms such as 'success' are counted. (3b) Targeted *phrase* methods, by contrast, employ dictionaries consisting of multi-word expressions that represent more complex constructs (Bochkay et al., 2023). Phrase-based approaches allow scholars to account for semantic context and syntactic specificity, and are especially useful when constructs are unlikely to be captured adequately by single terms. The latter method is the focus of our study, as shown in Fig. 1.

A dictionary-based NLP method starts with a theoretical inquiry to define the underlying construct. The theoretical underpinnings drive the dictionary choice (methodological inquiry) (Neuendorf, 2016). Scholars can either use a *built-in* dictionary, create their own (*customized dictionary*), or use a customized dictionary developed by prior studies (Neuendorf, 2016). *Built-in* dictionaries are an integral part of a software package and generally undergo rigorous reliability and validity assessments; thus, they are the preferred option to measure a construct, if available. The choice of dictionary then determines the software selection (i.e., inquiry of the measurement instrument) (McKenny et al., 2018; McKenny et al., 2013; Short et al., 2010).

Prior research provides guidelines for dictionary creation and validation (e.g.. Pandey & Pandey, 2019; Short et al., 2010). Hickman et al. (2020) describe important steps for the preprocessing of text data, and state that the outcomes of textual analysis depend highly on the choice of software. Dictionary-based NLP methods often rely on pre-designed software packages (Hickman et al., 2020; Short et al., 2010). However, the flexibility to adjust the software functionalities to the construct's idiosyncratic requirements remains limited (Duriau & Reger, 2004). Management scholars predominantly use software packages such as Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022) and DICTION (Hart, 2014; Hickman et al., 2020). LIWC offers reliable single-word matching and built-in dictionaries, but has limited ability to process multi-word phrases or construct-specific expressions. DICTION is optimized for political and rhetorical content using pre-defined word lists, but lacks phrase recognition and performs poorly with complex, domain-specific constructs. Authors of software packages often design dictionaries that become an integral part of their software (Neuendorf, 2016). LIWC and DICTION are the preferred option when using built-in dictionaries because their *off-the-shelf* dictionaries have undergone rigorous validity and reliability assessments (e.g., Pennebaker et al., 2015; Pennebaker & Francis, 1996; Short & Palmer, 2008). Scholars predominantly use LIWC to conduct sentiment analyses by employing the integrated sentiment dictionary (e.g., Gamache & McNamara, 2019; Hubbard et al., 2018; Love et al., 2017; Pfarrer et al., 2010). A dictionary-based sentiment analysis counts positive, neutral and negative words to capture the tone/sentiment of texts (Bochkay et al., 2023). DICTION is the preferred choice when measuring, for example, political leadership rhetoric because of its explicit development for political discourse (e.g., Bligh & Hess, 2007; Bligh et al., 2004; Bligh & Robinson, 2010; Davis & Gardner, 2012).

There is no prevailing software for customized dictionaries. Other software packages, such as McKenny et al.'s (2012) CAT Scanner, are used in cases of customized dictionaries. CAT Scanner supports user-defined dictionaries and can handle phrases and hyphenated terms, making it suitable for construct-specific textual analysis. The higher variability in software packages used in cases of dictionary customization indicates an increased challenge in matching the software with the construct requirements (Duriau & Reger, 2004).

Scholars acknowledge that prevailing software packages such as LIWC or DICTION reach their limits when examining constructs that go beyond standard textual analysis (McKenny et al., 2018). For instance, Nadkarni et al. (2019) develop a customized program because the

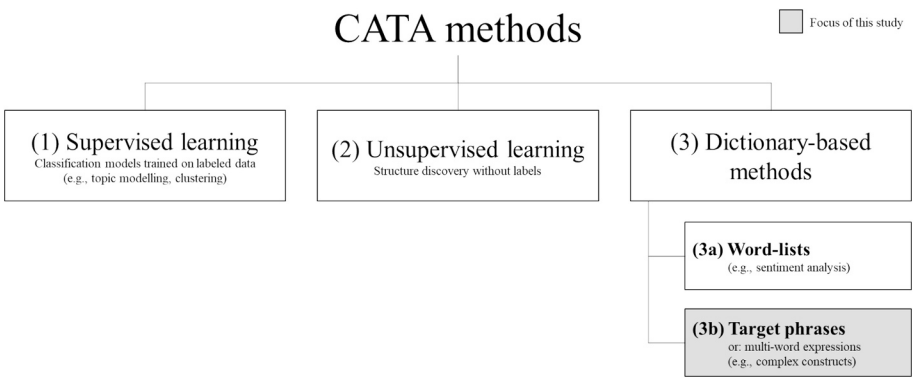

Fig. 1. Typology of CATA methods and positioning of this study.

coding and categorizing of numerical and non-numerical targeted words simultaneously is not possible with LIWC and DICTION.

Many studies do not explicitly discuss their choice of software. However, choosing an inappropriate software package may threaten the construct's reliability (e.g., McKenny et al., 2018) as an algorithm error may occur when the software's design choice does not match the construct (McKenny et al., 2018). Providing a rationale for the choice of software would help other scholars to decide for (or against) a particular program. Studies that do provide reasons for choosing a particular software commonly state that the selection was made by following prior research (Gamache & McNamara, 2019; Love et al., 2017; Rhee & Fiss, 2014). Few studies corroborate their choice of software by theory or match the software functionalities with the construct features (e.g., Guo et al., 2021; McKenny et al., 2018; Murphy & Ensher, 2008).

This study structures the software selection process to counteract the lack of precise guidelines. We respond to McKenny et al.'s (2018) call and contribute to textual analysis research by providing a deeper examination of potential algorithm errors.

