

The Contemporary Tax Journal

Volume 7
Issue 1 *The Contemporary Tax Journal Volume
7, No. 1 – Winter 2018*

Article 4

2-1-2018

Latent Semantic Analysis: A Big Data Opportunity for Tax Research

Paul D. Hutchison Ph.D.
University of North Texas, paul.hutchison@unt.edu

C. Elizabeth Plummer Ph.D., CPA
Texas Christian University, c.e.plummer@tcu.edu

Benjamin George Ph.D.
University of South Dakota, benjamin.george@usd.edu

Follow this and additional works at: <https://scholarworks.sjsu.edu/sjsumstjournal>



Part of the [Taxation Commons](#), [Taxation-Federal Commons](#), [Taxation-Transnational Commons](#), and the [Tax Law Commons](#)

Recommended Citation

Hutchison, Paul D. Ph.D.; Plummer, C. Elizabeth Ph.D., CPA; and George, Benjamin Ph.D. (2018) "Latent Semantic Analysis: A Big Data Opportunity for Tax Research," *The Contemporary Tax Journal*: Vol. 7 : Iss. 1 , Article 4.
<https://doi.org/10.31979/2381-3679.2018.070104> <https://scholarworks.sjsu.edu/sjsumstjournal/vol7/iss1/4>

This Article is brought to you for free and open access by the Lucas Graduate School of Business at SJSU ScholarWorks. It has been accepted for inclusion in *The Contemporary Tax Journal* by an authorized editor of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Latent Semantic Analysis: A Big Data Opportunity for Tax Research

by

Paul D. Hutchison, Ph.D.

Associate Professor
University of North Texas
Department of Accounting

C. Elizabeth Plummer, Ph.D., CPA

Professor
Texas Christian University
Neeley School of Business
Department of Accounting

Benjamin George, Ph.D.

Assistant Professor
University of South Dakota
Beacom School of Business
Department of Decision Sciences

ABSTRACT

The underlying purpose of this article is to discuss latent semantic analysis (LSA) and its application and research opportunities in tax research. LSA is a Big Data research methodology that can assist researchers in identifying the underlying themes/topics in data. To illustrate, the authors apply LSA to almost 2,300 abstracts of articles in *Advances in Taxation*, *The Journal of the American Taxation Association*, and *National Tax Journal* to identify the major topics in academic tax research over the period 1979-2015 and examine differences across journals and across time. The authors also discuss how researchers can apply LSA methodology to specific tax and accounting topics, providing research opportunities that heretofore may have previously been impractical.

INTRODUCTION

Over the past several decades, more and more data from company websites, electronic databases, court cases, regulatory filings, accounting standard-setting processes, blogs, and even tweets are made readily available for public perusal and research analyses. Current computer technology has evolved to the point that data manipulation and analysis can be easily accomplished using created computer programs or packaged software programs (e.g., Tableau). Due to the magnitude of Big Data available and the potential research questions that can be addressed using these resources, researchers from various academic fields have begun applying Big Data methodologies to determine meaning and understanding from patterns or signals in available data.

The purpose of the current article is to discuss a specific research methodology, latent semantic analysis (LSA), and its application and research opportunities within tax research for Big Data applications. To illustrate LSA's ability to analyze large amounts of textual data and extract patterns and themes, this study applies LSA to the abstracts of 2,293 articles from three tax journals over the time period of 1979 through 2015.¹ The primary 12 tax topical areas are identified, which are addressed by 84.4% of these articles. These tax topics are examined for

¹ In LSA research, the terms "topics" and "themes" are used interchangeably.

differences both across journals and across time. The study then discusses a number of research opportunities for LSA application in tax and accounting research.

The remainder of this article is organized as follows. The next section discusses prior research that analyzes unstructured (textual) data, and describes LSA in particular. Then, results from applying LSA to tax journal abstracts are presented. Finally, the article concludes with suggestions for future tax research opportunities.

LITERATURE REVIEW

Data content can be disseminated in many forms: words, phrases, sentences, or narratives; numbers; and visuals, yet its meaning is left for the reader to interpret and understand based upon their own knowledge and contextual reference points. Over the years, content analysis is one of many approaches that has been used to examine data. Abbott & Monsen (1979, p. 504) define this form of analysis as:

a technique for gathering data that consists of codifying qualitative information in anecdotal and literary form into categories in order to derive quantitative scales of varying levels of complexity.

With content analysis, a principal characteristic of the data is that they are objective, systematic, and reliable (Krippendorff, 1980). The goal of content analysis, as applied to textual data, is to analyze message content by systematically enumerating, coding, and classifying words and phrases (McConnell, Haslem, & Gibson, 1986). This research approach can have varying levels of complexity from a simple counting methodology (Hutchison, White, & Daigle, 2004; Hutchison & White 2004; Hutchison & White, 2003) to more sophisticated statistical approaches. It can also be used to determine meaning/understanding from the amount of content or actual space occupied on a printed page (Gray, Kouhy, & Lavers, 1995).

Qualitative content analysis is one method used to analyze textual data, which focuses on language as communication with attention to the content and contextual meaning of the text (Hsieh & Shannon, 2005). It seeks to go beyond merely counting words, to closely examining language for the purpose of classifying large amounts of text into categories which represent similar meanings, and can represent either explicit or inferred communication (Weber, 1990).

Financial and accounting researchers have utilized qualitative content analysis methodology with accounting data to ascertain associations between narrative content and firm performance, valuation, or stock returns. For example, this approach has been applied to the president's letter in the annual report (McConnell et al., 1986; Kohut & Segars, 1992), earnings announcement press releases (Hoskin, Hughes, & Ricks, 1986; Frances, Schipper, & Vincent, 2002), and MD&A (Tennyson, Ingram, & Dugan, 1990; Callahan & Smith, 2004). Results based upon numerous research studies suggest that narrative disclosures do have information relevance.

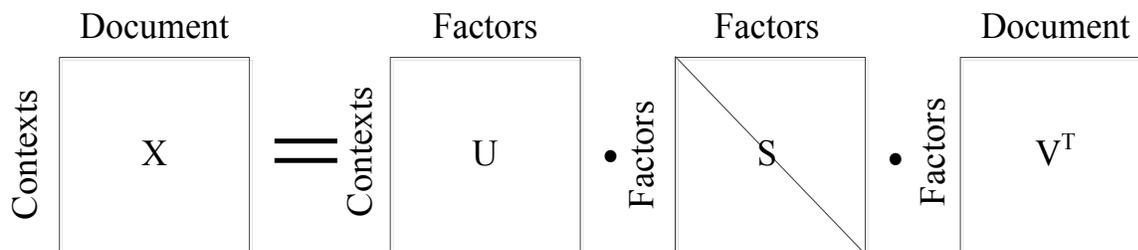
Latent Semantic Analysis

Within the area of content analysis, semantic analysis seeks to examine the linguistic or semantic elements of narrative data by relating syntactic structures. This approach reviews the words and/or linguistic structure of the disclosure narratives rather than the subject matter of the data. Specifically, it relates "syntactic structures, from the levels of phrases, clauses, sentences, and paragraphs to the level of the writing as a whole, to their language-independent meanings" (Goddard, 2013).

Using semantic analysis, Henry (2005) examined earnings press releases using linguistic tone and other stylistic attributes and found that they affect the market reaction to earnings

announcements. Extending Henry (2005), Davis, Piger, & Sedor (2007) further examined earnings announcements and determined that managers use optimistic and pessimistic linguistic tone to provide useful information to the reader. Using a semantic approach, Yuthas, Rogers, & Dillard (2002) examined management disclosures in the annual report (both the President's Letter and MD&A) and determined managers communicated something about their own credibility to investors along with information about the firm. Cho, Roberts, & Patten (2010) applied semantic analysis to environmental reports, and results suggest that language and verbal tone are used to manage stakeholder impressions.

Currently, researchers are seeking to extend semantic analysis research by examining words and their context for underlying meaning by using LSA. From a research perspective, LSA can be viewed as a statistical model of word usage that allows the researcher to draw comparisons of semantic similarity between pieces of textual information (Foltz, 1996). This methodology assumes "there is some underlying, or 'latent' structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate and define this latent structure" (Foltz, 1996). Specifically, LSA initially generates a matrix of occurrences of each word in each document under study and then uses singular-value decomposition (SVD), a technique closely related to eigenvector decomposition and factor analysis (see Figure 1). A key advantage of LSA methodology is that it can match two pieces of textual information even if they have no words in common (Foltz, 1996). (A detailed discussion of LSA can be examined in the Appendix of this article.)

Figure 1: Singular Value Decomposition (SVD)²

LSA was pioneered by researchers in psychology, information retrieval, and bibliometrics and has only recently been applied in the business discipline. One study used LSA to identify the intellectual core of the information systems discipline and identified five core research areas (Sidorova, Evangelopoulos, Valacich, & Ramakrishnan, 2008). Similarly, LSA has also been used to identify the intellectual core of real estate research (Winson-Geideman & Evangelopoulos, 2013b) and operations management research (Kulkarni, Apte, & Evangelopoulos, 2014), and to provide a comprehensive overview of the business processes literature and suggest directions for future research (Sidorova & Isik, 2010). Based upon article citations, LSA has also been utilized to chart the intellectual body of scholarly works based upon a specific author (F.W. Taylor) and textbook (*The Principles of Scientific Management*) (Evangelopoulos, 2011).

Quantitative Techniques and Journal Analysis

Recent research studies have used quantitative techniques to identify the top research themes across journals in accounting information systems (AIS) (Moffitt, Richardson, Snow, Weisner, & Wood, 2016) and real estate (Winson-Geideman & Evangelopoulos, 2013a). Moffitt et

² Adapted from Ashton, Evangelopoulos, and Prybutok 2014.

al. (2016) used text-mining techniques applied to AIS article abstracts for the period 1986-2014, to identify and examine the top research themes across three leading AIS journals (*Journal of Information Systems*, *International Journal of Accounting Information Systems*, and *Journal of Emerging Technologies in Accounting*) over this time period. Their text-mining techniques included part-of-speech tagging, pattern matching, term weighting, and search engine querying. Winson-Geideman & Evangelopoulos (2013a) used LSA applied to article abstracts from three leading real estate journals (*Real Estate Economics*, *Journal of Real Estate Finance and Economics*, and *Journal of Real Estate Research*) to determine and compare each journal's intellectual contribution to the real estate discipline.

LSA APPLIED TO TAX JOURNAL ABSTRACTS

Descriptive Statistics

LSA was selected as an appropriate method for the current research study's analysis to illustrate its application as a Big Data research methodology. This study applies LSA to article abstracts from three academic tax journals (*Advances in Taxation (AIT)*, *The Journal of the American Taxation Association (JATA)*, and *National Tax Journal (NTJ)*) over the period 1979-2015, and then uses the results to provide insights into the most significant topical areas during the period examined. LSA provides an efficient method for content analysis that combines rigorous statistical techniques and scholarly judgement to categorize a large amount of unstructured, textual data (Kulkarni, Apte, & Evangelopoulos, 2014).

The current study's data are abstracts of tax research articles, as the abstract can be considered the best representation of the articles' research topic (Winson-Geideman &

Evangelopoulos, 2013b). To allow comparisons among journals, the authors chose June 1979, the release date of the first edition of JATA, as the start date for this study, and December 2015 as the end date. This research utilized article abstracts from AIT for the period 1987 to 2015; JATA for the period 1979 to 2015; and NTJ for the period 1979 to 2015. Accordingly, the abstracts cover a 37-year period, although AIT was only published during 29 of the 37 years examined. Abstracts were obtained electronically for JATA and NTJ from the EBSCO Host Business Source Complete database and hand-collected from published editions for AIT and then digitized.

The quantities of abstracts in each tax journal vary due to the number of volumes, issues, and articles published annually (see Table 1, Panel A). AIT typically publishes one edition per year. However, during the years under study, there were eight years in which no editions were published.³ JATA published two editions per year during the study period, yet there were also supplements published in 11 years.⁴ NTJ regularly published four editions per year during the period examined.