3. Methods

3.1. Construct selection

This study develops guidelines that may assist future scholars in choosing which textual analysis software to employ. We use VBM sophistication as an illustrative construct. VBM is an integrated management approach that aligns shareholders' and managers' interests to foster long-term value-enhancing strategies (Firk et al., 2019; Ittner & Larcker, 2001; Martin et al., 2009). The cornerstone of VBM is the establishment of value-based metrics that link a firm's key value drivers with performance measures to ensure value creation above the cost of capital (Burkert & Lueg, 2013; Firk et al., 2019). As an already established construct, VBM is suitable to use in the development of a systematic software selection process (Burkert & Lueg, 2013; Firk et al., 2021; Fiss & Zajac, 2004; Mavropulo et al., 2021), and its use in previous studies precludes confounding effects as a result of theoretical or methodological inconsistencies. Moreover, we can use an already developed VBM dictionary: Wobst et al. (2022) developed a customized dictionary containing VBM-related phrases. Prior dictionaries rely predominantly on word- instead of phrase-level analyses (Pandey & Pandey, 2019), but the use of a phrase-level dictionary enhances the generalizability of the study's findings.

3.2. Construct measurement

In line with the typology of CATA methods outlined above, this study applies a dictionary-based approach, and more specifically a targeted phrase method. This methodological choice reflects the complexity of the VBM construct, which is best captured through multi-word expressions. This design aligns with prior research that highlights the importance of matching the unit of analysis –words versus phrases – with the complexity of the construct (Bochkay et al., 2023; McKenny et al., 2018).

We investigated the algorithmic fit between the most commonly used software and the text-based VBM sophistication measures. We started by comparing the interrater reliability of the chosen software packages. Consequently, we performed a one-way analysis of variance (ANOVA) with a subsequent pairwise comparison. Then, we assessed the convergent validity of the text-based measures to ensure that the software suited the construct's theoretical underpinnings. The correlation between two measures of theoretically related constructs is known as convergent validity (Campbell, 1959; Short et al., 2010); in order to assess the convergent validity of the different text-based measures, the analysis requires an external benchmark and we are grateful for access to a subsample of Firk et al.'s (2019) manually coded VBM sophistication measure. Firk et al. (2019) develop a composite VBM sophistication proxy that synthesizes five binary indicators. The binary indicators account for the depth of VBM across firms' internal hierarchies. Their sample comprises European non-financial firms listed in the STOXX® Europe Total Market Index, and they manually collected information from corporate annual reports in the years 2005 to 2014. Firk et al. (2021) used the same VBM sophistication information to investigate the decision-making role of VBM from 2005 to 2016. See Firk et al. (2019) for further details on their sample and the construction of the VBM sophistication proxy. We adopted the sample used by Wobst et al. (2022): because a cross-sectional data set is sufficient to compare different software, we focused on the year 2010, the beginning year of their sample. Including only non-financial firms which offered annual reports downloadable in English, we collected these reports from the investor relations sections of their websites. The final sample comprises 297 firms.

We began by assessing the eligibility of LIWC and DICTION for measuring the VBM sophistication construct. We complemented the prevailing software by using the CAT Scanner (McKenny et al., 2012). The CAT Scanner is free of charge and used in peer-reviewed studies (e.g., Evert et al., 2018; McKenny et al., 2018; Murray & Fisher, 2022; Vaupel et al., 2022). Neuendorf (2016) outlines the option to create custom programs as an alternative to standard software. Custom programs have the advantage of being highly flexible and allowing a better understanding of the different functionalities (Short et al., 2018). To complement the chosen software, we developed a customized Python-based algorithm. Our Python-based solution is designed to flexibly capture complex phrases, process hyphenated expressions, and deliver output with high resolution, matching all required construct features. The algorithm uses existing Python libraries such as sci-kit learn to ensure reliability. Sci-kit learn is a machine-learning library that includes various algorithms to perform, for example, classifications, regressions, and topic modeling (Pedregosa et al., 2011). Our algorithm is

easily adaptable for other dictionary-based analyses. The output is a document-term matrix that displays the dictionary's raw and total word counts. In summary, we investigated the algorithmic fit of the VBM sophistication construct using four different software packages (LIWC, DICTION, CAT Scanner, and a customized Python-based algorithm).

The variable VBM LIWC is a continuous variable that gives the percentage of total dictionary words in each report, based on the total number of words (standard output of LIWC) (Boyd et al., 2022). We calculated the variables VBM Python (CAT Scanner, DICTION) comparably by taking the sum of dictionary counts divided by the total word count and displaying them as percentage values. Weighting the raw counts by the total word counts counterbalances the varying length of the reports (Loughran & McDonald, 2016).

4. Results

4.1. Descriptive results

Table 1 displays the descriptive statistics for the VBM sophistication constructs. The LIWC-based measure reveals the highest mean value (0.055) and the DICTION-based measure the lowest (0.003). All variables vary substantially. LIWC reveals the highest variation between 0.000 and 0.208. Deviations in mean values may occur either because the software packages differ in counting the dictionary phrases and/or because of differences in total word counts. The diverging descriptive results indicate that the four software packages use differing algorithms. We perform an ANOVA and assess the convergent validity to determine whether the differences in mean values are statistically and economically significant. Below, we outline some reasons for the observed (dis) agreements between the software and develop a systematic software selection process.

4.2. Interrater reliability: Krippendorff's alpha

Panel A of Table 2 displays the Krippendorff's alphas for the results generated for the VBM measure by the employed software packages. Krippendorff's alpha calculates the interrater reliability of coders that can be humans or algorythms (Krippendorff, 2018). This measure is calculated by comparing the observed agreement among coders to the agreement that would be expected by chance. A value close to 1 indicates high agreement among coders, while a value close to 0 indicates agreement no better than chance. Negative values suggest that the agreement is worse than would be expected by chance, indicating issues with the coding scheme or coder performance. Krippendorff's alpha has been employed to evaluate the efficacy of software in accurately measuring constructs within the field of management (McKenny et al., 2018).

The results for our Python-based algorithm and CAT Scanner show high interrater reliability of 0.961. This result is comparable with the findings of McKenny et al. (2018), who compared the interrater reliability for the constructs of entrepreneurial orientation, market orientation, and organizational ambidexterity, for which they report Krippendorff alphas from 0.88 to 0.90 using the software packages LIWC 2007 and DICTION 5.

However, DICTION performs substantially worse in our study than a random guess, evidenced by its negative Krippendorff alphas with all other measurements. This finding reinforces our theoretical argument that highly complex constructs in management, such as VBM, necessitate the use of phrases and hyphenated words for accurate measurement. Despite DICTION's established efficacy in measuring simpler, more easily definable constructs in areas such as financial sentiment analysis (Bochkay et al., 2023), its application to more intricate constructs warrants critical evaluation.

Krippendorff alphas for LIWC, when compared with Python or CAT Scanner, are adequate but not outstanding, with values ranging from 0.452 to 0.569, which fall significantly below the benchmarks set by established studies (McKenny et al., 2018). Researchers might question the appropriate course of action in such a situation. Therefore, we proceed with an analysis of the economic materiality in the following subsection.

4.3. Economic materiality: ANOVA and convergent validity results

Panel B of Table 2 presents the results for the convergent validity of the software measurements of the VBM construct, comparing them with a one-way ANOVA to the hand-collected VBM sophistication benchmark derived from field data (Firk et al., 2019). We included a post-hoc analysis with Bonferroni correction to examine pairwise differences between the software (Mooi et al., 2018). The results indicate a

Table 1
Descriptive statistics.

	N	Mean	Std dev.	Min	Median	Max
VBM sophistication benchmark	297	1.380	1.333	0.000	1.000	5.000
VBM Python	297	0.028	0.021	0.000	0.024	0.130
VBM DICTION	297	0.003	0.006	0.000	0.001	0.055
VBM CAT Scanner	297	0.025	0.019	0.000	0.021	0.116
VBM LIWC	297	0.055	0.041	0.000	0.046	0.208

Notes: LIWC = Linguistic Inquiry and Word Count. The table represents descriptive statistics for the value-based management (VBM) sophistication construct using different software. VBM sophistication benchmark = manually developed VBM sophistication index by Firk et al. (2019). VBM Python (CAT Scanner; DICTION; LIWC) are continuous variables that give the percentage of total dictionary words in each report, based on the total number of words.

Table 2
Convergent validity and Krippendorff's alpha.

	VBM sophistication benchmark	VBM Python	VBM DICTION	VBM CAT Scanner	VBM LIWC
Panel A: Krippendorff's alpha					
		VBM Python	VBM DICTION	VBM CAT Scanner	VBM LIWC
VBM Python		–			
VBM DICTION		−0.317	–		
VBM CAT Scanner		0.961	−0.293	–	
VBM LIWC		0.569	−0.364	0.452	–
Panel B: Convergent validity results					
VBM sophistication benchmark	1.000				
VBM Python	0.575	1.000			
VBM DICTION	0.346	0.476	1.000		
VBM CAT Scanner	0.583	0.978	0.468	1.000	
VBM LIWC	0.566	0.954	0.367	0.974	1.000

Notes: LIWC = Linguistic Inquiry and Word Count.
Panel A presents the Krippendorff's alpha values, indicating the interrater reliability among different software packages for the VBM construct. Higher values signify stronger agreement. Sample interpretation: A Krippendorff's alpha of 0.961 indicates a very high level of agreement (96.1%) between the VBM measurements generated by the Python algorithm and CAT Scanner. Such a high alpha value implies that the two methods can be reliably used interchangeably or in conjunction, as their measurements are almost identical.
Panel B presents the Spearman rank-order correlation results of the value-based management (VBM) sophistication construct using different textual analysis software packages. All correlation coefficients are statistically significant at the 1 % level or less. VBM sophistication benchmark = manually developed VBM sophistication index by Firk et al. (2019). VBM Python (CAT Scanner; DICTION; LIWC) are continuous variables that give the percentage of total dictionary words in each report, based on the total number of words.

statistically significant overall effect of mean differences between the software packages $F(3; 1{,}184) = 210.37$, $p < 0.000$, $R^2 = 0.348$. The post-hoc analysis shows that the text-based VBM sophistication measures differ significantly in mean values between all software packages except that of the Python-based algorithm and CAT Scanner ($M = 0.003$; $p = 0.665$).[1]

We performed a Spearman's rank-order correlation between the text-based measures and Firk et al. (2019) alternative VBM sophistication benchmark. The correlation analysis should reveal whether the differences in mean values between the software packages also materialize economically. If the software packages employ comparable algorithms, we would expect similar correlation results between the different software packages and the alternative benchmark. In such a case, statistically significant differences would not materialize economically. Panel B in Table 2 presents the correlation results. The results yield positive and statistically significant correlations between all text-based VBM sophistication measures and the alternative benchmark. The correlation results provide a similar and strong effect size for the Python, CAT Scanner, and LIWC-based measures (Python: 0.575, CAT Scanner: 0.583, LIWC: 0.566). The statistically significant differences in mean value between the Python-based algorithm and LIWC do not translate into economic significance. Consequently, the mediocre Krippendorff alphas do not diminish the economic materiality of the LIWC measure and should not overly concern researchers. The results are similar for LIWC, CAT Scanner, and our Python-based algorithm.

Regarding the DICTION-based measure, the correlation result supports the ANOVA results. The DICTION-based measure yields a significantly lower correlation with the alternative benchmark (0.346); this low correlation suggests a significant algorithm error. We conclude that the Python algorithm, LIWC, and CAT Scanner provide a suitable fit for measuring the highly complex, text-based VBM sophistication construct, whereas DICTION induces a substantial algorithm error in measuring the VBM construct.