³ AIT did not publish editions in 1988, 1991, 1995, 2005, 2007, 2009, 2011, and 2013.

⁴ JATA Supplements were issued for the years 1995 through 2005.

Table 1

Panel A
Journals, Volumes, and Issues Examined
1979-2015

| Journal | Volumes | | Issues | |
|---|----------------|----------------|---------------|----------------|
| | Number | Percent | Number | Percent |
| <i>Advances in Taxation</i> | 22 | 23.0% | 22 | 8.7% |
| <i>The Journal of the American Taxation Association</i> | 37 | 38.5% | 84 | 33.2% |
| <i>National Tax Journal</i> | 37 | 38.5% | 147 | 58.1% |
| TOTAL | 96 | 100.0% | 253 | 100.0% |

Panel B
Journal Abstracts⁵
1979-2015

| Journal | Initial Sample | Deleted Abstracts | Final Sample | Final percent |
|---|-----------------------|--------------------------|---------------------|----------------------|
| <i>Advances in Taxation</i> | 188 | 1 | 187 | 8.2% |
| <i>The Journal of the American Taxation Association</i> | 1,149 | 695 | 454 | 19.8% |
| <i>National Tax Journal</i> | 2,007 | 355 | 1,652 | 72.0% |
| TOTAL | 3,344 | 1,051 | 2,293 | 100.0% |

Table 1, Panel A provides a summary of the volumes and issues examined. Overall, there were 96 total volumes published from 1979 through 2015: 22 for AIT (23.0%), 37 for JATA (38.5%), and 37 for NTJ (38.5%). There were also 253 issues related to these journal volumes: 22 for AIT (8.7%), 84 for JATA (33.2%) and 147 for NTJ (58.1%). Thus, NTJ published more total issues than AIT and JATA combined.

Panel B of Table 1 provides information on the sample selection. For all three tax journals

⁵ Abstracts for book reviews, committee reports, doctoral research summaries, tax software reviews, editor reports, awards, discussions, replies, comments, remarks, corrections, introductions, and article summaries were not included in this study. Each abstract used in this study represents one article.

combined, there were 3,344 article abstracts initially collected: 188 for AIT (5.6%); 1,149 for JATA (34.4%); and 2,007 (60.0%) for NTJ. Because this study's goal is to examine tax topics in academic articles, the authors chose to limit the journal abstracts to those that focused primarily upon academic research. Accordingly, certain abstracts were deleted and removed from the sample. These included book reviews, committee reports, doctoral research summaries, tax software reviews, editor reports, awards, discussions, replies, comments, remarks, corrections, introductions, and article summaries. Deletions were 1 for AIT (0.1%), 695 for JATA (66.1%), and 355 for NTJ (33.8%). Please note that JATA and NTJ typically publish summaries of papers in the journal's issue, and JATA publishes a larger variety of items other than academic articles when compared to AIT and NTJ. The final sample for this study includes 2,293 abstracts: 187 for AIT (8.2%); 454 for JATA (19.8%); and 1,652 for NTJ (72.0%).

Results and Analysis

For this study, the authors applied LSA to the full sample of abstracts to identify the major topics in academic tax research over the past 37 years (1979 to 2015). SAS Enterprise Miner 14.2 and SAS Text Miner 14.2 were used to segregate text topics through a series of steps. Initially, the data file containing the full-length text abstracts were imported into SAS Enterprise Miner 14.2 from an Excel spreadsheet and then were converted from ASCII text format into a SAS table and SAS data file. This SAS dataset was then used for the natural language processing and textual analysis with SAS Text Miner 14.2 to generate the subsequent results (see Figure 2).

The horizontal axis of Figure 2 corresponds to the number of tax research topics extracted using the LSA methodology, while the vertical axis indicates the percentage of abstracts represented by those topics. Figure 2 shows that LSA identified nine tax research topics that are

addressed by 78.89% of the journal abstracts, while 22 topics are addressed in 93.94% of the abstracts. For tractability, the authors chose to limit the current study's analysis to the 12 strongest tax research topics identified, since these primary topics are addressed by 84.39% of the 2,293 journal abstracts examined.

Figure 2

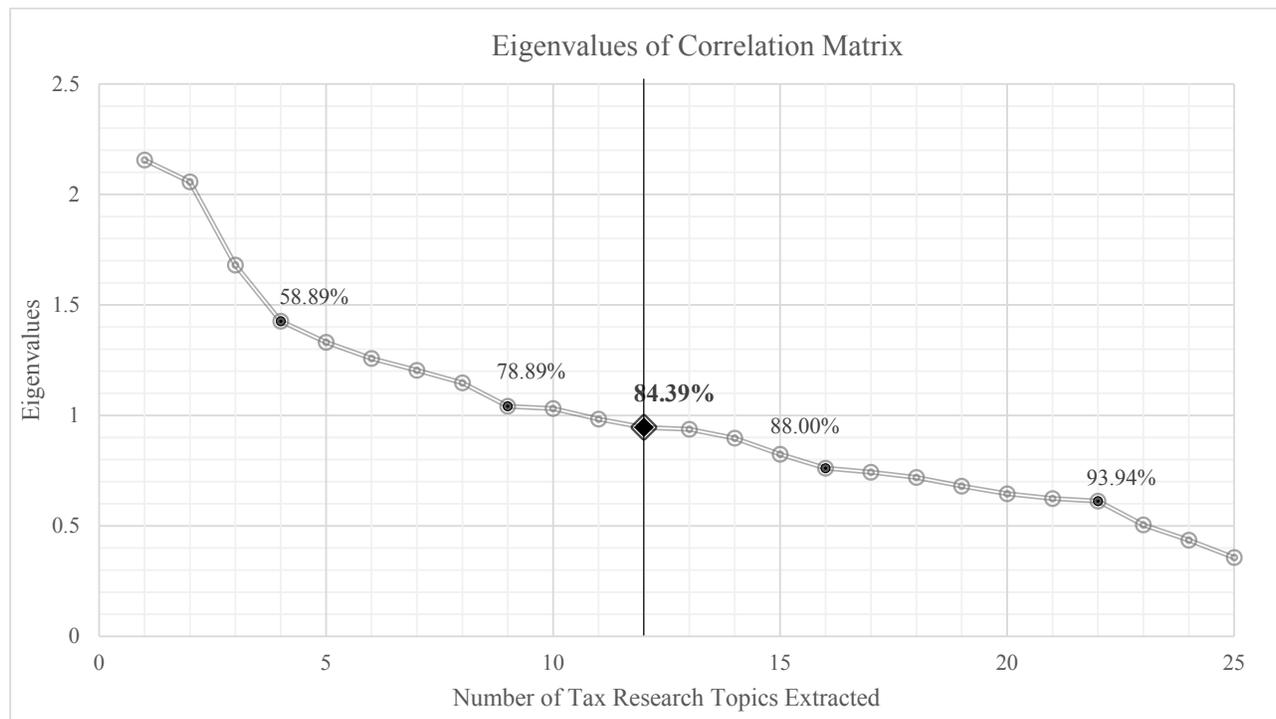


Table 2 presents a list of the strongest 12 Tax Research Topics identified from analyzing the full sample of abstracts.⁶ Topics are ranked in order of the number of published papers related to a topic. It is important to remember that the number of published papers does not necessarily indicate the relative importance of that topic to a discipline. Articles are, in part, a function of data availability and preferences of journal editors, as well as other factors

⁶ For each set of Terms identified, two of the authors independently determined a Tax Topic. Then they reconciled their differences after multiple passes to determine an agreed upon Tax Topic for each set of Terms.

(Winson-Geideman & Evangelopoulos, 2013a). In addition, this research examines abstracts, which do not capture article length, complexity, or incremental contribution to the literature (e.g., a paper's innovativeness). Nevertheless, this study's focus is to identify the major topics in academic tax research over the time period studied, and the number of published articles related to a tax topic provides a reasonable proxy of research effort and focus.

Table 2 indicates that there are 3,286 articles associated with the 12 Tax Research Topics identified. The LSA heuristic used in this study allows each article abstract to be associated with more than one topic, which results in more total documents associated with the 12 topics (N=3,286) than the source number of documents (N=2,293). Approximately half of the paper abstracts (49.2%) loaded on only one topic, while 35.8% loaded on two topics. The remaining 15% loaded on more than two topics. The most common reason for a two-factor loading was that a paper was associated with "Tax policy and tax reform," and also one of the other topics. Examination of these articles shows a paper that discusses "Tax policy and reform" in the context of (for example) the "Taxation of capital gains." There was also significant two-factor loading for "State and local taxes" and "Property taxes," which shared about one-third of the papers in common.

Table 2
Tax Research Topics: 12-Factor Solution for Full Sample¹
1979-2015

| | Terms | Tax Research Topics | No. of Articles |
|----|---|--|------------------------|
| 1 | +policy,+system,+taxation,+reform,economic | Tax policy and tax reform | 359 |
| 2 | +elasticity,+estimate,+model,+price,+rate | Price elasticity | 318 |
| 3 | +firm,+foreign,earnings,corporate,+corporation | Corporations and foreign earnings | 317 |
| 4 | +bond,+asset,+interest,+deduction,+tax-exempt | Effect of taxes on investment and financing | 317 |
| 5 | +income,+credit,+family,+taxpayer,+system | Income taxation of individuals | 314 |
| 6 | +state,local,+government,federal,fiscal | State and local taxes | 299 |
| 7 | +taxpayer,+compliance,+professional,+client,+audit | Taxpayer compliance | 273 |
| 8 | capital gains, +stock,+investor,+capital gain tax,+rate | Taxation of capital gains | 261 |
| 9 | +property,+school,property tax,+property tax,local | Property taxes | 253 |
| 10 | +sale,+sale tax,+state,+revenue,+business | State taxation of sales revenues and business income | 229 |
| 11 | +security,+retirement,+plan,social,+pension | Retirement plans, including social security | 192 |
| 12 | +health,insurance,care,+cost,+price | Health insurance | 154 |
| | Total Articles | | 3,286 |

¹ Ranked by number of articles.

Table 2 shows that “Tax policy and tax reform” is associated with the greatest number of articles (N=359 articles). This means that 15.7% of the 2,293 articles published in the three tax journals examined during 1979-2015 focused on tax policy and tax reform. This is not surprising for two reasons. First, as discussed above, tax policy and reform provides a broad research area which encompasses more specific topics. Second, tax legislation (both proposed and enacted) provides researchers with interesting questions and opportunities for impactful research. The next five topics identified in Table 2 are each associated with approximately 300 or more academic tax articles, which means that each topic is associated with more than 13% of the articles published in the three journals during 1979-2015. These topics are (in order): “Price elasticity,” “Corporations and foreign earnings,” “Effect of taxes on investment and financing,” “Income taxation of individuals,” and “State and local taxes.”

Table 3 presents a list of the top 12 Tax Research Topics, but ranks each of the topics according to the number of published articles within shorter time periods (i.e., 7 to 10 years). This table thus shows how the relative importance of the topics has changed over time. The trend is consistent with proposed and enacted tax legislation during this period. The earliest time period (1979-1985) shows that state and local tax research topics were dominant, ranking first (State and local taxes) and third (Property taxes) in number of published articles. This reflects state-level reforms and legislation occurring prior to and during this period. For example, in 1978, California voters passed the highly-controversial Proposition 13 with a 64.8% voter majority, which initiated sweeping changes to the

California property tax system.¹ Although a single-state referendum, Proposition 13 drew national attention—Howard Jarvis (who co-sponsored the bill) made the cover of *Time* magazine after its passage. The “tax revolt” in California reflected taxpayer sentiments across the nation and was followed by tax relief initiatives in other states (e.g., Oregon, Washington, Colorado, and Massachusetts). Some people speculate that Proposition 13 was the prelude to the Reagan income tax cuts in 1981 (Moore, 1998).