5. Discussion and conclusion

This study investigates the algorithmic fit between different textual analysis software (Hickman et al., 2020; McKenny et al., 2018). The results indicate that the Python-based algorithm, CAT Scanner, and LIWC fit the VBM sophistication construct. DICTION seems to be a misfit. The VBM sophistication construct requires a software package that: (1) enables the integration of customized dictionaries, (2) processes phrases and hyphenated words, and (3) produces an output that displays enough decimal places (up to three) in terms of relative values (dictionary count/total word count) or raw counts of dictionary- and total words. The third criterion is relevant because the high number of total words in relation to VBM-specific phrases in annual reports requires more than two decimal places to reveal the concept properly (Wobst et al., 2022).

DICTION is not able to meet all three criteria because it can only count words and hyphenated words (Hart, 2014); it is not capable of counting phrases. Thus, DICTION's design does not fit the construct's features and leads to an algorithm error. Thereby, our study highlights a broader implication: researchers should not be compelled to use software packages simply because they are 'the most established' ones, such as DICTION. Editors, reviewers, funding agencies, conference participants, or thesis supervisors should refrain from demanding such tools as a benchmark when the construct's linguistic features clearly call for a different, possibly innovative and lesser known tool. Our findings provide an empirical basis for resisting misplaced expectations.

[1] DICTION and CAT Scanner: $M = -0.022$, $p = 0.000$; LIWC and CAT Scanner: $M = 0.030$, $p = 0.000$;LIWC and DICTION: $M = 0.052$, $p = 0.000$; Python and DICTION: $M = 0.025$, $p = 0.000$;Python and LIWC: $M = -0.027$, $p=0.000$.

The three other software packages fulfil the three salient features. More subtle differences (e.g., different word counting conventions) between LIWC, CAT Scanner, and the Python-based algorithm may explain the statistically significant differences in mean values observed (e.g., different treatment of hyphens). The subtle algorithmic differences do not materialize economically because the correlation results are similar. However, subtle differences still require investigation to prevent an algorithm error.

5.1. Systematic software selection guidelines

Fig. 2 presents a systematic software selection process (step 1–7) to allow generalizability to other constructs. (1) The software selection process suggests starting with a theoretical inquiry (defining the construct). This includes assessing the complexity of the construct, as it influences the linguistic structure required for its valid measurement. Constructs such as sentiment or tone are often captured using single terms (e.g., "risk", "opportunity"), whereas more complex constructs like VBM or organizational ambidexterity rely on multi-word expressions (e.g., "value creation above capital cost", "simultaneous exploration and exploitation"). For complex constructs, phrase-based dictionary methods are generally more appropriate than single-word approaches.

(2) Scholars should then choose whether to use a built-in dictionary (which will have already undergone rigorous testing) or to create a customized one. In cases of built-in dictionaries, scholars tend to use the software it is bundled with. For theory-driven research involving complex or domain-specific constructs, customized dictionaries offer superior alignment with construct definitions. In some cases, researchers may need to develop new dictionaries entirely—especially for emerging or multi-dimensional constructs without established terminology. For such cases, Short et al. (2010) outline a structured approach to dictionary development, which includes concept specification, item generation, and validation. In cases with non-English corpora of text, Heyden et al. (2015) demonstrate how adapting keyword lists across languages requires both semantic precision and theoretical anchoring. These steps ensure alignment between the construct's theoretical definition and the dictionary's linguistic representation.

(3) In cases of dictionary customization, we suggest performing an initial software screening. This comprises a matching of the salient construct features with the software functionalities. Salient features refer to the design of the dictionary, such as the use of single- or hyphenated words or phrases, while software functionalities describe the design choice of the software – such as its capability to count phrases. Systematic screening ensures that salient incompatibilities – such as software ignoring phrase boundaries or mishandling hyphenated terms – are identified early. Researchers might also want to document the rationale behind excluding software that lacks critical functionalities and are this mis-aligned with their construct.

(4) The purpose of the screening is to identify software packages that are functionally compatible with the salient features of the construct. This compatibility must be evaluated with respect to technical capabilities such as phrase recognition, handling of hyphenated terms, and dictionary input formats. Publicly available documentation, prior research, or software manuals serve as important sources for evaluating these capabilities. In our context, for example, DICTION and LIWC lacked phrase-level processing, which is critical for constructs like value-based management. By contrast, CAT Scanner and the Python-based algorithm allowed for phrase recognition, supported custom dictionaries, and handled hyphenated expressions reliably. This screening step ensures that only those software options proceed to empirical evaluation which can, in principle, process the construct as theoretically defined.

(5) This should be followed by a quantitative comparison that investigates whether subtle software functionalities may cause potential algorithm errors. Different approaches are available to determine algorithm errors such as Krippendorf's alpha, convergent validity, ANOVA, and Kendall's W (McKenny et al., 2018; Morris, 1994; Short et al., 2010).

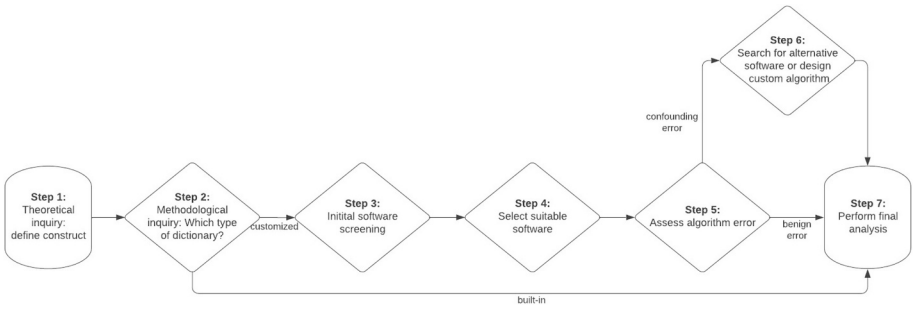

Fig. 2. Overview of a systematic software selection process.

(6) An algorithm error can be either benign or confounding. Benign errors refer to algorithm errors that are not economically meaningful. Confounding errors run the risk of distorting subsequent analyses (Xu et al., 2020).

(7) Scholars may choose any one of the selected software programs and perform the final analysis if no confounding error is present. If the results diverge significantly, ideally they will choose the software with the highest convergent validity results. A confounding error requires searching for an alternative software, or designing (and testing) a customized program.

A systematic software selection process offers several advantages. First, it reduces the likelihood of an inappropriate and arbitrary software selection. A systematic selection aligns the software functionalities with the construct's theoretical underpinnings and minimizes measurement errors (McKenny et al., 2018; Nunnally & Bernstein, 1994). Second, a systematic software selection process increases the study's transparency and replicability if reported adequately. We recommend including a rationale for choosing the underlying software that links the software selection with theory. Third, a systematic software selection process may save time if already conducted by prior research for similar constructs. For example, our illustrative software selection process may serve as orientation for future research and comparable constructs.

6. Contributions

This study contributes to research and practice in several ways. First, it contributes to textual analysis research by introducing a customizable, *Python-based dictionary* with extended features that captures complex constructs. Management research increasingly relies on the dictionary-based NLP method (Bochkay et al., 2023; Loughran & McDonald, 2016) and, although a variety of readily available software packages are available to perform such a NLP task (Boyd et al., 2022; Hart, 2014; McKenny & Short, 2012; Short et al., 2010), not all software is capable of measuring every construct adequately (McKenny et al., 2018). Some constructs require more advanced features, such as phrase recognition or the ability to detect concepts of low prevalence. Our tool addresses these challenges. By explicitly situating our approach within the targeted phrase subcategory of dictionary-based CATA methods, we further clarify the methodological classification within the broader CATA landscape.

Second, we provide researchers with a *decision-making tool* for software selection, thereby providing solutions extant problems (McKenny et al., 2018). The absence of guidelines when to use customized dictionaries leaves scholars navigating 'murky waters' when searching for a suitable software. Our guidelines structure the software selection process with the overall goal of minimizing an algorithmic misfit, and are generalizable to a wide variety of constructs and disciplines. Our findings further show that once a suitable methodological family is selected, results are consistent across tools within that family, even without technical customization. At the same time, we quantify the risk of selecting an unsuitable tool – often chosen due to disciplinary convention – by showing the extent to which such a misfit can distort results and undermine construct validity. Examples include (leadership) characteristics (Anglin et al., 2018), corporate sustainability (Mansouri & Momtaz, 2022), and corporate culture (Li et al., 2020; Pandey & Pandey, 2019). Thereby, we complement the strategic typology of method selection proposed by Herhausen et al. (2025), who offer a conceptual matrix to guide whether to predict, classify, or explore textual data. Our framework extends this logic by empirically operationalizing software selection once a CATA-based approach is chosen—particularly in research settings with difficult, theory-driven constructs. Herhausen et al. (2025) address text analysis at a task level; we add depth at the tool level by testing and validating four software packages across fit criteria. We thereby offer the technical execution layer implied in their broader framework, along with a transition path toward AI-supported construct measurement.

Third, this study offers a *reporting tool* that systemizes the software choice and enhances the transparency of textual analysis research. Prior studies seldomly document their rationale for choosing a particular software (McKenny et al., 2018). A missing rationale impairs the study's transparency and replicability. More transparent software selection processes may also aid future scholars in familiarizing themselves with the key functionalities of different software.

Fourth, it extends CATA to the *field* of corporate governance (McKenny et al., 2018). We demonstrate which software is suitable to measure the established and complex VBM sophistication construct reliably. Matching construct features to software functionalities enables tool selection for both complex constructs like VBM and simpler ones like tone or sentiment. Future scholars can build on this study when investigating other governance systems with differing sophistication, such as various forms of stakeholder- or sustainable management (Wobst et al., 2025). The framework is not limited to VBM and can support construct measurement in any domain involving complex, theory-driven, or low-prevalence textual patterns (Gaur & Kumar, 2018).

Fifth, the guidelines may help *practitioners* in selecting a suitable software. Practitioners increasingly use textual analysis to process unstructured data. For example, practitioners employ textual analysis in their investment decisions (Chan et al., 2020): minimizing an algorithmic misfit reduces noise and ensures astute investment decisions (McKenny et al., 2018).

6.1. Limitations and research propositions: Navigating the changing textual landscape of artificial intelligence and large language models

This study has several limitations that present promising opportunities for future research. For instance, the study's findings are only partially generalizable to other NLP methods, such as machine-learning methods. Different machine-learning libraries may follow different algorithms (Gevorkyan et al., 2019). Future research could investigate how the choice of machine-learning algorithm affects algorithm errors. In the following, we extend the limitations to conceptual propositions to guide subsequent inquiry into software selection, construct measurement, and the evolving role of tools emplyoing AI and LLMs in management research.

"Practice" wants to meet "Theory", not its own avatar.

Recent contributions in management and strategy execution (Mahlendorf et al., 2023; Qiu et al., 2023; Wobst et al., 2025) explicitly show that researchers continue to seek construct-valid, transparent, and theoretically anchored measures. VBM is a paradigmatic example of such a construct: multi-dimensional, domain-specific, and embedded in practitioner language. Textual analysis in this field therefore still requires ex ante theoretical decisions on constructs instead of data-mining the often simplified represenatations in practice or applied research (Wobst et al., 2025). This holds true regardless of whether the downstream text analysis is performed by a human, a dictionary, or an AI. Hermann and Puntoni (2024) emphasize that generative AI applications depend on domain-specific tasks, which reinforces our claim that theory-based construct design remains central, even when AI tools are applied. Thus, our focus on dictionary-based tools does not imply that supervised, qualitative approaches are obsolete. Tools such as NVivo remain essential when researchers work with exploratory data, emergent constructs, or context-specific meanings that resist predefined coding schemes. In such cases, iterative coding and researcher interpretation should enable construct refinement.

7. Proposition 1: Structured constructs will remain foundational in management research, even as AI advances

LLMs still depend on prompts.