¹ Among other changes, Proposition 13 generally caps tax rates at 1% of a property’s assessed value, and limits real property tax assessment increases to 2% per year unless a property is sold.

Table 3
Decade-wise Analysis of Tax Research Topics¹

| | 1979-1985 (7 years) | 1986-1995 (10 years) | 1996-2005 (10 years) | 2006-2015 (10 years) |
|----|--|--|--|--|
| 1 | State and local taxes | Tax policy and tax reform | Tax policy and tax reform | Corporations and foreign earnings |
| 2 | Effect of taxes on investment and financing | Effect of taxes on investment and financing | Income taxation of individuals | Price elasticity |
| 3 | Property taxes | Price elasticity | Corporations and foreign earnings | Tax policy and tax reform |
| 4 | Income taxation of individuals | Taxpayer compliance | Taxpayer compliance | Income taxation of individuals |
| 5 | Tax policy and tax reform | Taxation of capital gains | Effect of taxes on investment and financing | Taxpayer compliance |
| 6 | Price elasticity | Income taxation of individuals | Taxation of capital gains | Taxation of capital gains |
| 7 | State taxation of sales revenues and business income | State and local taxes | Price elasticity | State taxation of sales revenues and business income |
| 8 | Corporations and foreign earnings | Property taxes | Retirement plans, including social security | Property taxes |
| 9 | Retirement plans, including social security | State taxation of sales revenues and business income | State and local taxes | State and local taxes |
| 10 | Taxation of capital gains | Corporations and foreign earnings | State taxation of sales revenues and business income | Health insurance |
| 11 | Taxpayer compliance | Retirement plans, including social security | Property taxes | Effect of taxes on investment and financing |
| 12 | Health insurance | Health insurance | Health insurance | Retirement plans, including social security |

¹ Ranked by number of articles within the time period examined.

In each of the next two-time periods examined (1986-1995 and 1996-2005), “Tax policy and tax reform” was the dominant topic, with differences in the rankings of more specific topical areas. Most notably, “Income taxation of individuals” and “Corporations and foreign earnings” increased in their ranking during 1996-2005 period, while research related to state taxation dropped in the rankings. These relative changes in research focus reflect significant federal tax legislation occurring before and during this period, with the most notable being the Tax Reform Act of 1986 (TRA86). TRA86 made significant changes to the *Internal Revenue Code*, including lowering the top individual marginal tax rate from 50% to 28%.¹ Individual income tax rates were further changed by both increases (1990 and 1993) and decreases (2001 and 2003) in subsequent tax legislation.² Importantly, TRA86 also increased attention on the ability of multinational corporations to avoid or defer tax by transferring intangible property. Specifically, the Conference Committee Report on TRA86 recommended that the U.S. Treasury make a comprehensive study of the §482 regulations, with specific attention as to whether the regulations should be modified (these are commonly referred to as the transfer pricing regulations).³ The U.S. Treasury and Internal Revenue Service (IRS) released their study in 1988. Then again in 1990, Congress directed the IRS to examine recent legislation aimed at increasing §482 compliance and to recommend further modifications (OBRA90, Committee Report 4821: Study of IRC §482). The IRS released that report in 1992.⁴

¹ Other significant tax legislation in the 1980's included the Economic Recovery Tax Act of 1981, Tax Equity and Fiscal Responsibility Act of 1982, and the Deficit Reduction Act of 1984. See <http://www.taxpolicycenter.org/laws-proposals/> for a summary of major federal tax legislation from 1940 through the current year.

² Most notably, these include the Omnibus Budget Reconciliation Act (OBRA) of 1990, OBRA 1993, the Economic Growth and Tax Relief Reconciliation Act of 2001, and the Jobs and Growth Tax Relief Reconciliation Act of 2003.

³ IRC §482 authorizes the U.S. Treasury Secretary to allocate income and deductions among commonly controlled businesses “in order to prevent evasion of taxes or clearly to reflect the income” of the businesses, and the §482 regulations provide the applicable standards and methodologies for determining transfer prices.

⁴ See Green (1993) for a more detailed discussion.

As evidenced by the last column in Table 3, the topic of “Corporations and foreign earnings” continues to increase in importance. There have been more papers published during 2006-2015 on this topic than any other topic identified.⁵ It is also worth noting that the ranking of “Health insurance” increased slightly during the last decade. Given the passage of healthcare reform in 2010 (Patient Protection and Affordable Care Act), and the Trump administration’s efforts to repeal and/or replace significant portions of this act, health insurance will likely continue to increase in research importance.⁶

Because each of the tax journals used in the current study has a different audience, Table 4 presents a list of the top 12 Tax Research Topics separately for each journal, ranked according to the number of articles in each journal for the full time period (1979-2015). Differences among the rankings demonstrate the topical preferences of each journal. AIT’s most highly-ranked topic is “Taxpayer compliance,” followed by “Corporations and foreign earnings.” Articles related to state and local taxation are ranked in the bottom half of the 12 identified tax research topics. In contrast, JATA’s most highly-ranked topic is “Corporations and foreign earnings,” followed by several federal tax topics. Articles regarding “Price elasticity” are ranked sixth, while topics involving state and local tax articles are among the bottom four (9, 10, 11). NTJ has the most diverse audience, including “academic, private sector, and government economists, accountants and attorneys, as well as business and governmental tax practitioners,” and the broadest editorial policy—to publish “research on government tax and expenditure policies.”⁷ NTJ’s audience and editorial policy is evident from the article rankings using LSA. Tax research topics involving state

⁵ In 2004, Congress directed the U.S. Treasury to conduct three studies dealing with foreign-related earnings: the effectiveness of the §482 transfer pricing rules, with an emphasis on transactions involving intangible property; income tax treaties; and inversion transactions (e.g., American Jobs Creation Act of 2004, Committee Report 4821.099).

⁶ Specifically, the Patient Protection and Affordable Care Act of 2010 and Health Care and Education Reconciliation Act of 2010.