LLMs may eventually displace classical dictionary-based tools in textual analysis. Yet, they will not do this as an alternative, but as a technical development: LLMs do not eliminate the need for theoretical construct work that this paper outlays. As recent studies in finance show (Novy-Marx & Velikov, 2025), using LLMs at scale depends on high-quality prompts – often several hundred per paper – which must be grounded in construct-specific terminology. Rezazadeh et al. (2025) show that even GenAI-driven startups rely on prompt workflows grounded in domain knowledge, confirming that semantic clarity remains essential in AI-based textual analysis. Our approach offers exactly this kind of preparatory work. Researchers can use our method not only to select current tools but also to design AI prompts that align with complex constructs, ensuring semantic focus and reproducibility.

8. Proposition 2: Theoretical construct work and dictionary design will directly inform future prompt engineering for AI applications

Prompt engineering is pre-testing by another name.

LLM-based, downstream workflows still require rigorous pre-testing of prompts, just as dictionary-based workflows require validation of terms and software compatibility. Prompt engineering has already emerged as one of the most actively discussed generative AI use cases in business practice (Chan & Choi, 2025). Researchers' construct choices are often politically embedded, highlighting the importance of acknowledging the changing neo-classical theorization of VBM (Wobst et al., 2025), or being transparent about a study's chosen framing / paradigms in ethics research (Dzhengiz & Hockerts, 2022). Researchers will not (be able to) outsource the accountability for this embedded meaning to, only seemingly, neutral, data-driven tools. This is particularly true in management fields like governance at large, ethics, and leadership, where interpretability, transparency, and replicability are essential (Martin, 2019; Mittelstadt et al., 2016). Thereby, our study provides not just a tool for today's researchers but delivers a methodologically sound stepping stone for LLM-enhanced research, especially in areas where full automation is not yet ethically or empirically viable.

9. Proposition 3: Prompt engineering will inherit the validation challenge from dictionary-based methods

Validation goes beyond internal consistency.

Our study validates the VBM dictionary only through internal consistency and convergent validity. However, the same feature-based alignment logic – linking construct characteristics to software functionalities – can also support predictive validation (e.g., linking VBM scores to ESG outcomes, audit quality, or firm value) and discriminant validation (e.g., distinguishing VBM from adjacent constructs). Our approach provides a systematic entry point for evaluating whether textual measures reflect the theoretical properties they are intended to capture. This includes extensions to domain-specific sentiment analysis, where researchers move beyond generic polarity and aim to capture nuanced affective constructs (e.g., regulatory optimism or ethical concern). It also applies across textual sources, as differences in structure, spontaneity, and authorship—such as between annual reports and earnings calls—may influence the choice and validation of software. This makes it applicable to software packages beyond the ones selected in this study.

10. Proposition 4: Systematic software selection guidelines offer a foundation for multiple forms of construct validation

Algorithm error persists.

Our study should not be seen as an endpoint, but as a bridge between dictionary-based CATA and the next generation of prompt-based, AI-assisted analysis. Even in traditional settings, algorithmic errors arise at multiple stages—for example, during preprocessing. Stemming algorithms differ in how they reduce word forms, and the choice of algorithm can introduce non-trivial distortions in term counts and construct representation (Hull, 1996). These errors are often opaque and dataset-dependent, underscoring the need for validation and tool awareness beyond surface-level features.

While LLMs may reduce some downstream limitations (e.g., phrase recognition), they introduce new risks: hallucinations, prompt sensitivity, data leakage, synthetic coherence, overfitting, domain drift, or output instability (Nguyen et al., 2022; Novy-Marx & Velikov, 2025). Our framework, which links construct features with the evaluation of software tools, offers a transparent and replicable process that future LLM workflows will also require (similar: Mariani & Dwivedi, 2024; Nguyen et al., 2022). It can inform *how* scholars choose LLM tools, *which* prompts are worth trusting, and *how* to verify conceptual alignment between input and output.

11. Proposition 5: Algorithmic errors will not disappear with LLMs – They will just look different

11.1. Conclusion

In summary, this study compares different software packages and develops guidelines for selecting suitable software. The guidelines suggest a structured and transparent software selection process. We hope the study will encourage scholars to choose their software in line with theory and to report their software choice transparently.

Ethical approval and informed consent statements: not applicable

Data availability statement: Data is proprietary.

Funding

Not applicable.

CRediT authorship contribution statement

Janice Wobst: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rainer Lueg:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express our gratitude to Sebastian Firk, the editor Mariano L.M. Heyden, Henning Schröder, as well as the two anonymous reviewers for their valuable feedback. We appreciate the feedback from participants at the Academy of Management Annual Meeting (Seattle, Washington, USA) 2022, The European Academy of Management EURAM Conference (Winterthur/Zurich, Switzerland) 2022, and at ACMAR Annual Conference for Management Accounting Research (Vallendar, Germany), 2023. We are very grateful to Sebastian Firk and his colleagues, who shared their proprietary data and knowledge with us for validation purposes.

Data availability

The data that has been used is confidential.

References

Anglin, A. H., Wolfe, M. T., Short, J. C., McKenny, A. F., & Pidduck, R. J. (2018). Narcissistic rhetoric and crowdfunding performance: A social role theory perspective. *J. Bus. Ventur., 33*(6), 780–812. https://doi.org/10.1016/j.jbusvent.2018.04.004

Babbie, E. (2020). *The Practice of Social Research* (15th ed.). Boston, MA: Cengage Learning.

Bisbe, J., Batista-Foguet, J.-M., & Chenhall, R. (2007). Defining management accounting constructs: A methodological note on the risks of conceptual misspecification. *Acc. Organ. Soc., 32*(7–8), 789–820.

Bligh, M. C., & Hess, G. D. (2007). The power of leading subtly: Alan Greenspan, rhetorical leadership, and monetary policy. *Leadersh. Q., 18*(2), 87–104.