⁷ See NTJ’s editorial policy (NTJ 2016).

and local taxation are highly-ranked among the topics identified (1, 5, 6), as well as “Tax policy and tax reform” (2) and “Price elasticity” (3)—this last topic being a research area more commonly pursued by economists than accountants.

Table 4
Tax Research Topics by Journal⁸
1979-2015

| | <i>AIT</i> | <i>JATA</i> | <i>NTJ</i> |
|-----------------------|--|--|--|
| 1 | Taxpayer compliance | Corporations and foreign earnings | State and local taxes |
| 2 | Corporations and foreign earnings | Taxpayer compliance | Tax policy and tax reform |
| 3 | Effect of taxes on investment and financing | Effect of taxes on investment and financing | Price elasticity |
| 4 | Tax policy and tax reform | Taxation of capital gains | Income taxation of individuals |
| 5 | Income taxation of individuals | Income taxation of individuals | Property taxes |
| 6 | Taxation of capital gains | Price elasticity | State taxation of sales revenues and business income |
| 7 | Price elasticity | Tax policy and tax reform | Effect of taxes on investment and financing |
| 8 | Property taxes | Retirement plans, including social security | Retirement plans, including social security |
| 9 | Retirement plans, including social security | State taxation of sales revenues and business income | Taxation of capital gains |
| 10 | State taxation of sales revenues and business income | Property taxes | Health insurance |
| 11 | Health insurance | State and local taxes | Corporations and foreign earnings |
| 12 | State and local taxes | Health insurance | Taxpayer compliance |
| Total Articles | 302 (9.2%) | 683 (20.8%) | 2,301 (70.0%) |

Focusing upon the most recent time period in the study (2000-2015), Table 5 provides a ranking of the Tax Research Topics separately for each academic tax journal. The rankings for AIT and JATA are quite similar to their rankings for the full time period, while NTJ’s rankings show the

⁸ Ranked by number of articles in each journal.

most movement. Most notably, articles focused on “Corporations and foreign earnings” moved from a ranking of 11th (1979-2015) to 6th (2000-2015). The topics of “Tax policy and tax reform” remains a main focus of NTJ’s articles, and “Price elasticity” is ranked first in this time period.

Table 5

**Tax Research Topics⁹
2000-2015**

| | <i>AIT</i> | <i>JATA</i> | <i>NTJ</i> |
|----|--|--|--|
| 1 | Taxpayer compliance | Corporations and foreign earnings | Price elasticity |
| 2 | Corporations and foreign earnings | Taxpayer compliance | Tax policy and tax reform |
| 3 | Income taxation of individuals | Taxation of capital gains | Income taxation of individuals |
| 4 | Tax policy and tax reform | Effect of taxes on investment and financing | State and local taxes |
| 5 | Taxation of capital gains | Income taxation of individuals | State taxation of sales revenues and business income |
| 6 | Effect of taxes on investment and financing | Price elasticity | Corporations and foreign earnings |
| 7 | Price elasticity | Tax policy and tax reform | Retirement plans, including social security |
| 8 | Retirement plans, including social security | Retirement plans, including social security | Health insurance |
| 9 | State taxation of sales revenues and business income | State taxation of sales revenues and business income | Property taxes |
| 10 | Property taxes | Property taxes | Effect of taxes on investment and financing |
| 11 | Health insurance | State and local taxes | Taxation of capital gains |
| 12 | State and local taxes | Health insurance | Taxpayer compliance |

TAX RESEARCH OPPORTUNITIES AND CONCLUSIONS

As illustrated in this article, LSA is a powerful text analysis methodology that can be used in tax and accounting research, providing research opportunities that may have previously been deemed impractical. Essentially, LSA can be applied to any textual data—including verbal, print,

⁹ Ranked by number of articles in each journal.

or electronic, as well as textual data obtained from narrative responses, open-ended survey questions, interviews, and focus groups (Kondracki & Wellman, 2002). There are widely available technologies now to convert verbal data into textual data, making the application of LSA almost unlimited.

For tax researchers specifically, LSA can be applied to judicial decisions to determine (for example) differences and commonalities in reasoning used by different courts on the same issue, or by the same court across different issues. Transcripts of Congressional debates regarding tax legislation are now generally available and can be analyzed across political affiliation, geographic representation, or person-specific characteristics. Similar analysis could be contemplated with IRS rulings and notices, and Congressional committee reports. Further, most of these opportunities for textual analysis apply equally at the state level as well.¹⁰

For tax and accounting researchers more generally, LSA can be applied to firms' annual reports and SEC filings to examine (for example) the narrative in the firm's tax footnote, or transcripts of earnings conference calls to determine when and in what context tax issues are discussed. State and local governments also issue their own annual reports—a *Comprehensive Annual Financial Report* (CAFR), or a *Popular Annual Financial Report* (PAFR), to which LSA could be applied. CAFR's closely resemble a firm's annual report—including an MD&A section, auditor's report, financial statements, and detailed footnotes. As one would expect, taxation is a theme that permeates these CAFR's. Opportunities for analysis also applies to comment letters, which are available from both the Financial Accounting Standards Board (FASB) and Governmental Accounting Standards Board (GASB). These letters can be examined across entity type and issue, and it is worth noting that several significant statements have been issued recently that deal with

¹⁰ For example, transcripts from the Illinois General Assembly are available at: <http://www.ilga.gov/senate/transcripts/default.asp>.

tax issues. For example, GASB Statement No. 77 requires state and local governments to disclose in their CAFRs any tax abatements affecting their revenue-raising abilities, effective 2016 or 2017, depending on the government's fiscal year-end. The GASB received 298 comment letters related to this statement (Francis, 2015). Thus, both comment letters and CAFRs provide textual information for analysis.

This research study has sought to discuss LSA, provide an application of LSA in tax research by using article abstracts from three academic tax journals (*Advances in Taxation*, *The Journal of the American Taxation Association*, and *National Tax Journal*), discuss those results, and suggest research opportunities for LSA application in tax research. While a counting methodology can be used for textual analysis, LSA extends that methodology by providing theory with an underpinning of statistics to reach its results. And importantly, LSA allows data to be assigned to multiple topics/themes. Because LSA can be applied to voluminous amounts of unstructured data (Big Data), it provides significant and interesting research opportunities that may not previously have been practical.