Bligh, M. C., Kohles, J. C., & Meindl, J. R. (2004). Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks. *Leadersh. Q., 15*(2), 211–239.

Bligh, M. C., & Robinson, J. L. (2010). Was Gandhi "charismatic"? Exploring the rhetorical leadership of Mahatma Gandhi. *Leadersh. Q., 21*(5), 844–855.

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? *Contemp. Account. Res., 40*(2), 765–805. https://doi.org/10.1111/1911-3846.12825

Boyd, R., Ashokkumar, A., Seraj, S., & Pennebaker, J. (2022). *The Development and Psychometric Properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Burkert, M., & Lueg, R. (2013). Differences in the Sophistication of Value-based Management - the Role of top executives. *Manag. Account. Res., 24*(1), 3–22. https://doi.org/10.1016/j.mar.2012.10.001

Campbell. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull., 56*(2), 81–105.

Chan, H.-L., & Choi, T.-M. (2025). Using generative artificial intelligence (GenAI) in marketing: Development and practices. *J. Bus. Res., 191*, Article 115276.

Chan, Y., Hogan, K., Schwaiger, K., & Ang, A. (2020). ESG in factors. *The Journal of Impact and ESG Investing, 1*(1), 26–45.

Davis, K. M., & Gardner, W. L. (2012). Charisma under crisis revisited: Presidential leadership, perceived leader effectiveness, and contextual influences. *Leadersh. Q., 23*(5), 918–933.

Duriau, V. J., & Reger, R. (2004). *Choice of Text Analysis Software in Organization Research: Insight from a Multi-dimensional Scaling (MDS) analysis. Paper presented at the Le poids des mots: Actes de la 7e édition de Journées internationales d'Analyse statistique des Données Textuelles*.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organ. Res. Methods, 10*(1), 5–34. https://doi.org/10.1177/1094428106289252

Dzhengiz, T., & Hockerts, K. (2022). Dogmatic, instrumental and paradoxical frames: A pragmatic research framework for studying organizational sustainability. *Int. J. Manag. Rev., 24*(4), 501–534.

Evert, R. E., Payne, G. T., Moore, C. B., & McLeod, M. S. (2018). Top management team characteristics and organizational virtue orientation: An empirical examination of IPO firms. *Bus. Ethics Q., 28*(4), 427–461. https://doi.org/10.1017/beq.2018.3

Firk, S., Richter, S., & Wolff, M. (2021). Does value-based management facilitate managerial decision-making? an analysis of divestiture decisions. *Manag. Account. Res., 51*(2), Article 100736.

Firk, S., Schmidt, T., & Wolff, M. (2019). Exploring Value-based Management sophistication: The role of potential economic benefits and institutional influence. *Contemp. Account. Res., 36*(1), 418–450. https://doi.org/10.1111/1911-3846.12402

Fiss, P. C., & Zajac, E. J. (2004). The diffusion of ideas over contested terrain: The (non) adoption of a shareholder value orientation among German firms. *Adm. Sci. Q., 49*(4), 501–534.

Gamache, D. L., & McNamara, G. (2019). Responding to bad press: How CEO temporal focus influences the sensitivity to negative media coverage of acquisitions. *Acad. Manag. J., 62*(3), 918–943.

Gaur, A., & Kumar, M. (2018). A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research. *J. World Bus., 53*(2), 280–289.

Gevorkyan, M. N., Demidova, A. V., Demidova, T. S., & Sobolev, A. A. (2019). Review and comparative analysis of machine learning libraries for machine learning. *Discrete and Continuous Models and Applied Computational Science, 27*(4), 305–315. https://doi.org/10.22363/2658-4670-2019-27-4-305-315

Godfrey, P. C., & Hill, C. W. L. (1995). The problem of unobservables in strategic management research. *Strateg. Manag. J., 16*(7), 519–533. https://doi.org/10.1002/smj.4250160703

Guo, W., Sengul, M., & Yu, T. (2021). The impact of executive verbal communication on the convergence of investors' opinions. *Acad. Manag. J., 64*(6), 1763–1792. https://doi.org/10.5465/amj.2019.0711

Hart, R. P. (2014). DICTION 7: Help Manual Digitext, Inc., Version 2/26/2014.

Herhausen, D., Ludwig, S., Abedin, E., Haque, N. U., & de Jong, D. (2025). From words to Insights: Text analysis in business research. *J. Bus. Res., 198*, Article 115491.

Hermann, E., & Puntoni, S. (2024). Artificial intelligence and consumer behavior: From predictive to generative AI. *J. Bus. Res., 180*, Article 114720.

Heyden, M. L., Oehmichen, J., Nichting, S., & Volberda, H. W. (2015). Board background heterogeneity and exploration-exploitation: The role of the institutionally adopted board model. *Glob. Strateg. J., 5*(2), 154–176.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text preprocessing for text mining in organizational research: Review and recommendations. *Organ. Res. Methods, 25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Hubbard, T. D., Pollock, T. G., Pfarrer, M. D., & Rindova, V. P. (2018). Safe bets or hot hands? how status and celebrity influence strategic alliance formations by newly public firms. *Acad. Manag. J., 61*(5), 1976–1999. https://doi.org/10.5465/amj.2016.0438

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *J. Am. Soc. Inf. Sci., 47*(1), 70–84. https://doi.org/10.1002/(SICI)1097-4571(199601)47:1<70::AID-ASI7>3.0.CO;2-#

Ittner, C. D., & Larcker, D. F. (2001). Assessing empirical research in managerial accounting: A Value-based Management perspective. *Journal of Accounting & Economics, 32*(1–3), 349–410.

Klein, H. (2014). Text analysis software: Classified. Retrieved from http://textanalysis.info/pages/text-analysis-software-classified.php.

Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology* (4th ed.). Thousand Oaks, CA: Sage.

Li, K., Mai, F., Shen, R., & Yan, X. (2020). Measuring corporate culture using machine learning. *The Review of Financial Studies*. https://doi.org/10.1093/rfs/hhaa079

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Love, E. G., Lim, J., & Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal, 60*(4), 1462–1481.

Mahlendorf, M. D., Martin, M. A., & Smith, D. (2023). Innovative Data – Use-cases in Management Accounting Research and Practice. *European Accounting Review, 32*(2), 547–576. https://doi.org/10.1080/09638180.2023.2213258

Mansouri, S., & Momtaz, P. P. (2022). Financing sustainable entrepreneurship: ESG measurement, valuation, and performance. *Journal of Business Venturing, 37*(6), Article 106258. https://doi.org/10.1016/j.jbusvent.2022.106258

Mariani, M., & Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research, 175*, Article 114542.

Marshall, J. D., Yammarino, F. J., Parameswaran, S., & Cheong, M. (2022). Using CATA and machine learning to operationalize old constructs in new ways: An illustration using U.S. governors' COVID-19 press briefings. *Organizational Research Methods, 26*(4), 705–750. https://doi.org/10.1177/10944281221098607

Martin, J. D., Petty, J. W., & Wallace, J. (2009). Shareholder value maximization — is there a role for Corporate Social Responsibility? *Journal of Applied Corporate Finance, 21*(2), 110–118. https://doi.org/10.1111/j.1745-6622.2009.00232.x

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850.

Mavropulo, O., Rapp, M. S., & Udoieva, I. A. (2021). Value-based management control systems and the dynamics of working capital: Empirical evidence. *Management Accounting Research, 52*, Article 100740.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management, 44*(7), 2909–2933. https://doi.org/10.1177/0149206316657594

McKenny, A. F., & Short, J. C. (2012). CAT Scanner Manual. *Retrieved from*. http://www.amckenny.com/CATScanner.

McKenny, A. F., Short, J. C., & Newman, S. M. (2012). CAT Scanner (Version 1.0) [Software]. *Available from http://www.catscanner.net/*.

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods, 16*(1), 152–184. https://doi.org/10.1177/1094428112459910

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), Article 2053951716679679.

Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). *Market Research the Process, Data, and Methods using Stata* (1 ed.). Singapore: Springer.
Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management, 20*(4), 903–931. https://doi.org/10.1016/0149-2063(94)90035-3
Murphy, S. E., & Ensher, E. A. (2008). A qualitative analysis of charismatic leadership in creative teams: The case of television directors. *The Leadership Quarterly, 19*(3), 335–352. https://doi.org/10.1016/j.leaqua.2008.03.006
Murray, A., & Fisher, G. (2022). When more is less: Explaining the curse of too much capital for early-stage ventures. *Organization Science, 34*(1), 246–282. https://doi.org/10.1287/orsc.2021.1568
Nadkarni, S., Pan, L., & Chen, T. (2019). Only timeline will tell: Temporal framing of competitive announcements and rivals' responses. *Academy of Management Journal, 62*(1), 117–143.
Neuendorf, K. A. (2016). *The Content Analysis Guidebook*. Thousand Oaks: Sage Publications.
Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). Artificial intelligence in business: A literature review and research agenda. *Communications of the Association for Information Systems, 50*(1), 7.
Novy-Marx, R., & Velikov, M. Z. (2025). AI-powered (finance) scholarship: National Bureau of Economic Research.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.
Pandey, S., & Pandey, S. K. (2019). Applying Natural Language Processing Capabilities in Computerized Textual Analysis to measure Organizational Culture. *Organizational Research Methods, 22*(3), 765–797. https://doi.org/10.1177/1094428117745648
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015 [LIWC manual]*. Austin, TX: LIWC.net.
Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion, 10*(6), 601–626. https://doi.org/10.1080/026999396380079
Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal, 53*(5), 1131–1152.
Qiu, F., Hu, N., Liang, P., & Dow, K. (2023). Measuring management accounting practices using textual analysis. *Management Accounting Research, 58*, Article 100818.
Rezazadeh, A., Kohns, M., Bohnsack, R., António, N., & Rita, P. (2025). Generative AI for growth hacking: How startups use generative AI in their growth strategies. *Journal of Business Research, 192*, Article 115320.
Rhee, E. Y., & Fiss, P. C. (2014). Framing Controversial Actions: Regulatory Focus, source credibility, and Stock Market Reaction to Poison Pill Adoption. *Academy of Management Journal, 57*(6), 1734–1758. https://doi.org/10.5465/amj.2012.0686
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.
Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior, 5*(1), 415–435. https://doi.org/10.1146/annurev-orgpsych-032117-104622
Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727–752. https://doi.org/10.1177/1094428107304534
Vaupel, M., Bendig, D., Fischer-Kreer, D., & Brettel, M. (2022). The role of share repurchases for firms' social and environmental sustainability. *Journal of Business Ethics, 183*(2), 401–428. https://doi.org/10.1007/s10551-022-05076-3
Wobst, J., Röttger, P., & Lueg, R. (2022). Textual analysis and complex construct development in management accounting: The case of Value-based Management sophistication. Working paper, Leuphana University Lüneburg: Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335126.
Wobst, J., Tanikulova, P., & Lueg, R. (2025). Value-based management: A review of its conceptualizations and a research agenda toward sustainable governance. *Journal of accounting literature, 47*(1), 150–200. https://doi.org/10.1108/JAL-11-2022-0123
Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. *Journal of Management, 46*(7), 1257–1274. https://doi.org/10.1177/0149206319862027

Janice Wobst is a researcher specializing in managerial accounting, with a particular focus on value-based management and sustainable governance. She earned her doctorate summa cum laude from Leuphana University of Lüneburg in 2024, where she also served as a Research Associate from December 2020 to May 2024.

Rainer Lueg is a Full Professor of Finance and Accounting. Previously, he worked as a consultant for McKinsey & Co. His current research revolves around strategic performance management systems with a specific focus on value-orientation and sustainability management.