REFERENCES

- Abbott, W. F., and R.J. Monsen. 1979. On the Measurement of Corporate Social Responsibility: Self-Reported Disclosures as a Method of Measuring Corporate Social Involvement, *Academy of Management Journal* 22 (3): 501-515.
- Ashton, T., N. E. Evangelopoulos, and V. Prybutok. 2014. Extending Monitoring Methods to Textual Data: A Research Agenda, *Quality & Quantity* 48 (4): 2277-2294.
- Callahan, C. M., and R. E. Smith. 2004. Firm Performance and Management's Discussion and Analysis Disclosures: An Industry Approach, working paper, University of Arkansas.
- Cho, C., R. W. Roberts, and D. Patten. 2010. The Language of U.S. Corporate Environmental Disclosure, *Accounting, Organizations, and Society* 11 (1): 1-16.
- Davis, A. K., J. Piger, and L. M. Sedor. 2007. Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases, AAA 2008 Financial Accounting and Reporting Section (FARS) paper.
- Deerwester, S., S. T. Dumais, and T. K. Landauer. 1990. Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science* 41 (6): 391-407.
- Evangelopoulos, N. E. 2011. Citing Taylor: Tracing Taylorism's Technical and Sociotechnical Duality Through Latent Semantic analysis, *Journal of Business and Management* 17 (1): 57-74.
- Foltz, P. W. 1996. Latent Semantic Analysis of Text-Based Research, *Behavior Research Methods, Instruments, & Computers* 28 (2): 197-202.
- Frances, J., K. Schipper, and L. Vincent. 2002. Expanded Disclosures and the Increased Usefulness of Earnings Announcements, *The Accounting Review* 77 (3): 515-546.
- Francis, N. 2015. GASB 77: Reporting Rules on Tax Abatements, *State and Local Finance Initiative Brief*, Urban Institute, downloaded on March 23, 2017 at: <http://www.urban.org/research/publication/gasb-77-reporting-rules-tax-abatements>.
- Goddard, C. 2013. *Semantic Analysis: An Introduction* (2nd ed.). New York, N.Y.: Oxford Press.
- Gray, R., R. Kouhy, and S. Lavers. 1995. Methodological Themes; Constructing a Research Database of Social and Environmental Reporting by UK companies, *Accounting, Auditing & Accountability Journal* 8 (2): 78-101.
- Green, R. A. 1993. The Future of Source-Based Taxation of the Income of Multinational Enterprises, *Cornell Law Review* 79 (18): 18-86.
- Henry, E. 2005. Are Investors Influenced by How Earnings Press Releases Are Written? *Journal of Business Communication* 45 (4): 363-407.
- Hoskin, R. E., J. S. Hughes, and W. E. Ricks. 1986. Evidence on the Incremental Information Content of Additional Firm Disclosures Made Concurrently With Earnings, *Journal of Accounting Research* 24 (3): 1-32.
- Hsieh, H-F, and S. E. Shannon. 2005. Three Approaches to Qualitative Content Analysis, *Qualitative Health Research* 15 (9): 1277-1288.
- Hutchison, P. D., and C. G. White. 2003. *The Journal of the American Taxation Association* 1979-2000: Content, Participation, and Citation Analyses, *The Journal of the American Taxation Association* 25 (1): 100-121.
- Hutchison, P. D., and C. G. White. 2004. Academic Tax Articles: Productivity and Participation Analyses 1980-2000, *Advances in Taxation* 16: 183-199.
- Hutchison, P. D., C. G. White, and Daigle, R. J. 2004. *Advances in Accounting Information Systems and International Journal of Accounting Information Systems: The First Ten Volumes (1992-*

- 2003), *International Journal of Accounting Information Systems* 5(3): 341-365.
- Kohut, G. F., and A. H. Segars. 1992. The President's Letter to Stockholders: An Examination of Corporate Communication Strategy, *Journal of Business Communication* 29 (1): 7-21.
- Jolliffe, I. T. 2002. *Principal Component Analysis* (2nd ed.). *Springer Series in Statistics*. New York, NY: Springer-Verilog New York.
- Kondracki, N. L. and N. S. Wellman. 2002. Content Analysis: Review of Methods and Their Applications in Nutrition Education, *Journal of Nutrition Education and Behavior* 34 (4): 224-230.
- Krippendorff, K. 1980. *Content Analysis: An Introduction to its Methodology*. New York, NY: Sage Publishing, Inc.
- Kulkarni, S. S., U. M. Apte, and N. E. Evangelopoulos. 2014. The Use of Latent Semantic Analysis in Operations Management Research, *Decision Sciences* 45 (1): 971-994.
- McConnell, D., J. A. Haslem, and V. R. Gibson. 1986. The President's Letter to Stockholders: A New Look, *Financial Analysts Journal* 42 (5): 66-70.
- Moffitt, K. C., V. J. Richardson, N. M. Snow, M. M. Weisner, and D. A. Wood. 2016. Perspectives on Past and Future AIS Research as the *Journal of Information Systems* Turns Thirty, *Journal of Information Systems* 30 (3): 157-171.
- Moore, S. 1998. Proposition 13: Then, Now and Forever, CATO Institute (July 30, 1998), downloaded on March 23, 2017 at: <https://www.cato.org/publications/commentary/proposition-13-then-now-forever>.
- Sidorova, A., and O. Isik. 2010. Business Process Research: A Cross-Disciplinary Review, *Business Process Management Journal* 16 (4): 566-597.
- Sidorova, A., N. E. Evangelopoulos, J. S. Valacich, and T. Ramakrishnan. 2008. Uncovering the Intellectual Core of the Information Systems Discipline, *MIS Quarterly* 32 (3): 467-482.
- Tennyson, M., R. Ingram, and M. Dugan. 1990. Assessing the Information Content of Narrative Disclosures in Explaining Bankruptcy, *Journal of Business Finance & Accounting* 17 (3): 391-410.
- Weber, R. P. 1990. *Basic Content Analysis*. Beverly Hills, CA: Sage.
- Winson-Geideman, K., and N. E. Evangelopoulos. 2013a. Research in Real Estate, 1973-2010: A Three-Journal Comparison, *Journal of Real Estate Literature* 21 (2): 255-267.
- Winson-Geideman, K., and N. E. Evangelopoulos. 2013b. Topics in Real Estate Research, 1973-2010: A Latent Semantic Analysis, *Journal of Real Estate Literature* 21 (1): 59-76.
- Yuthas, K.R., R. Rogers, and J. F. Dillard. 2002. Communicative Action and Corporate Annual Reports, *Journal of Business Ethics* 41 (1/2): 141-157.
- Zhu, M. and A. Ghodsi. 2006. Automatic Dimensionality Selection From the Scree Plot Via the Use of Profile Likelihood, *Computational Statistics and Data Analysis* 51 (2): 918-930.

Appendix

LATENT SEMANTIC ANALYSIS (LSA)

LSA is a natural language processing technique that extracts concepts from a matrix of terms to construct a pattern of words (Deerwester, Dumais, & Landauer, 1990; Kulkarni et al. 2014) and is bounded by the fundamental assumption that words that are similar in meaning will occur in similar sections of text. A collection of contexts, identified as either individual words or a collection of terms, is extracted from the data under investigation. Due to the central assumption of LSA, contexts with similar meaning will therefore occur in similar meaning documents. Through a process similar to traditional factor analysis, utilizing singular-value decomposition (SVD), LSA then establishes the common factors that represent the underlying concepts exhibited within the data and then are compiled in a term frequency matrix, which contains counts per terms.

A matrix containing counts of contexts per individual textual data segment (i.e., document, paragraph, or other designation of granularity) is constructed from the data. SVD is used to reduce the number of rows that represent contexts within the documents while preserving the similarity structure among columns representing each separate document (see Figure 1).

In an approach similar to principal component analysis, SVD generates simultaneous principal components for two sets of variables, the contexts (U) and the documents (V^T). These results produce two separate sets of factor loadings, one for each set of matrices, with each latent semantic factor associated with a set of corresponding high-loading terms and a set of high-loading documents. These can be interpreted to develop a fundamental word usage and association pattern, which are termed Factors or Themes. In a similar manner to traditional factor analysis, the researcher can dictate the number of factors to be extracted within LSA and therefore

determine the level of granularity topics/themes that are identified.

The resulting LSA analysis approximates the influence of a word on the meaning of passages and vice versa. The similarity of the effects the words have on the passage is the interpretation of the derived relation between individual words, not the frequency of words in a passage. LSA allows for a word vector to represent a mixture of senses in comparison to contextual usage and not disambiguity in passage meaning formation. Due to the similarities to traditional factor analysis, cross-loadings and thus overlapping of topics can occur, and should be expected due to the nature of the utilized data source, academic tax journal abstracts.

Dimension Reduction for Topic Extraction

Due to the complexity of high dimensional data, the determination of the specific number of topics to extract that will provide meaningful context to the compiled collection of documents is a complex exercise. The extant literature provides several methods for the investigation and reduction of dimensionality of component matrices. Many of the current techniques rely upon various facets of the dimension reduction technique, principal component analysis (PCA). The three primary methods in use are: *percent variance*, *scree plot analysis*, and *sequential testing* (Jolliffe, 2002). Each method is contingent upon the eigenvectors generated by the sample variance-covariance matrix of the data during PCA with the ordered eigenvalues determining the rank of each extracted context. The percent variance approach attempts to simplify the selection process to a heuristic that defines a cutoff value for which a predetermined percentage of documents within the corpus are included by the addition of each subsequent ranked context. In comparison, scree plot analysis is a visual interpretation of gaps or “elbows” within the plotted magnitudes of each eigenvector to determine natural breaks within the extracted topics. Finally, the sequential testing approach attempts to repeatedly test the difference between the ordered

eigenvalues in an attempt to locate statistically significant differences. Each of these three methods have their own benefits and shortcomings, which complicates the selection of an individual methodology.

The fundamental advantage of each of the three approaches is tied to the relative simplicity of analysis. As the approaches digress from a simple heuristic, to comparative visual analysis, to statistical tests, the ease in both application and interpretation declines. The primary detraction for both the percent variance approach and scree plot analysis is tied to the subjective nature of the determination of threshold or gaps. In addition, Zhu & Ghodsi (2006) highlight a problematic flaw concerning the validity of the sequential test approach as it relies on the assumption that the underlying data follow a multivariate normal distribution. Additionally Jolliffe expounds on this issue to surmise that "it's difficult to get even an approximate idea of the overall significance level because of the number of tests done is not fixed but random, and the tests are not independent of each other"(Jolliffe, 2002, sec. 6.1.4). Due to the variety of approaches available with no clear discernible predominant methodology and the heavy criticism for the parametric option, this study employed both an analysis of scree plots in conjunction with percentage of abstracts identified to determine the final number of topics extracted. Owing to the iterative nature of this methodology, this study includes the individual scree plot as well as the corresponding coverage values for reader perusal (see Figure 2).

Contact Information:

Paul D. Hutchison, Ph.D.
Associate Professor
University of North Texas
Department of Accounting
1155 Union Circle #305219
Denton, TX 76203-5017
(940) 565-3083
email: paul.hutchison@unt.edu

C. Elizabeth Plummer, Ph.D., CPA
Professor
Texas Christian University
Neeley School of Business
Department of Accounting
TCU Box 298530
Fort Worth, TX 76129
(817) 257-5104
email: c.e.plummer@tcu.edu

Benjamin George, Ph.D.
Assistant Professor
University of South Dakota
Beacom School of Business
Department of Decision Sciences
414 East Clark Street
Vermillion, SD 57069
(940) 367-1109
email: benjamin.george@usd.edu